PROVABLY EFFICIENT TEST-TIME ALIGNMENT VIA SHARED CANDIDATE RESPONSE POOLS

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time alignment enables a fixed language model to serve diverse user preferences without retraining. Existing studies typically assume that the preference is known when responses are generated, whereas in real-world serving systems, the preference may arrive only with the request. To amortize the response generation cost, language model serving systems can pre-generate a shared candidate response pool before an active preference is revealed, serving requests later via importance reweighting. This paper answers a critical question: how large must this shared response pool be to form a near-optimal serving policy for any admissible preference revealed later? Standard importance sampling struggles to solve this problem, as it requires uniformly bounded importance weights, characterized by Γ_∞. With only a finite second-moment coverage coefficient Γ_2, it suffers from linear dependence on the confidence factor $1/\delta$. To overcome this difficulty, we propose an algorithmic framework POOLALIGN. It employs a novel blockwise importance sampling scheme, equipped with within-block normalization and the median-of-means value estimator, to naturally control extreme importance weights without the strong requirement of finite Γ_∞. We prove that POOLALIGN guarantees the ϵ_F-optimality of the serving policy uniformly over prompts, a preference family, and a criteria function class $\mathcal{F}$, using a shared response pool of size $\widetilde{\mathcal{O}}(\Gamma\epsilon_F^{-2}\log(1/\delta))$ per prompt for any valid $\Gamma \geq \Gamma_2$. Crucially, this bound depends on Γ_2 instead of Γ_∞, and has logarthmic rather than linear dependence on $1/\delta$. Moreover, we establish a nearly matching lower bound to demonstrate the tightness of our pool size guarantee for POOLALIGN. Experiments on real LLM response pools validate our theoretical results.

1 INTRODUCTION

Language model systems serve users who prefer different notions of response quality (Achiam et al., 2023; Touvron et al., 2023; Grattafiori et al., 2024). One request may trade off helpfulness against harmlessness (Bai et al., 2022), another informativeness against brevity (Wang et al., 2024a), and a third a formal style over a casual one (Mysore et al., 2024). Training-time alignment such as RLHF (Christiano et al., 2017; Ouyang et al., 2022) and DPO (Rafailov et al., 2023) optimizes model parameters for one preference objective, and a different objective requires retraining. Test-time alignment instead keeps the model fixed and uses the objective at inference, for example through guided decoding (Ji et al., 2024).

In both cases, the preference objective is assumed to be available when the responses are produced. In realistic serving systems, however, the preference is user-dependent and arrives only with the request, while the prompt often reappears across requests or is known in advance. Many serving pipelines therefore separate generation from selection: generation runs ahead of time, and preference-conditioned selection runs at request time. This amortizes the cost of autoregressive response generation (Kwon et al., 2023; Zhong et al., 2024) through offline batching (Zhao et al., 2024b; Qiao et al., 2024) and response caching (Bang, 2023): responses are generated once and subsequently reweighted for later requests.

This pattern determines the information order. For each deployment prompt, the system first draws a *preference-agnostic* pool of responses from a proposal distribution. The active preference is revealed afterward, and the drawn responses may then be reweighted for the specific preference. Since we

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.

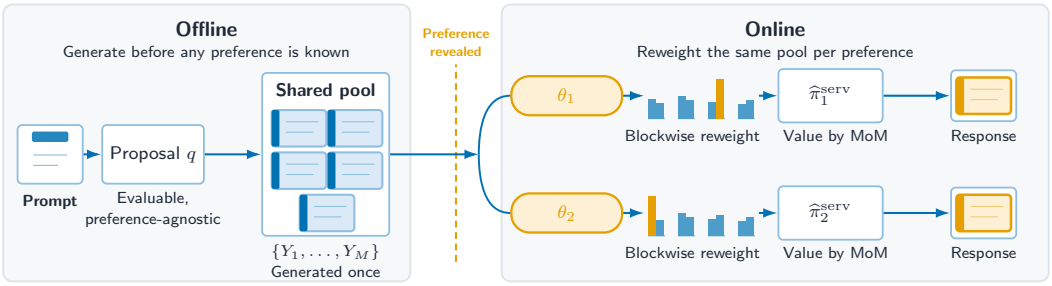


Figure 1: The post-generation preference revelation protocol. For each prompt, a pool of candidate responses is generated before an active preference is revealed. After revelation, our algorithm POOLALIGN uses the shared response pool to construct a preference-dependent serving policy.

do not generate fresh responses for every active preference, serving Q preferences using one shared pool of M responses requires only M generations rather than QM. We call this protocol *post-generation preference revelation* (Figure 1). After the preference is revealed, the system constructs a serving policy supported on the shared response pool by reweighting its responses according to the preference. In this paper, we answer a crucial question: how large must the shared response pool be for the serving policy to achieve near-optimality uniformly over all admissible preferences?

This question becomes technically difficult due to the order in which information arrives. First, since the target (optimal) policy for an active preference is revealed after the response pool is generated, a single high-probability event must hold simultaneously over the prompts, preference family, and criteria function class. This requires controlling the worst-case second moment of the target-to-proposal density ratio, denoted by Γ_2, rather than the per-target second moment that suffices only when the target policy is fixed before sampling (Beirami et al., 2025; Amini et al., 2025). The resulting challenge is to derive a guarantee that depends only on Γ_2, without requiring the stronger uniform bound Γ_∞ on the target-to-proposal density ratio. Second, a finite Γ_2 alone poses a concentration challenge: with finite Γ_∞, Bernstein-type analyses of self-normalized importance sampling (SNIS) yield logarithmic dependence on $1/\delta$, whereas using only a finite Γ_2 typically leads to linear dependence on $1/\delta$ through a second-moment bound (Agapiou et al., 2017; Boucheron et al., 2013).

We develop an algorithmic framework POOLALIGN to address these difficulties. The key idea is to partition the response pool into disjoint blocks, apply SNIS within each block, and average the resulting policies to form a serving policy. The within-block normalization keeps each block evaluation bounded regardless of the raw importance weights. This decomposes the error into two parts: Γ_2 controls the within-block bias, while the bounded and independent block evaluations enable concentration of their average via Bernstein's inequality. Since the within-block normalization cancels out the normalizer Z_θ of the target policy, we can estimate the optimal value $\psi_\theta = \beta \log Z_\theta$ separately by median of means (MoM) (Lugosi & Mendelson, 2019), which gives logarithmic dependence on $1/\delta$ under a finite second moment. A preference net then extends the guarantee from a finite preference set to a continuous family, and an optional pilot procedure selects the proposal and provides a coverage bound $\Gamma \geq \Gamma_2$ using pilot responses. We prove that a shared response pool of size $M = \widetilde{\mathcal{O}}(\Gamma \epsilon_F^{-2} \log(1/\delta))$ for any valid $\Gamma \geq \Gamma_2$ suffices to guarantee ϵ_F-optimality of the serving policy uniformly over a preference family and a criteria function class, with logarithmic rather than linear dependence on $1/\delta$. Here $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic factors but retains the dependence on δ. Moreover, our lower bound validates the tightness of this pool size guarantee for POOLALIGN. All proofs are deferred to Appendix.

Our contributions are summarized as follows.

- We formulate the post-generation preference revelation problem arising in modern language model serving systems, and investigate how large the candidate response pool must be to form a near-optimal serving policy for every admissible preference arriving later.
- An algorithmic framework POOLALIGN is proposed, which applies SNIS within partitioned blocks of the response pool, estimates the value of each block policy by median of means, and mixes the block policies into the serving policy. With a pool of size $M = \widetilde{\mathcal{O}}(\Gamma \epsilon_F^{-2} \log(1/\delta))$ per

prompt for any valid $\Gamma \geq \Gamma_2$, the output policy of POOLALIGN attains ϵ_F-optimality uniformly over a preference family, depending on Γ_2 instead of Γ_∞ and scaling logarithmically rather than linearly with $1/\delta$.

- We establish a lower bound on the required pool size M that matches our upper bound in terms of the coverage coefficient Γ_2, the accuracy parameter ϵ_F, and the joint-preference confidence factor $\log(N/\delta)$, demonstrating that POOLALIGN is near-optimal.
- We corroborate our theoretical results on real LLM response pools generated by Llama-3.1-8B-Instruct (Grattafiori et al., 2024) for 200 prompts from the HH-RLHF dataset (Bai et al., 2022).

2 RELATED WORK

Test-time alignment. Many works steer a frozen language model toward a prescribed reward or preference at inference. Existing approaches include sampling-importance resampling (Ji et al., 2024), twisted sequential Monte Carlo (Zhao et al., 2024a), QAlign (Faria & Smith, 2025), Soft Best-of-n (Verdun et al., 2025), and guided speculative inference (Geuter et al., 2026). Best-of-n generates candidates and selects among them according to a specified objective (Beirami et al., 2025; Amini et al., 2025). Other approaches control the decoding distribution directly (Liu et al., 2024; Mudgal et al., 2024) or train preference-conditioned models where the preference, such as a trade-off between helpfulness and safety, is chosen at deployment (Wang et al., 2024a; Yang et al., 2024; Zhou et al., 2024). All of these use the objective while candidates are generated or selected.

Importance sampling and proposal design. For self-normalized importance sampling, the second moment of the target-to-proposal density ratio, equivalently one plus a Pearson χ^2 divergence, controls the standard bias and mean-square error bounds (Agapiou et al., 2017). High-confidence analyses (Kuzborskij et al., 2021; Metelli et al., 2018; 2021), defensive mixtures (Hesterberg, 1995; Owen & Zhou, 2000) and proposal optimization (He & Owen, 2014; Roy & Evangelou, 2024) have been developed for a fixed target or a predeclared collection of targets. Truncation and winsorization address large importance weights (Ionides, 2008; Orenstein, 2022), while bias-reduced self-normalized importance sampling (BR-SNIS) reduces finite-sample bias through iterated sampling-importance resampling (Cardoso et al., 2022). In these works, the target is fixed before the sample is drawn.

Reward-agnostic data collection. Offline batching (Zhao et al., 2024b; Qiao et al., 2024) and response caching (Bang, 2023) amortize the cost of response generation in language model serving systems. Reward-free and reward-agnostic exploration similarly collects data before the downstream reward is specified, with coverage or concentrability determining how well policies can later be learned for different rewards (Jin et al., 2020; Li et al., 2024; 2023). This line of work similarly collects data before the objective is specified, but its samples are trajectories and the returned policy need not be supported on the collected samples.

We instead generate a shared pool of candidate responses before the active preference is revealed, and provide a uniform guarantee over the entire preference family for a serving policy supported on this pool. Our coverage quantity therefore measures the mismatch between the proposal and target response distributions uniformly over the preference family. We present a comparison with related work in terms of theoretical guarantees in Section A.

3 PROBLEM FORMULATION

In this section, we formalize the post-generation preference revelation problem.

3.1 PREFERENCE-CONDITIONED TARGET POLICY

A problem instance is characterized by a tuple $(\mathcal{X}, \mathcal{Y}, \pi_0, \beta, \Theta, \{s_\theta\}_{\theta\in\Theta})$. Here $\mathcal{X}$ is the prompt space, $\mathcal{Y}$ the full space of responses, π_0 a given reference policy, $\beta > 0$ a parameter controlling the weight of the KL-regularization term relative to π_0, and (Θ, d_Θ) a nonempty compact metric space indexing the preferences, where each preference θ specifies a sequence-level score function $s_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. The deployment prompts $\boldsymbol{x}_{1:n} \coloneqq (x_1, \ldots, x_n) \in \mathcal{X}^n$ are given and fixed, and we denote $[n] \coloneqq \{1, \ldots, n\}$.

Under deployment prompt x and preference θ, the goal is to maximize the following KL-regularized objective over policies π such that $\mathrm{supp}(\pi(\cdot \mid x)) \subseteq \mathrm{supp}(\pi_0(\cdot \mid x))$:

$$\mathcal{J}_{\theta,x}(\pi) \coloneqq \mathbb{E}_{y\sim\pi(\cdot|x)}[s_\theta(x,y)] - \beta D_{\mathrm{KL}}(\pi(\cdot \mid x)\|\pi_0(\cdot \mid x))\,. \tag{1}$$

Denote the partition function by $Z_\theta(x) \coloneqq \mathbb{E}_{y\sim\pi_0(\cdot|x)}[\exp(s_\theta(x,y)/\beta)]$ and assume that $0 < Z_\theta(x_j) < \infty$ for all $j \in [n]$ and $\theta \in \Theta$. For each prompt x, the unique maximizer of Eq. (1) is the Gibbs tilt, i.e., an exponential reweighting of π_0:

$$\pi^\star_\theta(y \mid x) = \frac{\exp(s_\theta(x,y)/\beta)}{Z_\theta(x)}\,\pi_0(y \mid x).$$

We call $\pi^\star_\theta$ the *target* of preference θ, and by calculation, the maximum *value* of $\mathcal{J}_{\theta,x}(\pi)$ is $\psi_\theta(x) \coloneqq \beta \log Z_\theta(x)$ (Korbak et al., 2022). The score function $s_\theta(x,y)$ can be a learned reward model, a DPO-style implicit reward signal (Rafailov et al., 2023), or a scalarization of multiple metrics.

3.2 Shared Response Pool Protocol

For each prompt x_j, the serving system fixes a *proposal* q_j, i.e., a distribution on responses. Denote $q = (q_1, \ldots, q_n)$, which may depend on π_0, β, and the score function family, but not on the active preference. We require the proposal to cover the support of the reference policy, i.e., $q_j(y) > 0$ whenever $\pi_0(y \mid x_j) > 0$, and define the sequence-level likelihood ratio $R_j(y) \coloneqq \pi_0(y \mid x_j)/q_j(y)$. Our algorithmic framework only requires this ratio to be evaluable for each sampled response, without imposing any additional structural assumptions on the proposal. In particular, choosing $q_j = \pi_0(\cdot \mid x_j)$ gives $R_j \equiv 1$. Then for each prompt x_j, the system draws a fresh response pool $\mathcal{C}_j \coloneqq (Y_{j,1}, \ldots, Y_{j,M})$ with $Y_{j,1}, \ldots, Y_{j,M} \overset{\text{i.i.d.}}{\sim} q_j$, and the response pools are drawn independently across prompts. The active preference θ is revealed only after the response pools $\mathcal{C}_{1:n} \coloneqq (\mathcal{C}_1, \ldots, \mathcal{C}_n)$ are realized. After preference revelation, we form a serving policy supported on the realized response pool, denoted by $\widehat{\pi}^{\mathrm{serv}}_{j,\theta} = \sum_{i=1}^{M} W_{j,\theta,i}\,\delta_{Y_{j,i}}$. Here the weights $W_{j,\theta,i} \geq 0$ sum to one and may depend on θ and the response pool, and δ_y denotes the Dirac distribution concentrated at y. The information order is

$$(\boldsymbol{x}_{1:n}, \pi_0, \beta, \Theta, \{s_\theta\}_{\theta\in\Theta}, q) \longrightarrow \mathcal{C}_{1:n} \longrightarrow \theta \longrightarrow \widehat{\pi}^{\mathrm{serv}}_{1:n,\theta}.$$

Each response pool is shared across preferences, but response pools are not shared across prompts.

3.3 Uniform Guarantee and Sample Complexity

Let $\mathcal{F}$ be a finite, prespecified class of functions $f : \mathcal{X} \times \mathcal{Y} \to [-1, 1]$, which encode the criteria used to evaluate the policy. For a serving policy supported on the response pool and an estimator $\widetilde{\psi}_{j,\theta}$ of the optimal value $\psi_\theta(x_j)$ formed from the same response pool, we consider the following two performance metrics:

$$\mathrm{Err}_{\mathcal{F}} \coloneqq \max_{j\in[n]} \sup_{\theta\in\Theta} \max_{f\in\mathcal{F}} \left|\mathbb{E}_{y\sim\widehat{\pi}^{\mathrm{serv}}_{j,\theta}}[f(x_j,y)] - \mathbb{E}_{y\sim\pi^\star_\theta(\cdot|x_j)}[f(x_j,y)]\right|, \tag{2}$$

$$\mathrm{Err}_{\psi} \coloneqq \max_{j\in[n]} \sup_{\theta\in\Theta} \left|\widetilde{\psi}_{j,\theta} - \psi_\theta(x_j)\right|. \tag{3}$$

Given tolerance parameters $\epsilon_F, \epsilon_\psi > 0$ and $\delta \in (0,1)$, we aim to achieve the guarantee:

$$\mathbb{P}(\mathrm{Err}_{\mathcal{F}} \leq \epsilon_F,\ \mathrm{Err}_{\psi} \leq \epsilon_\psi) \geq 1 - \delta. \tag{4}$$

The definitions of $\mathrm{Err}_{\mathcal{F}}$ and Err_{ψ} contain the suprema over θ, and thus the error guarantee holds simultaneously for every preference $\theta \in \Theta$. For fixed $(q, \Theta, \mathcal{F}, \boldsymbol{x}_{1:n})$ and tolerance parameters $(\epsilon_F, \epsilon_\psi, \delta)$, we define the *sample complexity* of an algorithm as the smallest per-prompt pool size M_0 such that, for every $M \geq M_0$, the serving policy and the value estimator that the algorithm outputs from a shared response pool of size M satisfy Eq. (4).

3.4 Coverage Coefficient

The difficulty of covering the target distribution with a response pool drawn from a proposal distribution is captured by the normalized target-to-proposal ratio:

$$U_\theta(x_j, y) \coloneqq \frac{\pi^\star_\theta(y \mid x_j)}{q_j(y)} = \frac{R_j(y)}{Z_\theta(x_j)} \exp\!\left(\frac{s_\theta(x_j,y)}{\beta}\right).$$

It holds that $\mathbb{E}_{y\sim q_j}[U_\theta(x_j, y)] = 1$. We use two worst-case quantities to characterize this ratio:

$$\Gamma_2(q, \Theta; \boldsymbol{x}_{1:n}) \coloneqq \max_{j\in[n]} \sup_{\theta\in\Theta} \mathbb{E}_{y\sim q_j}\left[U_\theta(x_j, y)^2\right] = 1 + \max_{j\in[n]} \sup_{\theta\in\Theta} \chi^2(\pi^\star_\theta(\cdot \mid x_j)\|q_j(\cdot))\,, \tag{5}$$

$$\Gamma_\infty(q, \Theta; \boldsymbol{x}_{1:n}) \coloneqq \max_{j\in[n]} \sup_{\theta\in\Theta} \sup_{y:\, q_j(y)>0} U_\theta(x_j, y). \tag{6}$$

The *coverage coefficient* Γ_2 is a nature measure of importance-sampling difficulty (Agapiou et al., 2017; Owen, 2013) and is finite whenever the target-to-proposal ratio is square-integrable. Bernstein-type SNIS analyses (Boucheron et al., 2013) instead require the stronger uniform *envelope* Γ_∞, which becomes infinite once some ratio is unbounded, e.g., when the proposal assigns too little mass to responses favored by the target. Our main guarantees depend on Γ_2 rather than Γ_∞, and the following proposition shows that the shared response pool can inherently lead to a large Γ_2.

Proposition 1 (Price of a shared response pool)**.** *Let $p_1, \dots, p_N$ be probability mass functions on a finite response space. Define the smallest worst-case second moment of the target-to-proposal ratio achievable by a proposal as $V_2 \coloneqq \min_q \max_{\ell\in[N]} \sum_y p_\ell(y)^2/q(y)$, with $0/0 \coloneqq 0$ and $p/0 \coloneqq +\infty$ for $p > 0$. Then, it holds that $1 \le V_2 \le N$: it is 1 when $p_1, \dots, p_N$ all coincide and N when their supports are pairwise disjoint, with the latter attained by $q = \frac{1}{N}\sum_\ell p_\ell$ (Propositions 26 and 27).*

No proposal escapes the factor N once the targets separate, and the optimal coverage coefficient for a shared response pool can be as large as the number of well-separated preferences. This price is unavoidable: our guarantee requires a pool size linear in Γ_2 (Corollary 3), and our lower bound (Theorem 5) shows that the linear dependence cannot be removed.

4 PoolAlign: Test-Time Alignment from a Shared Response Pool

In this section, we present PoolAlign, illustrated in Algorithm 1. PoolAlign generates a shared pool of candidate responses before the active preference is revealed, and then serves every preference from the shared pool, enjoying a guarantee with tight coverage dependence.

4.1 Blockwise Policy Construction

Since the target policy $\pi^\star_\theta$ tilts the reference policy by $\exp(s_\theta/\beta)$, the unnormalized importance weight of a candidate response under a revealed preference θ is

$$A_\theta(x_j, y) \coloneqq R_j(y)\, \exp\left(\frac{s_\theta(x_j, y)}{\beta}\right).$$

This weight satisfies $A_\theta(x_j, \cdot) = Z_\theta(x_j)U_\theta(x_j, \cdot)$ and $\mathbb{E}_{y\sim q_j}[A_\theta(x_j, y)] = Z_\theta(x_j)$, and determines the target policy up to the unknown scale $Z_\theta(x_j)$. Ordinary self-normalized importance sampling cancels out this scale by normalizing all M weights jointly, but a single large weight can then dominate the resulting distribution (Owen, 2013; Chatterjee & Diaconis, 2018), and a high-probability analysis relies on the envelope Γ_∞. In contrast, we first partition the M indices of candidate responses for each prompt into K balanced blocks $\mathcal{B}_{j,1}, \dots, \mathcal{B}_{j,K}$ (Line 1). Then, we normalize the weights within each block to obtain a block policy $\widehat{\pi}^{(k)}_{j,\theta}$, and average the resulting block policies to form $\widehat{\pi}^{\mathrm{blk}}_{j,\theta}$ (Lines 7–8):

$$\widehat{\pi}^{(k)}_{j,\theta} \coloneqq \frac{\sum_{i\in\mathcal{B}_{j,k}} A_\theta(x_j, Y_{j,i})\delta_{Y_{j,i}}}{\sum_{i\in\mathcal{B}_{j,k}} A_\theta(x_j, Y_{j,i})}, \qquad \widehat{\pi}^{\mathrm{blk}}_{j,\theta} \coloneqq \frac{1}{K}\sum_{k=1}^{K} \widehat{\pi}^{(k)}_{j,\theta}, \tag{7}$$

where a block policy $\widehat{\pi}^{(k)}_{j,\theta}$ with vanishing denominator is taken to be uniform over its candidate responses. For every $f : \mathcal{X}\times\mathcal{Y} \to [-1, 1]$, the block functional $\widehat{\pi}^{(k)}_{j,\theta} f \coloneqq \mathbb{E}_{y\sim\widehat{\pi}^{(k)}_{j,\theta}}[f(x_j, y)]$ lies in $[-1, 1]$ regardless of the within-block weights. Since the K block functionals are independent and all bounded in $[-1, 1]$, Bernstein concentration bounds their average without dependence on the envelope Γ_∞, while Γ_2 controls the bias introduced by the within-block normalization.

Algorithm 1 POOLALIGN: blockwise serving policy and median-of-means value estimate

Require: Fixed prompts $\boldsymbol{x}_{1:n}$, reference policy π_0, proposals $q_{1:n}$, score family $\{s_\theta\}_{\theta\in\Theta}$, β, M, $1 \le K \le M$, and odd $1 \le K_Z \le M$.

1: Fix two partitions of candidate response indices $\{\mathcal{B}_{j,k}\}_{k=1}^{K}$, $\{\mathcal{B}_{j,r}^{Z}\}_{r=1}^{K_Z}$, each containing disjoint blocks.
 # Pool construction (before preference revelation)
2: **for** $j \in [n]$ **do**
3: Draw $Y_{j,1:M} \overset{\text{i.i.d.}}{\sim} q_j$ and cache $R_{j,i} \leftarrow \pi_0(Y_{j,i} \mid x_j)/q_j(Y_{j,i})$ for all $i \in [M]$.
4: **end for**
 # Serving (for each revealed preference θ)
5: **for** $j \in [n]$ **do**
6: Set $A_{j,\theta,i} \leftarrow R_{j,i} \exp(s_\theta(x_j, Y_{j,i})/\beta)$ for all $i \in [M]$.
7: For each $k \in [K]$, normalize within block $\mathcal{B}_{j,k}$ to form $\widehat{\pi}_{j,\theta}^{(k)}$ as in Eq. (7)
8: Set $\widehat{\pi}_{j,\theta}^{\text{blk}} \leftarrow \frac{1}{K}\sum_{k=1}^{K} \widehat{\pi}_{j,\theta}^{(k)}$.
9: Set $\widetilde{Z}_{j,\theta}^{\text{MoM}}$ to the median of the K_Z blockwise weight means and $\widetilde{\psi}_{j,\theta}^{\text{MoM}} \leftarrow \beta \log \widetilde{Z}_{j,\theta}^{\text{MoM}}$ as in Eq. (8), where $\log 0 := -\infty$ and we report a zero median as a value estimation failure.
10: **end for**
11: **return** $\{(\widehat{\pi}_{j,\theta}^{\text{blk}}, \widetilde{\psi}_{j,\theta}^{\text{MoM}})\}_{j=1}^{n}$.

4.2 VALUE ESTIMATION BY MEDIAN OF MEANS

Since the normalization within blocks cancels out the scale $Z_\theta(x_j)$, the block policies alone do not identify the optimal value $\psi_\theta(x_j)$. To address this challenge, we use a second balanced partition of candidate response indices with an odd number K_Z of blocks, take the median of the blockwise weight means to estimate $Z_\theta(x_j)$, and construct an estimator of $\psi_\theta(x_j)$ as follows (Line 9):

$$\widetilde{Z}_{j,\theta}^{\text{MoM}} := \text{median}_{r\in[K_Z]} \left\{ \frac{1}{|\mathcal{B}_{j,r}^{Z}|} \sum_{i\in\mathcal{B}_{j,r}^{Z}} A_\theta(x_j, Y_{j,i}) \right\}, \qquad \widetilde{\psi}_{j,\theta}^{\text{MoM}} := \beta \log \widetilde{Z}_{j,\theta}^{\text{MoM}}. \tag{8}$$

Median of means (MoM) is a popular estimator that achieves logarithmic dependence on $1/\delta$ for unbounded random variables under a finite second-moment condition (Alon et al., 1999; Devroye et al., 2016; Lugosi & Mendelson, 2019).

4.3 THE POOLALIGN PROCEDURE

Algorithm 1 integrates the blockwise serving policy construction with the median-of-means value estimator, both using the same response pool drawn in Line 3. The two partitions $\{\mathcal{B}_{j,k}\}_{k=1}^{K}$ and $\{\mathcal{B}_{j,r}^{Z}\}_{r=1}^{K_Z}$ are fixed before any candidate response is drawn and thus do not depend on the realized weights. The block counts involve the following trade-offs: a larger K averages more block functionals but shrinks every block and inflates the within-block bias, while a larger K_Z improves the confidence of the median-of-means estimator at the cost of noisier block means. Let $\mathfrak{L}_{\text{P}}$ and $\mathfrak{L}_{\text{V}}$ denote the logarithmic factors for the policy and value in the error bound (Theorem 2). Given a valid bound of the coverage coefficient $\Gamma \ge \Gamma_2$, to balance the within-block bias, block fluctuation, and Bernstein error in the error bound, we choose

$$K \asymp \sqrt{\frac{M\mathfrak{L}_{\text{P}}}{\Gamma}}, \qquad \min_k |\mathcal{B}_{j,k}| \asymp \sqrt{\frac{\Gamma M}{\mathfrak{L}_{\text{P}}}}, \qquad K_Z \asymp \mathfrak{L}_{\text{V}}. \tag{9}$$

After rounding, K and K_Z must also satisfy the conditions in Theorem 2.

4.4 PROPOSAL SELECTION USING PILOT RESPONSES

It remains to choose a proposal with a small Γ_2. A natural choice is the defensive mixture (Hesterberg, 1995; Owen & Zhou, 2000), i.e., mixing several proposal distributions $q_0, \ldots, q_J$:

$$q_\alpha(y \mid x) := \sum_{k=0}^{J} \alpha_k \, q_k(y \mid x), \qquad q_0 = \pi_0,$$

where $\alpha_0, \ldots, \alpha_J$ are nonnegative and sum to one, and $\alpha_0 \geq a > 0$, which retains a fixed fraction of the reference policy. This ensures support coverage and provides a bound: if every score $s_\theta(x_j, \cdot)$ varies by at most B over the support of $\pi_0(\cdot \mid x_j)$, then $q_\alpha \geq a\pi_0$ implies $\Gamma_2 \leq \Gamma_\infty \leq a^{-1}e^{B/\beta}$.

Optionally, for each prompt, before generating the candidate response pool, we can first draw H pilot responses to select a good proposal and obtain a sharp data-dependent upper bound on Γ_2. We consider a finite set $\mathcal{A}$ of candidate proposals q_α and evaluate their coverage coefficients over representative preferences $\{\theta_\ell\}_{\ell=1}^N \subseteq \Theta$. Since this involves the unknown normalizer $Z_\theta(x_j)$, we estimate the required coverage quantities using two independent pilot folds (Chernozhukov et al., 2018). We then select the proposal with the smallest estimated coverage coefficient and obtain an upper bound Γ on its coverage coefficient. Using this pilot selection procedure, we generate a total of $H + M$ responses per prompt (Section C.4).

5 Theoretical Guarantees

In this section, we provide the theoretical guarantees for algorithm PoolAlign and a lower bound.

5.1 Uniform Guarantee and Sample Complexity

First, we quantify the price of serving an entire preference family from one response pool. We present the support and sampling conditions in Section D.1, and assume that a deterministic bound $\Gamma_2(q, \Theta; \boldsymbol{x}_{1:n}) \leq \Gamma < \infty$ is known when choosing the block counts. For a continuous preference family, we assume that nearby preferences assign similar scores to responses: letting d_Θ be the metric on Θ defined in Section 3.1, for some $L_s > 0$ and every y with $q_j(y) > 0$,

$$|s_\theta(x_j, y) - s_{\theta'}(x_j, y)| \leq L_s\, d_\Theta(\theta, \theta'), \qquad \theta, \theta' \in \Theta. \tag{10}$$

For any $\rho > 0$, let $\Theta_\rho \subseteq \Theta$ be a deterministic minimum-cardinality ρ-net fixed before response sampling, such that for every $\theta \in \Theta$, there exists some $\theta' \in \Theta_\rho$ satisfying $d_\Theta(\theta, \theta') \leq \rho$. Let $N_\rho \coloneqq |\Theta_\rho|$ denote the corresponding covering number of Θ (Wainwright, 2019). We assume that for some radius $R_\Theta > 0$, some effective dimension $d_{\mathrm{par}} \geq 0$ of the preference parameter space, and every $\rho > 0$,

$$\log N_\rho \leq d_{\mathrm{par}} \log(1 + R_\Theta/\rho). \tag{11}$$

Theorem 2 (Uniform error bound of PoolAlign)**.** *Suppose that Assumptions 6 and 7, Eq.* (10)*, and Eq.* (11) *hold. Choose integer $1 \leq K \leq M$ and odd integer $1 \leq K_Z \leq M$. Define the minimum block sizes $b \coloneqq \min_{j,k} |\mathcal{B}_{j,k}|$ and $b_Z \coloneqq \min_{j,r} |\mathcal{B}^Z_{j,r}|$, and let $\mathfrak{L}_{\mathrm{P}} \coloneqq \log(\frac{4nN_\rho(|\mathcal{F}|+1)}{\delta})$ and $\mathfrak{L}_{\mathrm{V}} \coloneqq \log(\frac{4nN_\rho}{\delta})$. Then, with probability at least $1 - \delta$, the outputs $\{(\widehat{\pi}^{\mathrm{blk}}_{j,\theta}, \widetilde{\psi}^{\mathrm{MoM}}_{j,\theta})\}_{j=1}^n$ of Algorithm 1 satisfy that*

$$\mathrm{Err}_{\mathcal{F}} \leq \underbrace{\frac{16\Gamma}{b}}_{\text{within-block bias}} + \underbrace{8\sqrt{\frac{\Gamma\mathfrak{L}_{\mathrm{P}}}{bK}}}_{\text{block fluctuation}} + \underbrace{\frac{4\mathfrak{L}_{\mathrm{P}}}{3K}}_{\text{bounded-range Bernstein error}} + 2\left(e^{2L_s\rho/\beta} - 1\right), \tag{12}$$

and, whenever $K_Z \geq 8\mathfrak{L}_{\mathrm{V}}$ and $\kappa_Z \coloneqq 2\sqrt{\Gamma/b_Z} \leq \frac{1}{2}$,

$$\mathrm{Err}_\psi \leq 2\beta\kappa_Z + 2L_s\rho.$$

Corollary 3 (Sample complexity)**.** *Given $\epsilon_F \in (0, 1]$, $\epsilon_\psi \in (0, 2\beta]$, and $\delta \in (0, 1)$, run Algorithm 1 with the block counts in Eq.* (9) *at confidence level $\delta/2$, using the ρ-net radii $\rho_F \asymp \beta\epsilon_F/L_s$ for the functional guarantee and $\rho_\psi \asymp \epsilon_\psi/L_s$ for the value guarantee, with the corresponding net sizes denoted by N_{ρ_F} and N_{ρ_ψ}, respectively. Then, $\mathrm{Err}_{\mathcal{F}} \leq \epsilon_F$ and $\mathrm{Err}_\psi \leq \epsilon_\psi$ hold simultaneously with probability at least $1 - \delta$ (i.e., Eq.* (4) *holds), provided that*

$$M = \widetilde{\mathcal{O}}\left(\Gamma \max\left\{\frac{\log N_{\rho_F} + \log(\frac{n(|\mathcal{F}|+1)}{\delta})}{\epsilon_F^2},\ \frac{\beta^2\big(\log N_{\rho_\psi} + \log(\frac{n}{\delta})\big)}{\epsilon_\psi^2}\right\}\right). \tag{13}$$

In Theorem 2, under the covering condition, the dependence on the preference family enters through its metric entropy $\log N_\rho$, and for a finite preference set, we simply take $N_\rho = |\Theta|$ without requiring the Lipschitz condition Eq. (10). Overall, balancing the first three error terms in Eq. (12) makes

the sample complexity scale linearly with Γ and quadratically with the inverse accuracy, yielding $\widetilde{\mathcal{O}}(\Gamma_2 \epsilon_F^{-2} \log(1/\delta))$ when Γ tightly upper bounds Γ_2.

Remark (comparison with one-block SNIS). Compared with one-block SNIS, our algorithm POOLALIGN provides a uniform error guarantee under a weaker coverage condition and achieves a sharper dependence on the confidence level. Specifically, first, our guarantee depends only on the second-moment coverage coefficient Γ_2, which can be upper bounded using the pilot data (Theorem 4). In contrast, standard one-block SNIS analyses require the stronger envelope Γ_∞, which may be infinite and cannot in general be upper bounded from finite samples. Second, with only a finite second moment, one-block concentration yields linear dependence on $1/\delta$ (Metelli et al., 2018), whereas our blockwise design achieves logarithmic dependence on $1/\delta$. This improvement is particularly important for our uniform guarantee, where confidence must hold simultaneously over prompts, preferences, and criteria functions. The blockwise construction separates the two sources of error: Γ_2 controls the bias introduced by within-block normalization, while the normalized block evaluations are bounded and independent, allowing Bernstein concentration for their average without requiring a finite Γ_∞.

5.2 PILOT-TO-SERVING GUARANTEE

Theorem 2 takes as input a coverage bound Γ. The pilot procedure in Section 4.4 provides such a bound by selecting the proposal from a finite candidate proposal set $\mathcal{A}$ using representative preferences $\{\theta_\ell\}_{\ell=1}^N \subseteq \Theta$. We assume that every score s_θ varies by at most B over the support of π_0, and the proposal q_{pil} used to sample pilot responses satisfies $q_{\mathrm{pil}} \geq a_{\mathrm{pilot}} \pi_0$. For each prompt x_j and preference θ_ℓ, let $G_{j\ell}(\alpha) \coloneqq \mathbb{E}_{y \sim q_\alpha(\cdot|x_j)}[(\pi^\star_{\theta_\ell}(y \mid x_j)/q_\alpha(y \mid x_j))^2]$, whose maximum value over (j, ℓ) measures the coverage of q_α.

Theorem 4 (Pilot guarantee)**.** *Let $\delta_{\mathrm{p}} \in (0,1)$, choose accuracy parameters $\eta_Z, \eta_G \in (0,1)$ for the normalizer $Z_{\theta_\ell}(x_j)$ and the coverage $G_{j\ell}(\alpha)$, respectively, and set $\eta \coloneqq \min\{\eta_Z, \eta_G\}$. Suppose that the H pilot responses for each prompt are split into two independent folds of size $h = H/2$, where $h = \Theta\big(\log(\frac{nN(|\mathcal{A}|+1)}{\delta_{\mathrm{p}}}) e^{2B/\beta} a_{\mathrm{pilot}}^{-1} a^{-1} \eta^{-2}\big)$. With probability at least $1 - \delta_{\mathrm{p}}$, the selected proposal $q_{\widehat{\alpha}}$ satisfies*

$$\max_{j,\ell} G_{j\ell}(\widehat{\alpha}) \leq \kappa_{\mathrm{p}} \min_{\alpha \in \mathcal{A}} \max_{j,\ell} G_{j\ell}(\alpha), \quad \text{where } \kappa_{\mathrm{p}} \coloneqq \frac{1+\eta_G}{1-\eta_G} \left(\frac{1+\eta_Z}{1-\eta_Z}\right)^2.$$

In other words, the worst-case coverage of the selected proposal $q_{\widehat{\alpha}}$ is within a factor κ_{p} of the best proposal in the candidate proposal set $\mathcal{A}$.

The factor κ_{p} tends to one as η_Z, η_G approach zero at the cost of a larger pilot fold size h. Combining this with Theorem 2 yields the end-to-end guarantee Corollary 14 for proposal selection and subsequent serving, where the pilot guarantee extends from the representative preferences to the full preference family Θ via a Lipschitz net, with the two failure probabilities added.

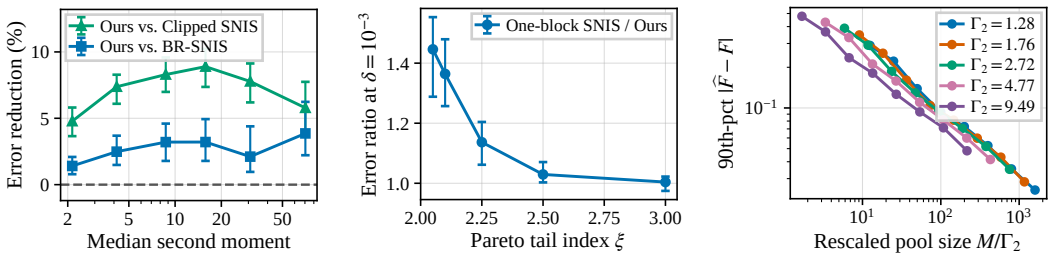
5.3 POOL SIZE LOWER BOUND

Now we show that Theorem 2 is tight. The lower bound considers a single prompt, a fixed proposal, a finite preference family, and an indicator function class $\mathcal{F}$, focusing on the functional error $\mathrm{Err}_{\mathcal{F}}$. A matching lower bound for the value error Err_ψ remains open.

Theorem 5 (Lower bound)**.** *Let $N \geq 2$, $0 < \epsilon \leq 1/4$, $\overline{\Gamma} \geq 2$ and $0 < \delta \leq 1/2$. There exists a family of instances for the post-generation preference revelation problem with $\Theta = \{\theta_1, \ldots, \theta_N\}$, a fixed proposal q, and indicator function class $\mathcal{F} = \{f_1, \ldots, f_N\}$, where the second moment of the target-to-proposal ratio Γ_2 lies in $[\overline{\Gamma} - 1, \overline{\Gamma}]$. On this family, any randomized map from M i.i.d. candidate responses to the serving policies supported on these responses that attains $\mathrm{Err}_{\mathcal{F}} \leq \epsilon$ with probability at least $1 - \delta$ requires*

$$M \geq \frac{3\overline{\Gamma}}{32\epsilon^2} \log\left(\frac{N}{2\delta}\right). \tag{14}$$

The construction uses N distinct rare events, one for each preference, forcing a single shared response pool to capture every rare event simultaneously. The sample complexity of POOLALIGN



(a) Real LLM response pools. (b) Finite Γ_2 but infinite Γ_∞. (c) Linear dependence on Γ_2.

Figure 2: Experimental results. (a) Error reduction of POOLALIGN over baselines on real LLM response pools under different coverage levels. (b) Heavy-tailed importance weights with finite Γ_2 and infinite Γ_∞. (c) Functional error versus the rescaled pool size M/Γ_2 under different Γ_2.

(Corollary 3) matches the lower bound in terms of the coverage coefficient, the squared inverse accuracy, and the joint-preference confidence factor $\log(N/\delta)$. On this family of instances, no strategy using one shared response pool can improve over POOLALIGN beyond logarithmic factors.

6 EXPERIMENTS

In this section, we evaluate our algorithm POOLALIGN in three settings: real LLM response pools generated by Llama-3.1-8B-Instruct (Grattafiori et al., 2024) for 200 prompts from the HH-RLHF dataset (Bai et al., 2022), synthetic Pareto-tailed instances, and synthetic Gaussian instances. We compare POOLALIGN with clipped SNIS, which caps the importance weights at a fixed empirical quantile, BR-SNIS (Cardoso et al., 2022), which reduces self-normalization bias via iterated resampling, and one-block SNIS. Experimental details and more results are presented in Section G.

Figure 2a uses real LLM response pools under seven helpfulness-safety preferences constructed using the reward model ArmoRM (Wang et al., 2024b). We vary the temperature parameter β to create different coverage levels, with the median target-to-proposal second moment over prompt-preference pairs ranging from 2.1 to 70.5, and report the error reduction of POOLALIGN over clipped SNIS and BR-SNIS at $\delta = 10^{-3}$. As shown in the figure, POOLALIGN consistently achieves lower error than both baselines across all coverage levels. Both baselines rely on a bounded importance-weight envelope, and their performance degrades relative to POOLALIGN as the coverage becomes more challenging.

In Figure 2b, we consider Pareto-tailed importance weights with tail index ξ, where Γ_2 is finite while Γ_∞ is infinite. The error ratio of one-block SNIS to POOLALIGN exceeds 1 and grows substantially as the importance weights become more heavy-tailed. This validates that POOLALIGN depends on Γ_2 rather than Γ_∞, and enjoys a significant advantage over one-block SNIS. Figure 2c uses Gaussian instances with a single prompt, $\mathcal{Y} = \mathbb{R}$, $q = \pi_0 = \mathcal{N}(0, 1)$, $s_\theta(y) = \theta y$ and $\beta = 1$, so that $\pi^\star_\theta = \mathcal{N}(\theta, 1)$ and $\Gamma_2(\theta) = e^{\theta^2}$. We plot the functional error against the rescaled pool size M/Γ_2. One can see that the five error curves under different Γ_2 almost coincide, which corroborates the linear dependence of the pool size on Γ_2 and is consistent with our theoretical results.

7 CONCLUSION

In this work, we formulate the post-generation preference revelation problem, where one shared response pool serves every preference revealed after response generation. We identify Γ_2, the worst-case target-to-proposal second moment, as the quantity governing the required size of the response pool. We design algorithmic framework POOLALIGN, which couples blockwise self-normalized importance sampling with median-of-means value estimation, and prove that a response pool of size $\widetilde{\mathcal{O}}(\Gamma\epsilon_F^{-2}\log(1/\delta))$ per prompt for any valid $\Gamma \geq \Gamma_2$ guarantees ϵ_F-optimality of the serving policy uniformly over a preference family. A nearly matching lower bound is also established.

AI USE STATEMENT

Generative AI tools were used for language editing, document organization, LaTeX assistance, and technical consistency and proof checks. The authors independently reviewed and verified all AI-assisted material and take responsibility for the paper's arguments, claims, proofs, citations, and final presentation.

ETHICS STATEMENT

Our language-model experiment uses publicly released HH-RLHF prompts and open model checkpoints to generate responses. It introduces no new human annotation or private data. Some source prompts concern harmful requests. They are used only to evaluate how accurately our algorithm approximates the target policy, rather than to assess or improve model safety.

REPRODUCIBILITY STATEMENT

This paper states the sampling model, information order, POOLALIGN algorithm, assumptions, and finite-sample guarantees clearly. Complete proofs and the auxiliary results are provided in the Appendix. Section G describes the detailed setup of our experiments, including the target policies, language-model response pools, evaluation protocols, random seeds, and aggregation procedures. We provide our code in supplementary materials to ensure the reproducibility of our experiments.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. URL https://arxiv.org/abs/2303.08774.

Sergios Agapiou, Omiros Papaspiliopoulos, Daniel Sanz-Alonso, and Andrew M. Stuart. Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, 32(3):405–431, 2017. doi: 10.1214/17-STS611.

Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1):137–147, 1999. doi: 10.1006/jcss.1997.1545.

Afra Amini, Tim Vieira, Elliott Ash, and Ryan Cotterell. Variational best-of-n alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=W9FZEQj3vv.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Fu Bang. GPTCache: An open-source semantic cache for LLM applications enabling faster answers and cost savings. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pp. 212–218, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.nlposs-1.24. URL https://aclanthology.org/2023.nlposs-1.24/.

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

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. doi: 10.1093/acprof:oso/9780199535255.001.0001.

Ralph A. Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Gabriel Cardoso, Sergey Samsonov, Achille Thin, Eric Moulines, and Jimmy Olsson. BR-SNIS: Bias reduced self-normalized importance sampling. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/04bd683d5428d91c5fbb5a7d2c27064d-Abstract-Conference.html.

Olivier Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4), 2012. doi: 10.1214/11-AIHP454.

Sourav Chatterjee and Persi Diaconis. The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2):1099–1135, 2018. doi: 10.1214/17-AAP1326. URL https://doi.org/10.1214/17-AAP1326.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4299–4307, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html.

Luc Devroye, Matthieu Lerasle, Gábor Lugosi, and Roberto I. Oliveira. Sub-Gaussian mean estimators. *The Annals of Statistics*, 44(6), 2016. doi: 10.1214/16-AOS1440.

Gonçalo Faria and Noah A. Smith. Sample, don't search: Rethinking test-time alignment for language models, 2025. URL https://arxiv.org/abs/2504.03790.

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

Jonathan Geuter, Youssef Mroueh, and David Alvarez-Melis. Guided speculative inference for efficient test-time alignment of LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=miNzDqDENd.

Aaron Grattafiori et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Hera Y. He and Art B. Owen. Optimal mixture weights in multiple importance sampling, 2014. URL https://arxiv.org/abs/1411.3954.

Tim Hesterberg. Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–194, 1995. doi: 10.1080/00401706.1995.10484303.

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

Mark R. Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:169–188, 1986. doi: 10.1016/0304-3975(86)90174-X.

Haozhe Ji, Pei Ke, Hongning Wang, and Minlie Huang. Language model decoding as direct metrics optimization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=488A64eOf6.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4870–4879. PMLR, 2020. URL https://proceedings.mlr.press/v119/jin20d.html.

Tomasz Korbak, Ethan Perez, and Christopher L. Buckley. RL with KL penalties is better viewed as Bayesian inference. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022. URL https://arxiv.org/abs/2205.11275.

Ilja Kuzborskij, Claire Vernade, Andras Gyorgy, and Csaba Szepesvari. Confident off-policy evaluation and selection through self-normalized importance weighting. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 640–648. PMLR, 2021. URL https://proceedings.mlr.press/v130/kuzborskij21a.html.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165. URL https://arxiv.org/abs/2309.06180.

Gen Li, Wenhao Zhan, Jason D. Lee, Yuejie Chi, and Yuxin Chen. Reward-agnostic fine-tuning: Provable statistical benefits of hybrid reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2426. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ade04fd4f26263f86b47ffb535c4cafb-Abstract-Conference.html.

Gen Li, Yuling Yan, Yuxin Chen, and Jianqing Fan. Minimax-optimal reward-agnostic exploration in reinforcement learning. In *Proceedings of the 37th Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 3431–3436. PMLR, 2024. URL https://proceedings.mlr.press/v247/li24a.html.

Tianlin Liu, Shangmin Guo, Leonardo Bianco, Daniele Calandriello, Quentin Berthet, Felipe Llinares-López, Jessica Hoffmann, Lucas Dixon, Michal Valko, and Mathieu Blondel. Decoding-time realignment of language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 31015–31031. PMLR, 2024. URL https://proceedings.mlr.press/v235/liu24r.html.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019. doi: 10.1007/s10208-019-09427-x.

Alberto Maria Metelli, Matteo Papini, Francesco Faccio, and Marcello Restelli. Policy optimization via importance sampling. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/6aed000af86a084f9cb0264161e29dd3-Abstract.html.

Alberto Maria Metelli, Alessio Russo, and Marcello Restelli. Subgaussian and differentiable importance sampling for off-policy evaluation and learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/4476b929e30dd0c4e8bdbcc82c6ba23a-Abstract.html.

Stanislav Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4), 2015. doi: 10.3150/14-BEJ645.

Sidharth Mudgal, Jong Lee, Harish Ganapathy, Yaguang Li, Tao Wang, Yanping Huang, Zhifeng Chen, Heng-Tze Cheng, Michael Collins, Trevor Strohman, Jilin Chen, Alex Beutel, and Ahmad Beirami. Controlled decoding from language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36486–36503. PMLR, 2024. URL https://proceedings.mlr.press/v235/mudgal24a.html.

Sheshera Mysore, Zhuoran Lu, Mengting Wan, Longqi Yang, Bahareh Sarrafzadeh, Steve Menezes, Tina Baghaee, Emmanuel Barajas Gonzalez, Jennifer Neville, and Tara Safavi. PEARL: Personalizing large language model writing assistants with generation-calibrated retrievers. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pp. 198–219, Miami, Florida, USA,

2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.customnlp4u-1.16. URL https://aclanthology.org/2024.customnlp4u-1.16.

Paulo Orenstein. Robust importance sampling with adaptive winsorization. *Bernoulli*, 28(4):2862–2873, 2022. doi: 10.3150/21-BEJ1440.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-2011. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract.html.

Art B. Owen. *Monte Carlo Theory, Methods and Examples*. 2013. URL https://artowen.su.domains/mc/. Book draft.

Art B. Owen and Yi Zhou. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143, 2000. doi: 10.1080/01621459.2000.10473909.

Yifan Qiao, Shu Anzai, Shan Yu, Haoran Ma, Shuo Yang, Yang Wang, Miryung Kim, Yongji Wu, Yang Zhou, Jiarong Xing, Joseph E. Gonzalez, Ion Stoica, and Harry Xu. ConServe: Fine-grained GPU harvesting for LLM online and offline co-serving. *arXiv preprint arXiv:2410.01228*, 2024. URL https://arxiv.org/abs/2410.01228.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Vivekananda Roy and Evangelos Evangelou. Selection of proposal distributions for multiple importance sampling. *Statistica Sinica*, 34(1):27–46, 2024. doi: 10.5705/ss.202021.0049.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. URL https://arxiv.org/abs/2307.09288.

Eric Veach and Leonidas J. Guibas. Optimally combining sampling techniques for Monte Carlo rendering. In *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '95)*, pp. 419–428, 1995. doi: 10.1145/218380.218498.

Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. Pareto smoothed importance sampling. *Journal of Machine Learning Research*, 25(72):1–58, 2024. URL https://jmlr.org/papers/v25/19-556.html.

Claudio Mayrink Verdun, Alex Oesterling, Himabindu Lakkaraju, and Flavio P. Calmon. Soft best-of-n sampling for model alignment. In *IEEE International Symposium on Information Theory*, pp. 1–6, 2025. doi: 10.1109/ISIT63088.2025.11195541.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. Arithmetic control of LLMs for diverse user preferences: Directional preference alignment with multi-objective rewards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8642–8655, Bangkok, Thailand, 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.468. URL https://aclanthology.org/2024.acl-long.468.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10582–10592, Miami, Florida, USA, 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.620. URL https://aclanthology.org/2024.findings-emnlp.620/.

Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 56276–56297. PMLR, 2024. URL https://proceedings.mlr.press/v235/yang24q.html.

Stephen Zhao, Rob Brekelmans, Alireza Makhzani, and Roger Baker Grosse. Probabilistic inference in language models via twisted sequential Monte Carlo. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 60704–60748. PMLR, 2024a. URL https://proceedings.mlr.press/v235/zhao24c.html.

Yilong Zhao, Shuo Yang, Kan Zhu, Lianmin Zheng, Baris Kasikci, Yang Zhou, Jiarong Xing, and Ion Stoica. BlendServe: Optimizing offline inference for auto-regressive large models with resource-aware batching. *arXiv preprint arXiv:2411.16102*, 2024b. URL https://arxiv.org/abs/2411.16102.

Yinmin Zhong, Shengyu Liu, Junda Chen, Jianbo Hu, Yibo Zhu, Xuanzhe Liu, Xin Jin, and Hao Zhang. DistServe: Disaggregating prefill and decoding for goodput-optimized large language model serving. In *18th USENIX Symposium on Operating Systems Design and Implementation*, pp. 193–210, 2024. URL https://arxiv.org/abs/2401.09670.

Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao Yang, Wanli Ouyang, and Yu Qiao. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10586–10613, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.630. URL https://aclanthology.org/2024.findings-acl.630.

A Extended Discussion of Related Work

Section 2 places our work at the intersection of test-time alignment, importance sampling, and reward-agnostic data collection. This appendix clarifies the comparison with the closest results along the dimensions that determine the scope of their guarantees.

Test-time alignment and blockwise normalization. Controlled decoding uses a learned prefix scorer in a blockwise decoding procedure that interpolates between best-of-K and tokenwise control (Mudgal et al., 2024). In that setting a block is a sequence of generated tokens within one response, and the scorer acts during generation. In PoolAlign, blocks partition an already realized candidate pool and are used for post-generation normalization, and the preference is supplied after the pool has been fixed. The two procedures therefore operate at different stages of the generation pipeline despite sharing the term "blockwise," which we use in the pool sense throughout.

Coverage assumptions. For self-normalized importance sampling, the density-ratio second moment

$$\mathbb{E}_q\big[U_\theta^2\big] = 1 + \chi^2(\pi_\theta^\star \| q) \tag{15}$$

is the natural quantity in standard bias and mean-square error bounds (Agapiou et al., 2017). It also controls the relative-entropy gap through $D_{\mathrm{KL}}(\pi_\theta^\star \| q) \le \log \mathbb{E}_q[U_\theta^2]$. A finite second moment does not imply a finite uniform envelope on the density ratio: a Gibbs tilt can assign appreciable mass to a target tail that is underweighted by the proposal, producing an unbounded yet square-integrable density ratio. Our analysis is designed for this square-integrable regime, which is what the blockwise construction of Section 4.1 exploits, whereas envelope-based analyses assume a uniform bound on the ratio.

A related line of work studies the nearly singular regime of ordinary importance sampling, where the required sample size can scale as $\exp(D_{\mathrm{KL}})$ (Chatterjee & Diaconis, 2018). That criterion describes a different source of difficulty, and neither condition implies the other; the two should be distinguished when interpreting coverage assumptions.

Theorem-level comparison. Three lines of work sit closest to Theorem 2, each differing along a specific axis.

Rényi-divergence confidence bounds. Metelli et al. (2018) obtain off-policy bounds in terms of an exponentiated Rényi-2 divergence, with polynomial dependence on $1/\delta$ from a Cantelli argument. Metelli et al. (2021) obtain logarithmic dependence under a transformed-ratio construction for a fixed target. Our blockwise construction makes each block contribution bounded and consequently permits Bernstein concentration under a finite second-moment assumption, leaving the weights themselves untouched. The resulting event is uniform over the preference family revealed after the pool has been sampled.

Confident off-policy evaluation. Kuzborskij et al. (2021) provide high-confidence value estimates and policy-selection guarantees for finite, prespecified policy families. Our guarantee concerns a serving policy supported on one realized candidate pool. The pool is fixed before the preference is revealed, and the supremum over a continuous preference space is controlled through a preference net followed by a Gibbs-perturbation argument (Section F).

Robust and bias-reduced estimators. Truncation and winsorization limit the influence of large importance weights (Ionides, 2008; Orenstein, 2022). Bias-reduced self-normalized importance sampling, abbreviated BR-SNIS, uses iterated sampling–importance resampling to reduce finite-sample bias (Cardoso et al., 2022). These methods address tail sensitivity or estimator bias for a target-specific estimate. Our objective is to construct one probability law on the realized pool that approximates every preference-indexed Gibbs target simultaneously, so the distinction lies in the combination of post-revelation sharing, candidate support, and a uniform guarantee based on a finite second moment. Section G.3 compares against both empirically.

Reward-agnostic data collection. Reward-free and reward-agnostic exploration collects trajectories before the reward function is specified, and coverage or concentrability conditions then determine which downstream rewards the collected data can support (Jin et al., 2020; Li et al., 2024;

Table 1: Comparison with nearby importance-sampling results. The columns indicate whether a method studies multiple targets, allows the target to be revealed after sampling, returns a policy supported on the realized pool, requires only a finite density-ratio second moment, and provides a uniform high-probability guarantee. Non-binary entries name what the referenced analysis provides instead: "MSE only" and "envelope" for the tail condition assumed, "optional" where candidate support is a modelling choice, "n/a" where no serving law is returned, "value only" where a scalar is certified, and "finite set" where the target family must be enumerated in advance. A cross mark records that the referenced analysis does not directly cover the corresponding setting; it does not preclude combining the underlying technique with an index-revealed target family.

Work	Multiple targets	Target revealed after sampling	Candidate support	Second moment	Uniform high-prob.
Classical SNIS (Agapiou et al., 2017)	✗	✗	✓	MSE only	✗
BR-SNIS (Cardoso et al., 2022)	✗	✗	✓	envelope	✓
Winsorized IS (Ionides, 2008; Orenstein, 2022)	✗	✗	optional	✓	✓
Multi-target selection (Roy & Evangelou, 2024)	✓	✗	n/a	✓	✗
Confident OPE (Kuzborskij et al., 2021)	finite set	✗	value only	✓	✓
POOLALIGN (ours)	✓	✓	✓	✓	✓

2023). Our setting has the same delayed-objective structure, but the samples are complete language-model responses rather than trajectories, the proposal determines coverage over responses, the target family consists of Gibbs reweightings indexed by user preferences, and the deployed policy is supported on the fixed pool. One structural consequence is that our coverage coefficient is a property of the *proposal* and is therefore certifiable before serving (Theorem 4), whereas concentrability in those works is a property of a fixed logged dataset. The result is a candidate-complexity problem for shared response pools, whose central object is a uniform post-revelation guarantee.

B FORMAL SETUP AND NOTATION

In this section, we restate Section 3 with the conventions used in the results and proofs. The three subsections below expand Sections 3.1, 3.2 and 3.4 in turn, the last of them also carrying the success criteria of Section 3.3. Policies and proposals on $\mathcal{Y}$ are written through their densities, and $\int \cdot \, \mathrm{d}y$ denotes integration over responses, which reads as a sum over $\mathcal{Y}$ when the response space is discrete. A ratio of two such densities is read as $p(y)/q(y) := +\infty$ when $p(y) > 0 = q(y)$ and as zero when both vanish, so that $\sup_y p(y)/q(y)$ and $\int p(y)^2/q(y) \, \mathrm{d}y$ are finite only when q covers p. We write δ_y for the distribution that places all of its probability on the response y.

Let $\mathcal{X}$ be a prompt space and let $\mathcal{Y}$ be the response space. A reference language model is a conditional distribution with density $\pi_0(y \mid x)$. We condition the candidate-sampling analysis on a fixed set of deployment prompts

$$\boldsymbol{x}_{1:n} = (x_1, \ldots, x_n). \tag{16}$$

We condition on these fixed prompts throughout. Extending the guarantee to a prompt population requires an additional generalization analysis.

B.1 PREFERENCE-CONDITIONED GIBBS POLICIES

Let Θ be a nonempty compact metric space with metric d_Θ, and let $\{s_\theta : \theta \in \Theta\}$ be a family of real-valued scores. The preference parameter θ is supplied by a user or an external controller. For a fixed prompt and preference, let $\beta > 0$ and suppose

$$0 < Z_\theta(x) := \int \exp(s_\theta(x, y)/\beta) \, \pi_0(y \mid x) \, \mathrm{d}y < \infty, \tag{17}$$

and define the scaled log-partition value $\psi_\theta(x) := \beta \log Z_\theta(x)$ and the Gibbs policy

$$\pi^\star_\theta(y \mid x) = \frac{\exp(s_\theta(x, y)/\beta)}{Z_\theta(x)} \pi_0(y \mid x). \tag{18}$$

To avoid an undefined $\infty - \infty$ when the expected score and relative entropy are both infinite, we formulate the KL-regularized problem in its Gibbs-relative extended-value form:

$$\sup_{\pi(\cdot|x)} \mathcal{J}_{\theta,x}(\pi), \qquad \mathcal{J}_{\theta,x}(\pi) := \psi_\theta(x) - \beta D_{\mathrm{KL}}(\pi(\cdot \mid x) \| \pi^\star_\theta(\cdot \mid x)) \,. \tag{19}$$

The supremum is over policies supported where $\pi_0(\cdot \mid x)$ is, and the divergence is $+\infty$ otherwise. Thus $\mathcal{J}_{\theta,x}$ is well defined under (17), is bounded above by $\psi_\theta(x)$, and is uniquely maximized by $\pi^\star_\theta(\cdot \mid x)$. Moreover, whenever $D_{\mathrm{KL}}(\pi(\cdot \mid x) \| \pi_0(\cdot \mid x)) < \infty$ and $s_\theta(x, \cdot)$ is $\pi(\cdot \mid x)$-integrable, a change of variables gives

$$\mathcal{J}_{\theta,x}(\pi) = \int s_\theta(x, y)\, \pi(y \mid x)\, \mathrm{d}y - \beta D_{\mathrm{KL}}(\pi(\cdot \mid x) \| \pi_0(\cdot \mid x)) \,. \tag{20}$$

Hence the extended-value formulation agrees with the familiar reward-minus-KL objective whenever its two terms are finite. This variational identity does not depend on how the score was learned. A Bradley–Terry reward model, a DPO-style implicit reward, a judge, or a calibrated vector of task metrics may all instantiate s_θ (Bradley & Terry, 1952; Rafailov et al., 2023).

The target policy is invariant to a promptwise additive shift $s_\theta(x, \cdot) \mapsto s_\theta(x, \cdot) + c_\theta(x)$, whereas the scaled log-partition shifts by $c_\theta(x)$. Thus value claims require a fixed, declared score gauge. For policy metrics, parameter covers should ideally be taken after quotienting out such redundant shifts; the score-space Lipschitz cover used below is always sufficient but can be conservative when different parameters induce the same target policy.

Two examples used throughout the paper are a singleton family $s_\theta = r$ and a continuous scalarization

$$s_\theta(x, y) = \theta^\top r(x, y), \qquad \theta \in \Delta_D. \tag{21}$$

Bounded sets of Lagrange multipliers can also be included, but optimizing such multipliers requires the direction dictated by the underlying constrained problem. We therefore treat every parameter in Θ only as an index of a target policy.

B.2 Shared Sampling before Preference Revelation

Before candidate generation, the system knows π_0, β, the score family $\{s_\theta\}_{\theta \in \Theta}$, and the proposal class, but not the particular active preference θ. It may use independent pilot data to choose a proposal. The proposal is then frozen, candidates are generated, and only afterward may an external user or controller reveal θ. Thus the proposal may depend on the declared score family, but not on the active preference revealed after candidate generation.

We distinguish two protocols. With a fixed proposal, whether supplied or analytically specified, the information order for each prompt is

$$(\pi_0, \beta, \{s_\theta\}, \Theta, q) \longrightarrow Y_{1:M} \longrightarrow \theta \longrightarrow \widehat{\pi}_{\theta,M}. \tag{22}$$

With certified proposal selection, an auxiliary pilot is used first:

$$(\pi_0, \beta, \{s_\theta\}, \Theta) \longrightarrow Y^{\mathrm{pilot}}_{1:H} \longrightarrow \widehat{q}\ \text{frozen} \longrightarrow Y^{\mathrm{serv}}_{1:M} \longrightarrow \theta \longrightarrow \widehat{\pi}_{\theta,M}. \tag{23}$$

The pilot responses are not part of the final policy support in the present analysis. In both protocols, "shared pool" means the fresh M-response pool for one fixed prompt is shared across preferences; it does not mean responses are shared across different prompts.

Our main probability statement is simultaneous: after one pool is generated, it must support every declared preference, including one selected adaptively after seeing the candidate pool. This models repeated service or a post-generation controller and leads to a $\mathbb{P}(\sup_{\theta \in \Theta} \mathrm{error}_\theta \le \epsilon)$ guarantee. If only one θ is drawn independently of the candidates, the weaker pointwise requirement $\sup_\theta \mathbb{P}(\mathrm{error}_\theta > \epsilon) \le \delta$ does not require the same preference-family union bound. These two operational quantifiers have different logarithmic costs and are not interchanged below.

Operational reuse and cost. For a fixed proposal, M counts generated responses per prompt. If Q service requests arrive after the pool is fixed, the same M responses support all Q revealed preferences, and the uniform event covers even an adaptive preference sequence without an additional factor depending on Q. Serving still incurs score and likelihood evaluation, QM preference-dependent weight computations, and sampling from the reweighted policy. In the scalarization

model $s_\theta = \theta^\top r$, the vector $r(x, Y_i)$ can be stored once and only its scalarization changes with θ. For the certified protocol, $H + M$—not M alone—is the total number of generated responses per prompt; all candidate-complexity results stated in terms of M concern the fresh serving pool.

Concretely, the asymmetry between the two costs is what makes the protocol worth its statistical price. Regenerating a preference-specific pool costs QM autoregressive decodes, each producing up to L tokens; sharing one pool costs M decodes plus QM weight updates, each of which is a stored-vector dot product and an exponential. At the scale of the study in Section G.3 ($M = 256$, $L \le 96$ new tokens, a scalarization $s_\theta = \theta^\top r$ with $r \in \mathbb{R}^2$ cached per candidate), the shared pool replaces $Q \times 256$ decodes of up to 96 tokens each by 256 such decodes plus $Q \times 256$ two-dimensional dot products. The ratio therefore grows linearly in the number of preferences served per prompt, and the entire statistical cost of the protocol is the replacement of a per-target second moment by Γ_2. Conversely, when $Q = 1$ there is nothing to amortize and the protocol reduces to ordinary test-time alignment.

The worst-case guarantee over Θ. Equation (4) places $\sup_{\theta \in \Theta}$ inside the probability, so the guarantee covers every preference that can arrive after the pool is seen, including one chosen adversarially. This is the form a serving system can act on: one realized pool is certified once and then reused for an arbitrary, possibly adaptive, sequence of requests, with no bound re-derived per request. The cost enters in a single place, the $\log N_\rho$ factor of Eq. (13), which equals $\log |\Theta|$ for a finite family and $d_{\mathrm{par}} \log(1 + R_\Theta/\rho)$ under Eq. (11). An average-case guarantee over a known preference distribution replaces that factor by a smaller one and Γ_2 by an averaged coverage coefficient, with the blockwise construction unchanged and only the union-bound step of Section F differing.

For each x_j, let $q_j = q(\cdot \mid x_j)$ be a proposal with $q_j(y) > 0$ whenever $\pi_0(y \mid x_j) > 0$, and let $R_j(y) := \pi_0(y \mid x_j)/q_j(y)$ be the likelihood ratio of Section 3.2, which we require to be evaluable and which the choice $q_j = \pi_0(\cdot \mid x_j)$ makes identically one. We take $R_j \ge 0$ everywhere, and $R_j(y) = 0$ wherever $\pi_0(y \mid x_j) = 0$. For every prompt, draw

$$Y_{j,1}, \dots, Y_{j,M} \overset{\text{iid}}{\sim} q_j, \qquad j \in [n], \tag{24}$$

independently across prompts. Define

$$A_\theta(x_j, y) := R_j(y) \exp(s_\theta(x_j, y)/\beta)\,, \tag{25}$$

$$\widehat{Z}_{\theta,M}(x_j) := \frac{1}{M} \sum_{i=1}^{M} A_\theta(x_j, Y_{j,i}), \qquad \widehat{\psi}_{\theta,M}(x_j) := \beta \log \widehat{Z}_{\theta,M}(x_j). \tag{26}$$

Here and below, $\log 0 := -\infty$. When the denominator is positive, the finite candidate pool induces the self-normalized empirical policy $\widehat{\pi}_{\theta,M}(\cdot \mid x_j)$, which is the one-block ($K = 1$, $b = M$) case of the blockwise policy of Eq. (7) and serves as the comparator throughout:

$$\widehat{\pi}_{\theta,M}(\cdot \mid x_j) = \sum_{i=1}^{M} \frac{A_\theta(x_j, Y_{j,i})}{\sum_{k=1}^{M} A_\theta(x_j, Y_{j,k})}\, \delta_{Y_{j,i}}. \tag{27}$$

If the denominator is zero, define $\widehat{\pi}_{\theta,M}$ to be the uniform distribution over the realized candidates. This fallback is independent of θ and makes the policy well defined in every case; the high-probability events below make it irrelevant to the stated bounds. Operationally, serving from (27) means sampling a candidate index with the displayed probabilities and returning its response. The policy bounds control conditional expectations under this randomized serving rule, not the realized metric of one draw or an argmax. Thus duplicate responses contribute the sum of their candidate-index weights. The same realized responses can be reweighted for every requested $\theta \in \Theta$. Importance reweighting and sampling-importance-resampling appear in controlled decoding (Ji et al., 2024); related test-time samplers target the same reward-tilted law (Faria & Smith, 2025). Our question is which acquisition laws make a finite pool uniformly reliable for a preference family and how the required budget depends on target overlap.

For analysis, normalize the raw weight by the unknown partition function:

$$U_\theta(x_j, y) := \frac{A_\theta(x_j, y)}{Z_\theta(x_j)} = \frac{\pi^\star_\theta(y \mid x_j)}{q_j(y)}. \tag{28}$$

Under $y \sim q_j$, $\mathbb{E} U_\theta(x_j, y) = 1$ and

$$\mathbb{E}_{q_j} U_\theta(x_j, Y)^2 = 1 + \chi^2(\pi^\star_\theta(\cdot \mid x_j) \| q_j)\,. \tag{29}$$

B.3 Candidate Coverage

The two coverage quantities of Eqs. (5) and (6) are, written out,

$$\Gamma_2(q, \Theta; \boldsymbol{x}_{1:n}) := \max_{j\in[n]} \sup_{\theta\in\Theta} \mathbb{E}_{Y\sim q_j} U_\theta(x_j, Y)^2, \tag{30}$$

$$\Gamma_\infty(q, \Theta; \boldsymbol{x}_{1:n}) := \max_{j\in[n]} \sup_{\theta\in\Theta} \sup_{y:\, q_j(y)>0} U_\theta(x_j, y). \tag{31}$$

Thus Γ_2 is a worst-prompt, worst-preference second moment, while Γ_∞ is a stronger envelope that permits a direct one-pool Bernstein analysis. The pointwise maximum is stronger than an average over prompts, but it matches the simultaneous per-prompt guarantee proved below. Our basic SNIS theorem uses both quantities. The robust block construction in Sections C.1 and D.2 instead assumes only a known upper bound on Γ_2 and allows $\Gamma_\infty = \infty$.

For a fixed proposal, Γ_2 quantifies worst-target proposal mismatch. The incremental cost of sharing is obtained by comparing it with a target-specific or unrestricted-design benchmark. The finite-space oracle results in Propositions 27 and 28 show that the optimal coefficient is one for identical targets and N for exactly disjoint targets; within the full-support Gibbs family, the latter value is approached as modes separate. The design objective therefore exploits target overlap while quantifying coverage under the admissible proposal class.

For an admissible proposal class $\mathcal{Q}$, the acquisition problem is summarized by the instance-dependent coefficient

$$\Gamma_2^\star(\mathcal{Q}, \Theta; \boldsymbol{x}_{1:n}) := \inf_{q\in\mathcal{Q}} \Gamma_2(q, \Theta; \boldsymbol{x}_{1:n}). \tag{32}$$

The unrestricted infimum is an oracle property of the target family. Restricting $\mathcal{Q}$ to sampleable, known-density language-model mixtures yields the implementable design target.

We measure prompt-averaged values on the fixed set by

$$\Psi_n(\theta) = \frac{1}{n}\sum_{j=1}^n \psi_\theta(x_j), \qquad \widehat{\Psi}_{n,M}(\theta) = \frac{1}{n}\sum_{j=1}^n \widehat{\psi}_{\theta,M}(x_j). \tag{33}$$

The estimator $\widehat{\psi}_{\theta,M}$ targets the optimum $\psi_\theta(x) = \sup_\pi \mathcal{J}_{\theta,x}(\pi)$ of (19) over all policies, which differs from the KL-regularized objective attained by a candidate-supported policy. For policy evaluation, let $\mathcal{F}$ be a nonempty finite set of functions $f : \mathcal{X}\times\mathcal{Y} \to [-1, 1]$. Some complexity restriction on $\mathcal{F}$ is essential; we use a finite class here. Without such a restriction, taking all bounded tests would recover total variation, which a serving policy supported on finitely many candidates cannot control against a target spread over far more responses. In contrast, expectations of a prespecified collection of rewards, safety scores, and task metrics can be approximated.

For a generic candidate-supported serving policy $\widehat{\pi}^{\mathrm{serv}}_{j,\theta}$ and value estimate $\widetilde{\psi}_{j,\theta}$, the two success criteria of Eqs. (2) and (3) read

$$\mathrm{Err}_{\mathcal{F}} := \max_{j\in[n]} \sup_{\theta\in\Theta} \max_{f\in\mathcal{F}} \left| \mathbb{E}_{\widehat{\pi}^{\mathrm{serv}}_{j,\theta}} f(x_j, Y) - \mathbb{E}_{\pi^\star_\theta(\cdot|x_j)} f(x_j, Y) \right|, \tag{34}$$

$$\mathrm{Err}_{\psi} := \max_{j\in[n]} \sup_{\theta\in\Theta} |\widetilde{\psi}_{j,\theta} - \psi_\theta(x_j)|. \tag{35}$$

The headline policy guarantee is the integral probability metric induced by $\mathcal{F}$, uniformly controlled by a candidate-supported randomized policy.

C Full Estimator and Certified Acquisition

In this section, we present the exact block construction, proposal family, pilot selector, and executable pseudocode summarized in Section 4; the two estimators are those of Sections 4.1 and 4.2.

C.1 Robust Block Aggregation

The ordinary policy in (27) is a useful baseline, but a high-probability analysis of that single ratio uses the envelope Γ_∞. To retain only the second-moment condition, partition each prompt's candidate indices, independently of the realized evaluation responses, into K disjoint, balanced blocks

$\mathcal{B}_{j,1}, \ldots, \mathcal{B}_{j,K}$. Let $b = \min_k |\mathcal{B}_{j,k}| = \lfloor M/K \rfloor$. For a revealed preference θ, define

$$\widehat{\pi}^{(k)}_{j,\theta} := \frac{\sum_{i \in \mathcal{B}_{j,k}} A_\theta(x_j, Y_{j,i}) \delta_{Y_{j,i}}}{\sum_{i \in \mathcal{B}_{j,k}} A_\theta(x_j, Y_{j,i})}, \qquad \widehat{\pi}^{\mathrm{blk}}_{j,\theta} := \frac{1}{K} \sum_{k=1}^{K} \widehat{\pi}^{(k)}_{j,\theta}. \tag{36}$$

If a block denominator is zero, its policy is defined to be the uniform distribution on that block. This fallback is fixed independently of θ and keeps the output candidate-supported. The same zero-weight blocks occur for every θ, because the exponential score factor is strictly positive and the remaining base likelihood ratio is common to the family.

The block policies alone cannot identify Z_θ: multiplying all raw weights by a common constant leaves every normalized block unchanged. We therefore estimate the partition function from the same candidates but with a second balanced deterministic partition. Let $\mathcal{B}^Z_{j,1}, \ldots, \mathcal{B}^Z_{j,K_Z}$ be value blocks, where K_Z is odd, and set

$$\widetilde{Z}^{\mathrm{MoM}}_{j,\theta} := \mathrm{median}_{r \in [K_Z]} \left\{ \frac{1}{|\mathcal{B}^Z_{j,r}|} \sum_{i \in \mathcal{B}^Z_{j,r}} A_\theta(x_j, Y_{j,i}) \right\}, \qquad \widetilde{\psi}^{\mathrm{MoM}}_{j,\theta} := \beta \log \widetilde{Z}^{\mathrm{MoM}}_{j,\theta}. \tag{37}$$

The policy and value partitions may reuse the same candidates; no estimate is used to set another estimate's weights. Their events are combined by a union bound rather than an independence claim.

The robust construction arranges classical SNIS and median-of-means (Jerrum et al., 1986; Alon et al., 1999; Catoni, 2012; Minsker, 2015) into one policy-valued interface for preferences revealed after candidate generation. Its confidence-dependent block count differs from the usual $K = O(\log(1/\delta))$. The policy bias, square-root fluctuation, and bounded block-level deviation balance at

$$b \asymp \sqrt{\frac{\Gamma M}{\mathfrak{L}_{\mathrm{P}}}}, \qquad K \asymp \sqrt{\frac{M \mathfrak{L}_{\mathrm{P}}}{\Gamma}}, \qquad K_Z \asymp \mathfrak{L}_{\mathrm{V}}, \tag{38}$$

where $\mathfrak{L}_{\mathrm{P}}(\rho, \delta)$ and $\mathfrak{L}_{\mathrm{V}}(\rho, \delta)$ are the policy and value log factors defined in Eqs. (52) and (53), written without their arguments when these are clear. Finite-sample conditions and the resulting envelope-free rate are given in Theorem 10 and Corollary 11.

Certified execution. Algorithm 1 gives the fixed-proposal core. For a finite-sample certificate, the evaluation class, confidence level, coverage bound, preference net (when needed), block counts, and index partitions must all be fixed independently of the serving candidates. One deterministic implementation enumerates the feasible integers $K \in [M]$ and minimizes the certified radius r_{blk} in (54), with a fixed tie-break. It chooses an odd K_Z satisfying (56). If either choice is infeasible, POOLALIGN remains executable, but the corresponding guarantee is not reported. Empirical ESS and weight-tail summaries are diagnostics rather than coverage certificates.

C.2 KNOWN-DENSITY PROPOSAL FAMILIES

Let $q_0(\cdot \mid x) = \pi_0(\cdot \mid x)$ and let $q_1, \ldots, q_J$ be proposal components whose sequence probabilities can be evaluated on every sampled response. We use the mixture

$$q_\alpha(\cdot \mid x) = \sum_{k=0}^{J} \alpha_k q_k(\cdot \mid x), \qquad \alpha \in \Delta_{J+1}, \quad \alpha_0 \geq a, \tag{39}$$

where $0 < a \leq 1$ is a defensive reference mass (Hesterberg, 1995; Owen & Zhou, 2000). Sampling first chooses a component and then draws a response, but the importance weight must use the *marginal* mixture density in (39), not only the density of the sampled component (Veach & Guibas, 1995).

The defensive constraint gives $q_\alpha \geq a\pi_0$. Suppose that, for each evaluation prompt, there is one common $\pi_0(\cdot \mid x_j)$-support S_j on which every score is finite and

$$\max_{j \in [n]} \sup_{\theta \in \Theta} \left\{ \sup_{y \in S_j} s_\theta(x_j, y) - \inf_{y \in S_j} s_\theta(x_j, y) \right\} \leq B, \tag{40}$$

then Proposition 30 shows that $\Gamma_\infty \leq a^{-1} e^{B/\beta}$. This yields a conservative, computable certificate. In an implementation, bounding B may require a declared reward normalization or clipping rule; the resulting target is then the Gibbs policy for that declared score, not for an unbounded latent reward.

C.3 Preference Anchors for Proposal Design

Anchors discretize the preference family for proposal design and certification; the selection rule is specified in Section C.4.

Choose a nonempty finite design set

$$\varnothing \neq \Theta_N = \{\theta_1, \dots, \theta_N\} \subset \Theta. \tag{41}$$

The candidate pool may subsequently be reweighted at an unseen θ. The anchors' geometric resolution is summarized by

$$h_N(\Theta_N) = \sup_{\theta \in \Theta} \min_{\ell \in [N]} d_\Theta(\theta, \theta_\ell). \tag{42}$$

Finite-space variational oracles clarify how target overlap changes the best possible coverage, but they require normalized target densities and generally are not tractable language-model samplers. Their square-root forms, overlap extremes, and full variational statements are deferred to Section E.5. The executable method instead searches the declared mixture catalogue below.

C.4 Split-Sample Proposal Selection

The certified variant uses the same fixed prompt values $x_{1:n}$ for proposal selection and evaluation, but independent response samples. Conditioning on the prompt values does not reuse candidate randomness. A design learned on disjoint prompt values transfers to the worst prompt under additional prompt regularity, which we do not assume here. Let $q_{\mathrm{pil}} = \sum_{k=0}^{J} \zeta_k q_k$ be a known-density defensive pilot proposal with $\zeta \in \Delta_{J+1}$ and $\zeta_0 \geq a_{\mathrm{pilot}} > 0$. For each fixed prompt x_j, we abbreviate the densities involved as

$$\begin{aligned} p_{0j}(y) &:= \pi_0(y \mid x_j), \qquad p_{j\ell}(y) := \pi^\star_{\theta_\ell}(y \mid x_j), \\ q_{\mathrm{pil},j}(y) &:= q_{\mathrm{pil}}(y \mid x_j), \quad q_{\alpha,j}(y) := q_\alpha(y \mid x_j). \end{aligned} \tag{43}$$

The proposal components, these densities, and the pilot proposal are fixed independently of the pilot responses (though they may depend on the fixed prompts or external data). Before observing pilot responses, fix a finite catalogue

$$\varnothing \neq \mathcal{A} \subset \{\alpha \in \Delta_{J+1} : \alpha_0 \geq a\}. \tag{44}$$

Draw $\widetilde{Y}_{j,1:H} \overset{\mathrm{iid}}{\sim} q_{\mathrm{pil}}(\cdot \mid x_j)$ independently of the serving pool. Split each prompt's pilot responses into two nonempty folds. For an anchor θ_ℓ, one fold estimates

$$\widehat{Z}^{(1)}_{j\ell} = \frac{1}{|I_{j,1}|} \sum_{i \in I_{j,1}} \frac{p_{0j}(\widetilde{Y}_{j,i})}{q_{\mathrm{pil},j}(\widetilde{Y}_{j,i})} \exp\Big(s_{\theta_\ell}(x_j, \widetilde{Y}_{j,i})/\beta\Big). \tag{45}$$

The other fold forms a plug-in estimate of

$$G_{j\ell}(\alpha) = \int \frac{p_{j\ell}(y)^2}{q_{\alpha,j}(y)} \, \mathrm{d}y, \tag{46}$$

through

$$\widehat{G}^{(2|1)}_{j\ell}(\alpha) = \frac{1}{|I_{j,2}|(\widehat{Z}^{(1)}_{j\ell})^2} \sum_{i \in I_{j,2}} \frac{p_{0j}(\widetilde{Y}_{j,i})^2 \exp(2 s_{\theta_\ell}(x_j, \widetilde{Y}_{j,i})/\beta)}{q_{\mathrm{pil},j}(\widetilde{Y}_{j,i}) q_{\alpha,j}(\widetilde{Y}_{j,i})}. \tag{47}$$

Let $\widehat{G}_{j\ell}$ denote the average of $\widehat{G}^{(2|1)}_{j\ell}$ and its fold-swapped analogue. This cross-fitting avoids using the same observation in its normalizer and second-moment term, although the integrand is defined as zero whenever its squared numerator is zero. A ratio with zero cross-fitted partition estimate is defined as $+\infty$. We select mixture weights by the robust finite-catalogue criterion

$$\widehat{\alpha} \in \arg\min_{\alpha \in \mathcal{A}} \max_{j \in [n], \ell \in [N]} \widehat{G}_{j\ell}(\alpha). \tag{48}$$

Because $\mathcal{A}$ is finite, we fix an order on it in advance and choose the first minimizer in that order, making the selector unambiguous. Under the declared score-oscillation and defensive-mixture conditions, Theorem 13 turns this statistic into a simultaneous multiplicative coverage UCB and gives

Algorithm 2 Optional certified proposal selection for POOLALIGN

Require: The preference-independent inputs of Algorithm 1, excluding the frozen proposals $q_{1:n}$ and block counts K, K_Z, with the score family (including any calibration) frozen independently of pilot and serving responses; evaluation class $\mathcal{F}$; anchors Θ_N; known-density components $q_{0:J}$ with $q_0 = \pi_0$; a predeclared finite catalogue $\mathcal{A}$; known-density pilot law q_{pil}; pilot size $H \geq 2$; defensive masses (a_{pilot}, a); score-oscillation bound B; pilot tolerances (η_Z, η_G); pilot and serving confidences $(\delta_{\mathrm{p}}, \delta_{\mathrm{e}})$; and, for continuous-family certification, L_s, ρ and a fixed ρ-net of size N_ρ.

1: Draw H independent pilot responses from $q_{\mathrm{pil}}(\cdot \mid x_j)$ at every fixed prompt and split them by a response-independent index partition.
2: Cross-fit anchor normalizers and coverage objectives using (45) and (47).
3: **if** (65) holds and all cross-fitted normalizers are positive **then**
4: Solve (48), set $\widehat{q} = q_{\widehat{\alpha}}$, and form the finite-anchor or continuous coverage certificate in (69)–(70).
5: **else**
6: Declare pilot-design failure, set $\widehat{\alpha} = (1, 0, \ldots, 0)$, and use the conservative certificate $\widehat{\Gamma} = e^{B/\beta}$ for $\widehat{q} = \pi_0$.
7: **end if**
8: Freeze $\widehat{q}$, tune K, K_Z using $\widehat{\Gamma}$ and δ_{e}, and run Algorithm 1 with a fresh, independent M-response serving pool.
9: Report total generation cost $H+M$ per fixed prompt and total failure probability at most $\delta_{\mathrm{p}}+\delta_{\mathrm{e}}$.

an oracle factor relative to the best member of $\mathcal{A}$. The UCB multiplier is common to the catalogue, so (48) also minimizes the worst UCB. The theorem does not cover optimization over the entire continuous simplex; that extension requires a separate uniform empirical process argument. After choosing $\widehat{\alpha}$, we freeze $\widehat{q} = q_{\widehat{\alpha}}$ and draw a fresh serving pool. Conditional on a valid pilot event, the serving responses are i.i.d. from a fixed proposal, and Corollary 14 composes the pilot and serving failure probabilities.

The pilot is an auxiliary design sample, not part of the returned policy's support. Consequently, the fixed-proposal theory uses one M-response serving pool per prompt, while certified selection uses $H + M$ generated responses per prompt: H pilot responses and M fresh serving responses. Reusing pilot responses in the final policy would require an adaptive or multiple-importance-sampling analysis not provided here. Pilot cost may be amortized across preferences served from the same realized prompt-specific pool, but not across unseen prompts under the present theorem.

C.5 Optional Certified Wrapper and Operational Boundaries

The frozen-proposal path in Algorithm 1 is the core algorithm and uses exactly one M-response serving pool per prompt. Algorithm 2 is an optional design wrapper; its pilot responses never enter the returned policy support.

The robust certificate needs a valid upper bound on Γ_2 to tune K; an uncorrected empirical second moment or ESS is not such a bound. The pilot UCB is valid only under the finite-catalogue, same-prompt, bounded-oscillation conditions in Theorem 13. The ordinary one-block certificate additionally needs Γ_∞. A fully deterministic but conservative fallback uses $\Gamma_2 \leq \Gamma_\infty \leq a^{-1}e^{B/\beta}$. Empirical ESS remains useful for failure diagnosis but is not a distribution-free coverage certificate. For a continuous-family certificate, the report must additionally declare d_Θ, L_s, ρ, and N_ρ; otherwise the certificate is restricted to a prespecified finite preference set and finite $\mathcal{F}$. Repeated response strings retain the sum of their candidate-index probabilities and are not deduplicated before normalization.

Likelihood-access boundary. If candidates are sampled directly from a black-box reference model, setting $q = \pi_0$ is valid because the reference-to-proposal weight equals one. A mixture of black-box generators with unavailable sequence probabilities does not support the correction π_0/q and falls outside the present theory. All mixture components must use a common response representation, including the tokenizer, EOS convention, and maximum length. Rejection retries, hidden

Result	Additional requirement	Certified object
One-block baseline	finite Γ_∞ envelope	uniform log-partition and bounded-functional error
Robust block law (main)	known finite Γ_2 bound	the same errors, without a bounded density ratio
Pilot oracle	bounded score oscillation; finite known-density catalogue	coverage UCB and fresh-pool end-to-end certificate
Fixed-proposal lower bound	paired finite target–function family	necessary M for simultaneous bounded-functional accuracy

Table 2: Assumptions and certified quantities for the principal results.

safety filters, and deduplication alter the sampling law unless they are included in the evaluated proposal density.

D Main Results and Design Limits

In this section, we present the formal statements of our main results, i.e., the robust upper bound (Theorem 10 and Corollary 11), the pilot certificate (Theorem 13 and Corollary 14), and the fixed-proposal lower bound (Proposition 15 and Theorem 16). These are the formal versions of Theorems 2, 4 and 5, and the one-block baselines they are compared against in Section 5.1 are stated in Section E. All results condition on the fixed prompts. Probabilities are over the pilot and/or serving candidate-pool randomness (and any explicitly stated policy randomization), and policy bounds concern conditional expectations under the candidate-supported serving policies shared across requested preferences. Evaluation classes, nets, proposals, and partitions are fixed before serving; pilot-selected objects are treated conditionally. Table 2 summarizes what each result assumes and what it certifies.

D.1 Assumptions for the Main Results

The main results use three conditions, each standard in the importance-sampling and test-time-alignment literature: a support condition, under which the targets are representable from the shared proposal; a second-moment (χ^2) coverage condition, the canonical measure of importance-sampling difficulty (Agapiou et al., 2017; Metelli et al., 2018); and a Lipschitz-in-preference regularity condition, under which a finite net controls the continuous family. Only the first two enter the fixed-proposal upper bound; the third is used solely to pass from a preference net to all of Θ.

Assumption 6 (Support and sampling)**.** *For every $j \in [n]$ and $\theta \in \Theta$, $0 < Z_\theta(x_j) < \infty$ and $q_j(y) > 0$ whenever $\pi_0(y \mid x_j) > 0$, so the proposal covers the reference policy and every target $\pi^\star_\theta(\cdot \mid x_j)$. Conditional on $\boldsymbol{x}_{1:n}$, the variables in* (24) *are independent and have the stated proposal laws.*

Support coverage is the minimal requirement for the target to be representable from the proposal at all, and holds automatically for the defensive mixtures of Section C.2, whose reference component keeps $q_j \geq a\pi_0(\cdot \mid x_j)$.

Assumption 7 (Uniform second-moment coverage)**.** *There is a known or analytically bounded constant $1 \leq \Gamma < \infty$ such that, for every $j \in [n]$ and $\theta \in \Theta$,*

$$\mathbb{E}_{q_j} U_\theta(x_j, Y)^2 \leq \Gamma. \tag{49}$$

The left-hand side is $1 + \chi^2(\pi^\star_\theta(\cdot \mid x_j) \| q_j)$, so Assumption 7 is equivalently the uniform bound $\chi^2(\pi^\star_\theta(\cdot \mid x_j) \| q_j) \leq \Gamma - 1$ on the target-to-proposal divergence. Controlling this second moment is the standard sufficient condition under which self-normalized importance sampling is analyzed (Agapiou et al., 2017; Metelli et al., 2018); it is implied by, but strictly weaker than, the uniform weight envelope of Assumption 17, and it is the only weight-tail condition the main block theorem requires.

Assumption 8 (Preference regularity)**.** *For a constant $L_s < \infty$,*

$$|s_\theta(x_j, y) - s_{\theta'}(x_j, y)| \leq L_s d_\Theta(\theta, \theta') \tag{50}$$

for all j, all $\theta, \theta' \in \Theta$, on one common set $\{y : q_j(y) > 0\}$. The covering number $\mathcal{N}(\Theta, d_\Theta, \rho)$ is finite for every $\rho > 0$.

Linear scalarizations of bounded reward coordinates, the standard multi-objective parameterization used in our experiments, satisfy (50) with L_s the reward range; the finite covering number then holds for any compact Θ (Vershynin, 2018). Only nearby preferences need nearby scores, which is why a ρ-net of Θ suffices in the proofs.

Let $N_\rho = \mathcal{N}(\Theta, d_\Theta, \rho)$. The stronger envelope assumption and the associated one-block deviation radius are deferred to Section E; neither is used by the main block theorem.

D.2 Envelope-Free Uniform Block Guarantee

The ordinary one-block policy is simple, but its high-probability radius also depends on a uniform importance-weight envelope. We now analyze the block policy and MoM value estimator from Section C.1, using Assumption 7 as the only weight-tail condition. The theorem also states the sampling conditions and, for continuous families, the required score regularity. The first lemma records the block-level bias and mean-square control that makes policy aggregation possible. Averaging over blocks does not reduce the bias, which is governed by the block size alone; what aggregation buys is variance reduction across independent blocks whose summands are bounded regardless of the weight tail. The lemma is a constant-explicit variant of the classical SNIS bias and MSE bounds of Agapiou et al. (2017).

Lemma 9 (One-block SNIS moments)**.** *Fix a prompt, a target $\pi_\theta^\star$, a proposal q, and b i.i.d. draws from q. Suppose q covers $\pi_\theta^\star$, set $U(y) = \pi_\theta^\star(y)/q(y) \geq 0$, so that $\mathbb{E}_q U = 1$, and suppose $\mathbb{E}_q U^2 \leq \Gamma$. Let P_b be the normalized weighted empirical policy. When all b weights vanish, use a fixed, block-local, target-independent fallback rule returning a candidate-supported distribution, such as the uniform distribution on that block. Then, for every $f : \mathcal{Y} \to [-1, 1]$,*

$$\left|\mathbb{E}[P_b f] - \mathbb{E}_{\pi_\theta^\star} f\right| \leq \frac{16\Gamma}{b}, \qquad \mathbb{E}\left[(P_b f - \mathbb{E}_{\pi_\theta^\star} f)^2\right] \leq \frac{32\Gamma}{b}. \tag{51}$$

For the balanced policy blocks in (36) and the balanced MoM value blocks in (37), let

$$\mathfrak{L}_{\mathrm{P}}(\rho, \delta) := \log \frac{4nN_\rho(|\mathcal{F}| + 1)}{\delta}, \tag{52}$$

$$\mathfrak{L}_{\mathrm{V}}(\rho, \delta) := \log \frac{4nN_\rho}{\delta}, \tag{53}$$

$$r_{\mathrm{blk}} := \frac{16\Gamma}{b} + 8\sqrt{\frac{\Gamma \mathfrak{L}_{\mathrm{P}}(\rho, \delta)}{bK}} + \frac{4\mathfrak{L}_{\mathrm{P}}(\rho, \delta)}{3K}, \tag{54}$$

$$\kappa_Z := 2\sqrt{\frac{\Gamma}{b_Z}}, \qquad b_Z := \min_{j \in [n],\, r \in [K_Z]} |\mathcal{B}_{j,r}^Z|. \tag{55}$$

Balanced partitions have prompt-independent block sizes, so b and b_Z do not depend on j.

Combining the one-block moments with Bernstein concentration across policy blocks, MoM concentration for raw weights, and a preference net gives the main uniform deviation bound.

Theorem 10 (Envelope-free uniform block guarantee)**.** *Suppose Assumptions 6 and 7 hold and $\mathcal{F}$ is a nonempty finite class of $[-1, 1]$-valued functions. For a continuous preference family, additionally suppose Assumption 8. Fix the block counts and the candidate-index partitions independently of the serving candidates, and use the same balanced, nonempty deterministic block partitions for all θ, with $1 \leq K \leq M$ and odd $1 \leq K_Z \leq M$, and suppose*

$$K_Z \geq 8\mathfrak{L}_{\mathrm{V}}(\rho, \delta), \qquad \kappa_Z \leq \frac{1}{2}. \tag{56}$$

Then, under the conditional law of the serving candidate pool given $\boldsymbol{x}_{1:n}$, with probability at least $1 - \delta$, all MoM estimates are positive and simultaneously

$$\max_{j \in [n]} \sup_{\theta \in \Theta} |\widetilde{\psi}_{j,\theta}^{\mathrm{MoM}} - \psi_\theta(x_j)| \leq 2\beta\kappa_Z + 2L_s\rho, \tag{57}$$

$$\max_{j \in [n]} \sup_{\theta \in \Theta} \max_{f \in \mathcal{F}} \left|\mathbb{E}_{\widehat{\pi}_{j,\theta}^{\mathrm{blk}}} f(x_j, Y) - \mathbb{E}_{\pi_\theta^\star(\cdot|x_j)} f(x_j, Y)\right| \leq r_{\mathrm{blk}} + 2\left(e^{2L_s\rho/\beta} - 1\right). \tag{58}$$

No finite value of Γ_∞ is required. For a finite preference set Θ_N, Assumption 8 is unnecessary: replace N_ρ by N and omit both extension terms.

The three terms in (54) have different origins: one-block SNIS bias, fluctuation across independent block policies, and the Bernstein envelope at the block-policy level. Balancing these terms determines the block count and yields the candidate-complexity form below.

Corollary 11 (Envelope-free candidate complexity)**.** *Suppose Assumptions 6 to 8 hold, $\mathcal{F}$ is a nonempty finite class of $[-1,1]$-valued functions, $L_s > 0$, and there are $R_\Theta > 0$ and $d_{\mathrm{par}} \geq 0$ such that, for every $\rho > 0$,*

$$\log \mathcal{N}(\Theta, d_\Theta, \rho) \leq d_{\mathrm{par}} \log(1 + R_\Theta/\rho). \tag{59}$$

Choose balanced blocks with

$$K \asymp \sqrt{\frac{M \mathfrak{L}_{\mathrm{P}}(\rho, \delta)}{\Gamma}}, \qquad K_Z \asymp \mathfrak{L}_{\mathrm{V}}(\rho, \delta). \tag{60}$$

For sufficiently large universal constants in the rounded block counts and in the following bounds, each choice is feasible under its corresponding complexity requirement and K_Z satisfies (56)*. Only the K_Z choice is needed for the value conclusion. Uniform value error at most ϵ_ψ for $0 < \epsilon_\psi \leq 2\beta$ follows from*

$$M \geq C \frac{\beta^2 \Gamma}{\epsilon_\psi^2} \left\{ d_{\mathrm{par}} \log\left(1 + \frac{4R_\Theta L_s}{\epsilon_\psi}\right) + \log \frac{4n}{\delta} \right\}, \tag{61}$$

where C is universal. Uniform bounded-functional error at most ϵ_F for $0 < \epsilon_F \leq 1$ follows from

$$M \geq C \frac{\Gamma}{\epsilon_F^2} \left\{ d_{\mathrm{par}} \log\left(1 + \frac{2R_\Theta L_s}{\beta \log(1 + \epsilon_F/4)}\right) + \log \frac{4n(|\mathcal{F}| + 1)}{\delta} \right\}. \tag{62}$$

If both displayed requirements hold at confidence $\delta/2$, then both conclusions hold simultaneously on one candidate pool with probability at least $1 - \delta$. The two conclusions use different net radii, so this invokes Theorem 10 twice, once at $(\rho_\psi, \delta/2)$ with $\rho_\psi = \epsilon_\psi/(4L_s)$ and once at $(\rho_F, \delta/2)$ with $\rho_F = \frac{\beta}{2L_s} \log(1 + \epsilon_F/4)$, and combines the two events by a union bound. The value and functional partitions may reuse the same candidates, since K_Z enters only the value conclusion and K only the functional one.

Thus blockwise normalization replaces the envelope-dependent term by a pure second-moment law. Tuning uses a valid deterministic upper bound on Γ, supplied, for example, by the pilot certificate below.

D.3 Certified Proposal Selection from Preference Anchors

We first transfer coverage from a finite anchor net to the full preference family, then certify the anchor risks from independent pilot data.

Proposition 12 (Anchor-to-family coverage transfer)**.** *Suppose Assumption 8 holds and every $\theta \in \Theta$ has an anchor θ_ℓ with $d_\Theta(\theta, \theta_\ell) \leq \rho$. For every prompt and proposal q_j,*

$$\int \left(\frac{\pi_\theta^\star(y \mid x_j)}{q_j(y)}\right)^2 q_j(y)\, \mathrm{d}y \leq e^{4L_s\rho/\beta} \int \left(\frac{\pi_{\theta_\ell}^\star(y \mid x_j)}{q_j(y)}\right)^2 q_j(y)\, \mathrm{d}y. \tag{63}$$

Each integral is $+\infty$ when q_j fails to cover the corresponding target. Thus a robust finite-anchor design controls the continuous-family second moment up to $e^{4L_s\rho/\beta}$ at the fixed prompts.

The next theorem converts the cross-fitted risks in (47) into simultaneous upper confidence bounds and an oracle inequality over the declared catalogue.

Theorem 13 (Finite-catalogue pilot coverage oracle)**.** *Use the same fixed prompts in the pilot and serving stages, with independent response samples, and suppose* (40) *holds. Let the pilot proposal satisfy $q_{\mathrm{pil}}(\cdot \mid x_j) \geq a_{\mathrm{pilot}} \pi_0(\cdot \mid x_j)$, let the nonempty catalogue $\mathcal{A}$ in* (44) *be fixed before observing pilot responses, and suppose the anchors, proposal components $q_{0:J}$ and their displayed density*

versions, pilot proposal, and fold partitions are likewise fixed independently of those responses. Suppose all displayed proposal densities are evaluated exactly, and fix $\delta_{\mathrm{p}} \in (0,1)$. *Set*

$$L_Z := \frac{e^{B/\beta}}{a_{\mathrm{pilot}}}, \qquad L_G := \frac{e^{2B/\beta}}{a_{\mathrm{pilot}} a}, \qquad u_{\mathrm{p}} := \log \frac{4nN(|\mathcal{A}|+1)}{\delta_{\mathrm{p}}}. \tag{64}$$

For $0 < \eta_Z, \eta_G < 1$, *let* $h := \min_{j\in[n],v\in\{1,2\}} |I_{j,v}|$ *and assume*

$$h \geq 8u_{\mathrm{p}} \max\left\{ \frac{L_Z}{\eta_Z^2}, \frac{L_G}{\eta_G^2} \right\}. \tag{65}$$

Then, under the conditional law of the pilot responses given $\boldsymbol{x}_{1:n}$, *with probability at least* $1-\delta_{\mathrm{p}}$, *every cross-fitted partition estimate is positive and, simultaneously for all* $j \in [n]$, $\ell \in [N]$, *and* $\alpha \in \mathcal{A}$,

$$\frac{1-\eta_G}{(1+\eta_Z)^2} G_{j\ell}(\alpha) \leq \widehat{G}_{j\ell}(\alpha) \leq \frac{1+\eta_G}{(1-\eta_Z)^2} G_{j\ell}(\alpha). \tag{66}$$

Consequently, the computable quantities

$$U_{j\ell}(\alpha) := \frac{(1+\eta_Z)^2}{1-\eta_G} \widehat{G}_{j\ell}(\alpha) \tag{67}$$

are simultaneous upper confidence bounds. The selector in (48) *obeys*

$$\max_{j,\ell} G_{j\ell}(\widehat{\alpha}) \leq \kappa_{\mathrm{p}} \min_{\alpha\in\mathcal{A}} \max_{j,\ell} G_{j\ell}(\alpha), \qquad \kappa_{\mathrm{p}} := \frac{1+\eta_G}{1-\eta_G} \left(\frac{1+\eta_Z}{1-\eta_Z} \right)^2. \tag{68}$$

The finite-anchor certificate is

$$\widehat{\Gamma}_{\mathrm{p},N} := \max_{j,\ell} U_{j\ell}(\widehat{\alpha}). \tag{69}$$

If, in addition, Assumption 8 holds and the anchors form a ρ*-net of* Θ, *then the continuous-family certificate*

$$\widehat{\Gamma}_{\mathrm{p}}(\rho) := e^{4L_s\rho/\beta} \widehat{\Gamma}_{\mathrm{p},N} \tag{70}$$

satisfies $\Gamma_2(q_{\widehat{\alpha}}, \Theta; \boldsymbol{x}_{1:n}) \leq \widehat{\Gamma}_{\mathrm{p}}(\rho)$ *on the same event.*

The pilot requirement is per prompt: if both folds have at least h responses, then the pilot cost is $H \geq 2h$ responses in addition to the M fresh serving responses. The result selects from the declared finite catalogue at the fixed prompt set; continuous-simplex optimization and transfer to unseen prompts require additional uniformity conditions. Its linear dependence on L_G follows from relative Bernstein concentration and $G_{j\ell}(\alpha) \geq 1$; an additive Hoeffding analysis would instead pay a quadratic envelope dependence.

Corollary 14 (Pilot-to-serving certificate)**.** *Under Theorem 13, fix* $\delta_{\mathrm{e}} \in (0,1)$, *freeze* $\widehat{q} = q_{\widehat{\alpha}}$, *and fix a serving budget* M. *For service restricted to the anchors, set* $\widehat{\Gamma} = \widehat{\Gamma}_{\mathrm{p},N}$. *In this branch, invoke the finite-service versions of the evaluation theorems, which do not require Assumption 8. For a continuous preference family, additionally suppose Assumption 8 holds and the anchors form a* ρ*-net, and set* $\widehat{\Gamma} = \widehat{\Gamma}_{\mathrm{p}}(\rho)$. *Conditional on the pilot data, choose robust block counts satisfying the conditions of Theorem 10 with* $\Gamma = \widehat{\Gamma}$ *and confidence* δ_{e}, *fix the corresponding deterministic partitions, and then draw the fresh serving pool. Then the simultaneous value and bounded-functional bounds* (57)–(58)*, with this certificate substituted for* Γ, *hold jointly over pilot and serving responses with probability at least* $1-\delta_{\mathrm{p}}-\delta_{\mathrm{e}}$. *The same composition applies to the one-block theorems when their additional envelope condition is certified.*

D.4 Necessity of Candidate Coverage

The support obstruction is clearest on a target-favored set that the realized pool may miss.

Proposition 15 (Missing-mass obstruction)**.** *Fix a prompt and preference, and let* $A \subseteq \mathcal{Y}$ *satisfy*

$$\pi_\theta^\star(A \mid x) = m_\star > 0, \qquad q(A \mid x) = r_\star \in (0,1). \tag{71}$$

For any (possibly randomized) policy whose support is contained in M i.i.d. candidates from q, the error for the bounded test function $f = \mathbf{1}_A$ is at least $m_\star$ with probability at least $(1 - r_\star)^M$. Therefore, any $(1 - \delta)$ guarantee with error strictly below $m_\star$ requires

$$M \geq \frac{\log(1/\delta)}{-\log(1 - r_\star)}. \tag{72}$$

Moreover, with the convention $\chi^2(P\|Q) = +\infty$ when P places mass where Q places none,

$$1 + \chi^2(\pi^\star_\theta \| q) \geq \frac{m_\star^2}{r_\star}. \tag{73}$$

Applying this obstruction to independent rare coordinates produces simultaneous dependence on coverage, accuracy, and the number of target–function pairs.

Theorem 16 (Joint target–function candidate lower bound)**.** *Fix $N \geq 2$, $0 < \epsilon \leq 1/4$, $\overline{\Gamma} \geq 2$, and $0 < \delta \leq 1/2$. For a single fixed prompt x and a finite preference set $\Theta = \{\theta_1, \ldots, \theta_N\}$ (equipped, for example, with the discrete metric), there is a finite, full-support Gibbs family with one fixed proposal q, targets $p_\ell = \pi^\star_{\theta_\ell}(\cdot \mid x)$, and $\mathcal{F} = \{f_1, \ldots, f_N\}$ consisting of N indicator functions such that*

$$\overline{\Gamma} - 1 \leq \mathbb{E}_q \left(\frac{p_\ell}{q} \right)^2 \leq \overline{\Gamma} \qquad (\ell \in [N]). \tag{74}$$

Here M counts all i.i.d. q-responses that may enter the output support. For every possibly randomized map from those candidates to policies $\widehat{p}_1, \ldots, \widehat{p}_N$ supported on those candidates, writing $\widehat{p}_\ell f$ for the expectation of f under $\widehat{p}_\ell$,

$$\mathbb{P}\left(\max_{\ell \in [N]} \max_{f \in \mathcal{F}} |\widehat{p}_\ell f - p_\ell f| \leq \epsilon \right) \leq \left[1 - (1 - r)^M \right]^N, \qquad r := \frac{4\epsilon^2}{\overline{\Gamma} - 1}. \tag{75}$$

The probability includes both candidate sampling and the map's internal randomness. Consequently, success probability at least $1 - \delta$ requires

$$M \geq \frac{\log\big(1/[1 - (1 - \delta)^{1/N}]\big)}{-\log(1 - r)} \geq \frac{3(\overline{\Gamma} - 1)}{16\epsilon^2} \log \frac{N}{2\delta} \geq \frac{3\overline{\Gamma}}{32\epsilon^2} \log \frac{N}{2\delta}. \tag{76}$$

The theorem is a single-prompt, fixed-proposal lower bound for candidate-supported (possibly randomized) policies. It does not cover adaptive proposal selection, post-revelation resampling, unrestricted policies, or log-partition estimation.

Thus the robust upper law $M = O(\Gamma \epsilon^{-2} \log(N(|\mathcal{F}| + 1)/\delta))$ is tight up to universal constants in its joint target–function logarithm on this fixed-proposal construction under only a second-moment restriction: by (74), its actual Γ_2 lies in $[\overline{\Gamma} - 1, \overline{\Gamma}]$. The construction's score oscillation is

$$B_{\mathrm{lb}} = \beta \log \frac{(\overline{\Gamma} - 1)(1 - r)}{2\epsilon(1 - 2\epsilon)}, \tag{77}$$

Thus the matching statement covers classes in which B may scale with $(\overline{\Gamma}, \epsilon)$; fixed-$B$ subclasses require a separate construction. Here $|\Theta| = |\mathcal{F}| = N$ and each hard target is paired with one hard evaluation function, so the construction captures their joint logarithmic complexity rather than separating preference entropy from function-class entropy. Parameter-space covering entropy is likewise representation-dependent: redundant parameterizations and promptwise additive score shifts can enlarge a score cover without changing the target policies.

Together, Proposition 15 and Theorem 16 are lower bounds for candidate-supported methods under explicit sampling oracles. They do not infer a minimax log-partition lower bound from the variance of one particular estimator. A matching value lower bound and a lower bound for unrestricted proposal design remain open.

E Supporting Results and Oracle Designs

In this section, we present comparison baselines, partial-coverage extensions, and finite-space proposal-design geometry. They support the main chain in Section D but are not required to state

its headline upper bound, pilot certificate, or fixed-proposal lower bound. Unless stated otherwise, probabilities in this section are conditional on the fixed prompts, and policy expectations are conditional on the realized candidate pool.

Assumption 17 (Uniform importance-weight envelope)**.** *There is a known or analytically bounded constant $\Gamma \le \Lambda < \infty$ such that, for every $j \in [n]$ and $\theta \in \Theta$,*

$$0 \le U_\theta(x_j, y) \le \Lambda \quad \textit{for every } y \textit{ with } q_j(y) > 0, \tag{78}$$

so that, for each j, one common set carries the envelope for all θ. Equivalently $\Gamma_\infty(q, \Theta; \boldsymbol{x}_{1:n}) \le \Lambda$, with Γ_∞ as in Eq. (31). The inequality $\Gamma \le \Lambda$ is always available from $U_\theta^2 \le \Lambda U_\theta$ and $\mathbb{E}U_\theta = 1$; a smaller second-moment bound gives a sharper result.

For $N_\rho = \mathcal{N}(\Theta, d_\Theta, \rho)$ and $u \ge 0$, define

$$\mathfrak{L}(\rho, \delta) := \log \frac{2nN_\rho}{\delta}, \tag{79}$$

$$r_M(u; \Gamma, \Lambda) := \sqrt{\frac{2\Gamma u}{M}} + \frac{2\Lambda u}{3M}. \tag{80}$$

The condition $r_M \le 1/2$ below guarantees that every estimated partition function is positive and that the logarithm is locally Lipschitz.

E.1 One-Block Log-Partition Approximation

Under the stronger envelope assumption, a direct Bernstein argument controls the one-block partition estimate uniformly.

Theorem 18 (Global-coverage value approximation)**.** *Suppose Assumptions 6 to 8 and 17 hold. Fix $\rho > 0$ and $\delta \in (0, 1)$, let $r = r_M(\mathfrak{L}(\rho, \delta); \Gamma, \Lambda)$, and assume $r \le 1/2$. Under the conditional law of the candidate pool given $\boldsymbol{x}_{1:n}$, with probability at least $1 - \delta$, every $\widehat{Z}_{\theta,M}(x_j)$ is positive and*

$$\max_{j \in [n]} \sup_{\theta \in \Theta} |\widehat{\psi}_{\theta,M}(x_j) - \psi_\theta(x_j)| \le 2\beta r + 2L_s\rho. \tag{81}$$

Consequently, on the same event,

$$\sup_{\theta \in \Theta} |\widehat{\Psi}_{n,M}(\theta) - \Psi_n(\theta)| \le 2\beta r + 2L_s\rho. \tag{82}$$

If the served preference set is exactly the finite set Θ_N, Assumption 8 is unnecessary: replace N_ρ by N in $\mathfrak{L}(\rho, \delta)$ and omit both $2L_s\rho$ extension terms.

Corollary 19 (One-block candidate complexity)**.** *Suppose Assumptions 6 to 8 and 17 hold, fix $\delta \in (0, 1)$, and suppose $R_\Theta > 0$, $d_{\mathrm{par}} \ge 0$, and, for every $\rho > 0$, $\log \mathcal{N}(\Theta, d_\Theta, \rho) \le d_{\mathrm{par}} \log(1 + R_\Theta/\rho)$. For $L_s > 0$ and $\epsilon > 0$, choose $\rho = \epsilon/(4L_s)$ and let $\eta_\epsilon = \min\{1/2, \epsilon/(4\beta)\}$. For a universal constant C, the sufficient per-prompt budget*

$$M \ge C \left(\frac{\Gamma}{\eta_\epsilon^2} + \frac{\Lambda}{\eta_\epsilon} \right) \left\{ d_{\mathrm{par}} \log\left(1 + \frac{4R_\Theta L_s}{\epsilon}\right) + \log\left(\frac{2n}{\delta}\right) \right\} \tag{83}$$

gives uniform value error at most ϵ. In the small-error regime $0 < \epsilon \le 2\beta$, the following simpler scaling is sufficient:

$$M \ge C \left[\left(\frac{\beta^2\Gamma}{\epsilon^2} + \frac{\beta\Lambda}{\epsilon} \right) \left\{ d_{\mathrm{par}} \log\left(1 + \frac{4R_\Theta L_s}{\epsilon}\right) + \log\left(\frac{2n}{\delta}\right) \right\} \right]. \tag{84}$$

For a finite preference set, replace the entropy term by $\log |\Theta|$.

The β^2 factor in the leading term follows from requesting absolute error in the value $\beta \log Z$. It disappears only if the error is normalized by β.

E.2 Nearest-Anchor Serving

If a system serves only the nearest anchor rather than directly reweighting at the requested preference, the anchor discretization error is explicit.

Corollary 20 (Candidate and anchor error)**.** *Suppose Assumption 8 holds. Fix an order on the anchors and let* $\mathrm{anc}(\theta)$ *be the first minimizer in that order of* $d_\Theta(\theta, \theta_\ell)$ *over* $\ell \in [N]$*. This tie-break is deterministic. If, on an event,*

$$\max_{j\in[n],\ell\in[N]} |\widehat{\psi}_{\theta_\ell,M}(x_j) - \psi_{\theta_\ell}(x_j)| \le \epsilon_M, \tag{85}$$

then

$$\max_{j\in[n]} \sup_{\theta\in\Theta} |\widehat{\psi}_{\theta_{\mathrm{anc}(\theta)},M}(x_j) - \psi_\theta(x_j)| \le \epsilon_M + L_s h_N(\Theta_N). \tag{86}$$

E.3 One-Block Evaluation Functions

We first record a perturbation fact used to extend a finite preference net to the continuum.

Lemma 21 (Bounded perturbations of Gibbs policies)**.** *Let* μ *be a base distribution on* $\mathcal{Y}$*, either a density such as the reference policy or a set of weights on finitely many responses such as a realized candidate pool, and let* $g, g' : \mathcal{Y} \to \mathbb{R}$ *have tilted masses* $Z_g := \mathbb{E}_\mu[e^{g/\beta}]$ *and* $Z_{g'} := \mathbb{E}_\mu[e^{g'/\beta}]$ *in* $(0,\infty)$*. Define the tilted policy* P_g *by* $\mathbb{E}_{P_g}[f] := \mathbb{E}_\mu[e^{g/\beta} f]/Z_g$*. If* $|g(y) - g'(y)| \le c$ *at every* y *carrying positive* μ*-weight, then for every* $f : \mathcal{Y} \to [-1, 1]$*,*

$$|\mathbb{E}_{P_g} f - \mathbb{E}_{P_{g'}} f| \le e^{2c/\beta} - 1. \tag{87}$$

For a finite $\mathcal{F}$, set

$$\mathfrak{L}_\mathcal{F}(\rho, \delta) := \log \frac{2nN_\rho(|\mathcal{F}| + 1)}{\delta}. \tag{88}$$

The Gibbs perturbation lemma and one-block concentration now give a uniform functional bound.

Theorem 22 (One-block bounded-functional approximation)**.** *Suppose Assumptions 6 to 8 and 17 hold and* $\mathcal{F}$ *is a prespecified nonempty finite class of* $[-1, 1]$*-valued functions. Fix* $\rho > 0$ *and* $\delta \in (0, 1)$*, and let* $r_\mathcal{F} = r_M(\mathfrak{L}_\mathcal{F}(\rho, \delta); \Gamma, \Lambda)$*. If* $r_\mathcal{F} \le 1/2$*, then, under the conditional law of the candidate pool given* $\boldsymbol{x}_{1:n}$*, with probability at least* $1 - \delta$*, all denominators in* (27) *are positive and*

$$\max_{j\in[n]} \sup_{\theta\in\Theta} \max_{f\in\mathcal{F}} \left| \mathbb{E}_{\widehat{\pi}_{\theta,M}(\cdot|x_j)} f(x_j, Y) - \mathbb{E}_{\pi^\star_\theta(\cdot|x_j)} f(x_j, Y) \right| \le 4r_\mathcal{F} + 2\left(e^{2L_s\rho/\beta} - 1\right). \tag{89}$$

If the served preference set is exactly the finite set Θ_N*, Assumption 8 is unnecessary: replace* N_ρ *by* N *and omit the exponential extension term.*

The theorem controls the integral probability metric induced by the declared evaluation class. KL and total-variation convergence are out of reach whenever the target spreads mass over far more responses than a finite pool can hold.

E.4 Partial Candidate Coverage

Global density-ratio bounds can be dominated by a small target tail. Partial coverage here means trimming a target tail while the support condition of Assumption 6 continues to hold, so $q_j(y) > 0$ at every response the target can produce. The tolerance τ pays for discarded target mass, while the moment conditions are imposed only on the retained region. The next result makes this tradeoff explicit for a finite preference set. For $j \in [n]$ and $\ell \in [N]$, choose a covered region $\mathcal{R}_{j\ell} \subseteq \mathcal{Y}$ independently of the serving candidates and write $\mathcal{R} := \{\mathcal{R}_{j\ell} : j \in [n], \ell \in [N]\}$. Define

$$m_{j\ell} := \pi^\star_{\theta_\ell}(\mathcal{R}^c_{j\ell} \mid x_j), \qquad \max_{j,\ell} m_{j\ell} \le \tau < 1. \tag{90}$$

Let

$$\overline{U}_{j\ell}(y) := U_{\theta_\ell}(x_j, y)\mathbf{1}\{y \in \mathcal{R}_{j\ell}\}, \tag{91}$$

$$\Gamma_2(\mathcal{R}) := \max_{j,\ell} \mathbb{E}_{q_j} \overline{U}_{j\ell}(Y)^2, \qquad \Gamma_\infty(\mathcal{R}) := \max\left\{1, \max_{j,\ell} \sup_{y:\, q_j(y)>0} \overline{U}_{j\ell}(y)\right\}. \tag{92}$$

These are the truncated analogues of (30) and (31), written with the region collection as their argument; the envelope is additionally floored at one, which only enlarges it and is used in the proof to bound the centered summands. The truncated partition estimate and policy are

$$\widehat{Z}^{\mathcal{R}}_{j\ell} := \frac{1}{M}\sum_{i=1}^{M} A_{\theta_\ell}(x_j, Y_{j,i})\mathbf{1}\{Y_{j,i} \in \mathcal{R}_{j\ell}\}, \tag{93}$$

$$\widehat{\pi}^{\mathcal{R}}_{j\ell} := \frac{\sum_{i=1}^{M} A_{\theta_\ell}(x_j, Y_{j,i})\mathbf{1}\{Y_{j,i} \in \mathcal{R}_{j\ell}\}\delta_{Y_{j,i}}}{\sum_{i=1}^{M} A_{\theta_\ell}(x_j, Y_{j,i})\mathbf{1}\{Y_{j,i} \in \mathcal{R}_{j\ell}\}}. \tag{94}$$

When the denominator in $\widehat{\pi}^{\mathcal{R}}_{j\ell}$ is zero, use a fixed candidate-supported fallback rule, such as the uniform distribution on the whole candidate pool. The event in the theorem below makes this convention irrelevant to its conclusion.

Theorem 23 (Partial-coverage upper bound)**.** *Suppose Assumption 6 holds for the finite target set Θ_N, $\mathcal{F}$ is nonempty and finite, fix $\delta \in (0,1)$, and suppose* (90)*–* (92) *hold. Set*

$$\mathfrak{L}_{N,\mathcal{F}} := \log\frac{2nN(|\mathcal{F}|+1)}{\delta}, \qquad r_{\mathcal{R}} := r_M(\mathfrak{L}_{N,\mathcal{F}}; \Gamma_2(\mathcal{R}), \Gamma_\infty(\mathcal{R}))\,. \tag{95}$$

If $r_{\mathcal{R}} \le (1-\tau)/2$, then, under the conditional law of the candidate pool given $\boldsymbol{x}_{1:n}$, with probability at least $1-\delta$, all truncated denominators are positive and simultaneously

$$\max_{j,\ell} |\beta \log \widehat{Z}^{\mathcal{R}}_{j\ell} - \psi_{\theta_\ell}(x_j)| \le \frac{2\beta r_{\mathcal{R}}}{1-\tau} + \beta\log\frac{1}{1-\tau}, \tag{96}$$

$$\max_{j,\ell}\max_{f\in\mathcal{F}} \left|\mathbb{E}_{\widehat{\pi}^{\mathcal{R}}_{j\ell}} f(x_j, Y) - \mathbb{E}_{\pi^\star_{\theta_\ell}(\cdot|x_j)} f(x_j, Y)\right| \le \frac{4r_{\mathcal{R}}}{1-\tau} + 2\tau. \tag{97}$$

Consequently, for a fixed proposal the metric-error upper envelope has the oracle form

$$\inf_{\substack{0\le\tau<1,\,\mathcal{R}:\\ \max_{j,\ell} m_{j\ell}\le\tau \\ r_{\mathcal{R}}\le(1-\tau)/2}} \left\{2\tau + \frac{4r_{\mathcal{R}}}{1-\tau}\right\}. \tag{98}$$

This oracle finite-budget upper envelope displays the trade-off directly: trimming can reduce the second moment and envelope but incurs target bias. At $\tau = 0$ it recovers the one-block, envelope-based finite-family global result. The optimizing regions are generally oracle objects. Selecting them on the serving pool invalidates the theorem; they must be declared analytically or selected on independent pilot data (Proposition 32). Pilot selection transfers the coverage quantities but not the trimmed mass: τ bounds a target-side mass, so the proposal diagnostics of Theorem 13 do not certify it and it requires either an analytic argument or its own certificate. Extending preference-dependent regions to a continuum requires complexity control for the joint class $\{(\theta, \mathcal{R}_\theta)\}$, not only Lipschitz scores.

The separate certificate the previous paragraph calls for is available from the same pilot sample. Although $m_{j\ell}$ is a target-side mass, it is the expectation under the pilot proposal of the untrimmed partition integrand restricted to the discarded region, so trimming (45) to that region and cross-fitting against the untrimmed estimate certifies it.

Proposition 24 (Pilot certificate for the trimmed mass)**.** *Adopt the pilot design and notation of Sections C.4 and D.3, suppose* (40) *holds and $q_{\mathrm{pil}}(\cdot \mid x_j) \ge a_{\mathrm{pilot}}\pi_0(\cdot \mid x_j)$, and let $\mathcal{R}$ be fixed independently of the pilot responses. Fix $\delta_\tau \in (0,1)$ and $0 < \eta_Z, \eta_\tau < 1$, put $u_\tau := \log\frac{8nN}{\delta_\tau}$, and assume*

$$h \ge \frac{8u_\tau L_Z}{3}\max\left\{\frac{1}{\eta_Z^2}, \frac{1}{\eta_\tau^2}\right\}. \tag{99}$$

For each fold $v \in \{1,2\}$ let $\widehat{Z}^{(v)}_{j\ell}$ be the untrimmed partition estimate of (45) *on $I_{j,v}$ and let*

$$\widehat{Z}^{c,(v)}_{j\ell} := \frac{1}{|I_{j,v}|}\sum_{i\in I_{j,v}} \frac{p_{0j}(\widetilde{Y}_{j,i})}{q_{\mathrm{pil},j}(\widetilde{Y}_{j,i})} \exp\Big(s_{\theta_\ell}(x_j, \widetilde{Y}_{j,i})/\beta\Big)\,\mathbf{1}\{\widetilde{Y}_{j,i} \notin \mathcal{R}_{j\ell}\} \tag{100}$$

be its trimmed counterpart. Set $\widehat{m}_{j\ell} := \max_{v \neq v'} \widehat{Z}^{c,(v)}_{j\ell} / \widehat{Z}^{(v')}_{j\ell}$. *Then, under the conditional law of the pilot responses given* $\boldsymbol{x}_{1:n}$, *with probability at least* $1 - \delta_\tau$, *every* $\widehat{Z}^{(v)}_{j\ell}$ *is positive and the pilot-determined quantity*

$$\widehat{\tau} := \max_{j \in [n], \ell \in [N]} \{(1 + \eta_Z)\widehat{m}_{j\ell} + \eta_\tau\} \tag{101}$$

satisfies $\max_{j,\ell} m_{j\ell} \leq \widehat{\tau}$ *and*

$$\widehat{\tau} \leq \frac{1 + \eta_Z}{1 - \eta_Z} \max_{j,\ell} m_{j\ell} + \left(\frac{1 + \eta_Z}{1 - \eta_Z} + 1 \right) \eta_\tau. \tag{102}$$

The certificate is additive rather than purely multiplicative in η_τ, which matches how τ is used: it enters (97) as the additive term 2τ, so only accuracy at the tolerance scale is needed, whereas a relative bound on an arbitrarily small discarded mass would require h to grow as $1/\max_{j,\ell} m_{j\ell}$. When $\widehat{\tau} < 1$ and the pilot event holds, Theorem 23 may be invoked with $\tau = \widehat{\tau}$ on fresh serving candidates, and Proposition 32 transfers the conclusion unconditionally. Running this certificate together with Theorem 13 costs $\delta_{\mathrm{p}} + \delta_\tau$ in total confidence, since the trimmed and untrimmed events are combined by a union bound; both use the same pilot responses, so no additional generation is required beyond taking h large enough for (65) and (99) at once.

E.5 Coverage Geometry and Finite-Space Oracle Benchmarks

The following elementary variational calculations are related to classical importance-sampling design (Agapiou et al., 2017; Veach & Guibas, 1995; He & Owen, 2014). They provide oracle benchmarks for coverage geometry; realizing them as LLM proposals additionally requires an implementable sampler.

Proposition 25 (Average-coverage oracle)**.** *Fix a prompt and target densities* $p_1, \ldots, p_N$ *on* $\mathcal{Y}$, *restricted to their union support* $\{y : p_\ell(y) > 0 \text{ for some } \ell\}$. *Interpret* $p_\ell(y)^2/q(y) = +\infty$ *when* $p_\ell(y) > 0 = q(y)$ *and as zero when both are zero. Let* $A = \sum_\ell \omega_\ell p_\ell^2$ *for weights* $\omega_\ell > 0$ *summing to one, and assume* $\int \sqrt{A(y)}\, \mathrm{d}y < \infty$. *Then a minimizer of*

$$\sum_{\ell=1}^{N} \omega_\ell \int \frac{p_\ell(y)^2}{q(y)}\, \mathrm{d}y \tag{103}$$

is

$$q^\star_{\mathrm{sqrt}}(y) = \frac{\sqrt{\sum_{\ell=1}^N \omega_\ell p_\ell(y)^2}}{\int \sqrt{\sum_{\ell=1}^N \omega_\ell p_\ell(y')^2}\, \mathrm{d}y'}. \tag{104}$$

Every minimizer agrees with $q^\star_{\mathrm{sqrt}}$ *at each* y *with* $A(y) > 0$. *Because the response space is restricted to the union support and every* ω_ℓ *is positive,* $A > 0$ *throughout the domain, so the minimizer is unique. If an ambient space larger than the union support is used, an optimizer must assign zero mass outside* $\{A > 0\}$.

Proposition 26 (Minimax second-moment oracle)**.** *Let* $p_1, \ldots, p_N$ *be probability mass functions on a finite response space* $\mathcal{S}$, *after removing points outside their union support. Interpret* $p_\ell(y)^2/q(y) = +\infty$ *when* $p_\ell(y) > 0 = q(y)$ *and as zero when both are zero. Then*

$$V_2 := \min_{q \in \Delta(\mathcal{S})} \max_{\ell \in [N]} \sum_{y \in \mathcal{S}} \frac{p_\ell(y)^2}{q(y)} = \max_{\lambda \in \Delta_N} \left(\sum_{y \in \mathcal{S}} \sqrt{\sum_{\ell=1}^N \lambda_\ell p_\ell(y)^2} \right)^2. \tag{105}$$

Every dual optimizer $\lambda^\star$ *collectively covers the union support, and*

$$q^\star_2(y) = \frac{\sqrt{\sum_{\ell=1}^N \lambda^\star_\ell p_\ell(y)^2}}{\sum_{z \in \mathcal{S}} \sqrt{\sum_{\ell=1}^N \lambda^\star_\ell p_\ell(z)^2}} \tag{106}$$

is primal optimal. It satisfies $\sum_y p_\ell(y)^2/q^\star_2(y) \leq V_2$ *for every* ℓ, *with equality for every active target* $\lambda^\star_\ell > 0$.

This equivalence aligns the oracle design with the worst-target coefficient in the approximation theorems. It is a finite-space structural benchmark; an unqualified minimax exchange over arbitrary infinite-dimensional densities would require additional compactness and integrability assumptions.

Proposition 27 (Target-overlap extremes)**.** *In the setting of Proposition 26, $1 \le V_2 \le N$ for every $p_1, \dots, p_N$. Both endpoints are attained: if all N targets are identical, then $V_2 = 1$; if their supports are pairwise disjoint, then $V_2 = N$, attained by $q = N^{-1} \sum_{\ell=1}^{N} p_\ell$.*

Exact disjointness is a generic target-family extreme. It cannot be attained by finite-score Gibbs tilts that share a full-support reference, because all such targets share that support. The following construction gives the corresponding near-separated law inside the main model.

Proposition 28 (Finite-score near-separated Gibbs family)**.** *Let $\mathcal{Y} = [N]$, $N \ge 2$, let π_0 be uniform, and fix $0 < \eta < (N-1)/N$. For*

$$s_\ell(y) = B\mathbf{1}\{y = \ell\}, \qquad B = \beta \log \frac{(N-1)(1-\eta)}{\eta}, \tag{107}$$

the Gibbs targets satisfy

$$p_\ell(\ell) = 1 - \eta, \qquad p_\ell(y) = \frac{\eta}{N-1} \quad (y \neq \ell). \tag{108}$$

Their minimax second moment is

$$V_2 = N \left\{ (1-\eta)^2 + \frac{\eta^2}{N-1} \right\}, \tag{109}$$

attained by the uniform proposal. Hence, for fixed N, $V_2 \to N$ as $\eta \downarrow 0$, while the required score oscillation grows as $B/\beta = \Theta(\log(N/\eta))$.

Thus sharing candidates is statistically inexpensive for overlapping targets and approaches a factor-N second-moment cost for separated Gibbs modes. For any proposal in Proposition 28, some distinguished response has proposal mass at most $1/N$; missing it produces indicator-function error $1 - \eta$ for the corresponding target. Therefore Proposition 15 forces $M = \Omega(N \log(1/\delta))$ for error below $1 - \eta$ under the worst preference revealed after generation. Under the symmetric minimax proposal, simultaneous realization of all N distinguished responses additionally exhibits the usual coupon-collector $N \log N$ effect. This near-separated construction alone does not capture the small-error law. Theorem 16 supplies the matching accuracy and joint finite target–function dependence; the corresponding value and unrestricted-design lower bounds remain open.

Proposition 29 (Worst-envelope oracle)**.** *Let $\bar{p}(y) = \max_{\ell \in [N]} p_\ell(y)$ and $C = \int \bar{p}(y)\, \mathrm{d}y$, and read ratios under the convention of Section B. Then $q_{\mathrm{env}} = \bar{p}/C$ satisfies*

$$\max_{\ell \in [N]} \sup_y \frac{p_\ell(y)}{q_{\mathrm{env}}(y)} = C \le N, \tag{110}$$

and no proposal has a smaller worst-target density-ratio envelope.

Proposition 30 (Defensive-mixture certificate)**.** *Suppose* (40) *holds and $q_\alpha(\cdot \mid x_j) \ge a\pi_0(\cdot \mid x_j)$ for all j. Then*

$$\sup_{j,\theta} \sup_{y:\, q_\alpha(y \mid x_j) > 0} \frac{\pi_\theta^\star(y \mid x_j)}{q_\alpha(y \mid x_j)} \le \frac{e^{B/\beta}}{a}. \tag{111}$$

Consequently, Assumptions 7 and 17 hold with $\Lambda = e^{B/\beta}/a$ and $\Gamma \le \Lambda$.

Proposition 31 (Coverage transfer)**.** *If two proposals satisfy $q(\cdot \mid x) \ge c\widetilde{q}(\cdot \mid x)$ for some $0 < c \le 1$, then for every target density p,*

$$\int \frac{p(y)^2}{q(y)}\, \mathrm{d}y \le \frac{1}{c} \int \frac{p(y)^2}{\widetilde{q}(y)}\, \mathrm{d}y, \qquad \sup_y \frac{p(y)}{q(y)} \le \frac{1}{c} \sup_y \frac{p(y)}{\widetilde{q}(y)}. \tag{112}$$

Thus a tractable proposal that multiplicatively covers an oracle design loses at most a factor $1/c$ in the two coverage quantities.

Proposition 32 (Validity after pilot selection)**.** *Let $\widehat{q}$ be any proposal selected using pilot data. For every pilot realization on the event above, suppose fresh serving candidates are independent draws from $\widehat{q}$ and all assumptions and radius conditions of the invoked theorem hold conditionally, with a pilot-determined bound Γ and, when needed, Λ, or with deterministic bounds such as Proposition 30. Then Theorems 10, 18, 22 and 23 hold under the conditional law of the fresh serving pool given the pilot data when their respective conditions are met, and hence also unconditionally. For Theorem 23 the region collection $\mathcal{R}$ must also be a function of the pilot responses, together with the trimmed-mass level τ. The latter bounds a target-side mass $m_{j\ell} = \pi^{\star}_{\theta_\ell}(\mathcal{R}^c_{j\ell} \mid x_j)$ rather than a proposal expectation, so it is not certified by the proposal diagnostics of Theorem 13 and must be supplied separately, either analytically or by the trimmed-mass certificate of Proposition 24. No guarantee is implied if the same candidates both select and evaluate the proposal without an additional adaptive-data analysis.*

F Proofs of the Main and Supporting Results

In this section, we present the proofs of our main results (Theorems 10, 13 and 16, Corollaries 11 and 14, and Proposition 15) and of the supporting results in Section E. We use the following two-sided Bernstein form throughout. If independent, centered variables $X_1, \dots, X_m$ satisfy $|X_i| \le c_{\mathrm{Ber}}$ everywhere and $\sum_i \mathbb{E}X_i^2 \le V$, then, for every $t > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{m} X_i\right| > \sqrt{2Vt} + \frac{c_{\mathrm{Ber}} t}{3}\right) \le 2e^{-t}. \tag{113}$$

One route is the sub-gamma bound $\log \mathbb{E}e^{\lambda \sum_i X_i} \le \lambda^2 V/[2(1-\lambda c_{\mathrm{Ber}}/3)]$ for $0 < \lambda < 3/c_{\mathrm{Ber}}$, applied also to $-X_i$. Some radii below use a larger common envelope for numerator and denominator events and are therefore conservative instances of (113).

Proof organization. The proofs below follow technical dependency rather than section order. The bounded-envelope baselines establish reusable concentration steps before the main envelope-free construction. An independent partial-coverage extension comes next; the anchor-transfer result precedes the pilot certificate because that certificate invokes it. The final two subsections collect independent proposal-design geometry and the main fixed-proposal lower bound, respectively.

F.1 Supporting One-Block Baselines

Proof of Theorem 18. Let $\mathcal{T}_\rho$ be a ρ-net of Θ with $|\mathcal{T}_\rho| = N_\rho$. For $j \in [n]$ and $\vartheta \in \mathcal{T}_\rho$, define

$$\widehat{R}_{j,\vartheta} := \frac{\widehat{Z}_{\vartheta,M}(x_j)}{Z_\vartheta(x_j)} = \frac{1}{M}\sum_{i=1}^{M} U_\vartheta(x_j, Y_{j,i}). \tag{114}$$

The summands are independent, have mean one and variance at most Γ, and satisfy $|U_\vartheta(x_j, Y_{j,i}) - 1| \le \Lambda$ because $0 \le U_\vartheta \le \Lambda$ and $\Lambda \ge 1$. We deliberately invoke Bernstein with the larger common envelope 2Λ, so the same denominator radius is compatible with the metric numerator events below. This gives the following conservative bound for every fixed pair (j, ϑ) and every $u > 0$:

$$\mathbb{P}\left(|\widehat{R}_{j,\vartheta} - 1| > \sqrt{\frac{2\Gamma u}{M}} + \frac{2\Lambda u}{3M}\right) \le 2e^{-u}. \tag{115}$$

Taking $u = \mathfrak{L}(\rho, \delta)$ and applying a union bound over nN_ρ pairs shows that, with probability at least $1 - \delta$,

$$\max_{j\in[n]} \max_{\vartheta\in\mathcal{T}_\rho} |\widehat{R}_{j,\vartheta} - 1| \le r. \tag{116}$$

On this event, $r \le 1/2$ implies $\widehat{R}_{j,\vartheta} \ge 1/2$. Hence every net-point partition estimate is positive and

$$|\widehat{\psi}_{\vartheta,M}(x_j) - \psi_\vartheta(x_j)| = \beta|\log \widehat{R}_{j,\vartheta}| \le \frac{\beta|\widehat{R}_{j,\vartheta} - 1|}{1 - r} \le 2\beta r. \tag{117}$$

It remains to leave the net. From Assumption 8, for any θ, θ' with $d_\Theta(\theta, \theta') \le \rho$,

$$e^{-L_s\rho/\beta} Z_{\theta'}(x_j) \le Z_\theta(x_j) \le e^{L_s\rho/\beta} Z_{\theta'}(x_j). \tag{118}$$

Therefore $|\psi_\theta(x_j) - \psi_{\theta'}(x_j)| \le L_s\rho$. The same sandwich holds for $\widehat{Z}_{\theta,M}$ and $\widehat{Z}_{\theta',M}$ because their nonnegative base weights $R_j(Y_{j,i})$ are identical. Positivity at a net point implies positivity throughout its neighborhood, and thus

$$|\widehat{\psi}_{\theta,M}(x_j) - \widehat{\psi}_{\theta',M}(x_j)| \le L_s\rho. \tag{119}$$

For each θ, choose $\vartheta \in \mathcal{T}_\rho$ within distance ρ and combine (117)–(119) with the triangle inequality. This proves (81). Averaging the pointwise inequality over the fixed prompts proves (82). If Θ_N is finite, take $\mathcal{T}_\rho = \Theta_N$, replace N_ρ by N in the union bound, and stop at the net-point bound; no Lipschitz extension is used. □

Proof of Corollary 19. Choose $\rho = \epsilon/(4L_s)$, so the extension term in (81) is $\epsilon/2$. Let $\eta_\epsilon = \min\{1/2, \epsilon/(4\beta)\}$ and require $r \le \eta_\epsilon$. Then $r \le 1/2$, and the log-partition term is at most $\epsilon/2$ whether $\epsilon \le 2\beta$ or $\epsilon > 2\beta$. Satisfying this requirement separately for the square-root and linear terms in (80) gives

$$M \ge C\left(\frac{\Gamma}{\eta_\epsilon^2} + \frac{\Lambda}{\eta_\epsilon}\right)\mathfrak{L}(\rho, \delta), \tag{120}$$

which proves (83). If $0 < \epsilon \le 2\beta$, then $\eta_\epsilon = \epsilon/(4\beta)$ and this becomes

$$M \ge C\left(\frac{\beta^2\Gamma}{\epsilon^2} + \frac{\beta\Lambda}{\epsilon}\right)\mathfrak{L}(\rho, \delta) \tag{121}$$

for a universal constant C. Substituting the entropy assumption and the chosen ρ yields (84). When Θ is finite, a union bound over its elements gives the stated replacement. □

Proof of Lemma 21. The score gap $|g - g'| \le c$ on the support of μ implies

$$e^{-c/\beta} \le \frac{Z_g}{Z_{g'}} \le e^{c/\beta}. \tag{122}$$

Consequently, the relative weight that P_g places on a response carrying positive μ-weight satisfies

$$e^{-2c/\beta} \le \varrho(y) := \frac{e^{g(y)/\beta}/Z_g}{e^{g'(y)/\beta}/Z_{g'}} \le e^{2c/\beta}. \tag{123}$$

Since $\mathbb{E}_{P_g}[f] = \mathbb{E}_{P_{g'}}[\varrho f]$ for every f, taking $|f| \le 1$ gives

$$|\mathbb{E}_{P_g} f - \mathbb{E}_{P_{g'}} f| \le \mathbb{E}_{P_{g'}} |\varrho(Y) - 1| \le e^{2c/\beta} - 1, \tag{124}$$

which proves the claim. □

Proof of Theorem 22. Fix a ρ-net $\mathcal{T}_\rho$ of Θ with $|\mathcal{T}_\rho| = N_\rho$. Fix j, $\vartheta \in \mathcal{T}_\rho$, and $f \in \mathcal{F}$. Write

$$\widehat{N}_{j,\vartheta,f} = \frac{1}{M}\sum_{i=1}^{M} U_\vartheta(x_j, Y_{j,i}) f(x_j, Y_{j,i}), \qquad \mu_{j,\vartheta,f} = \mathbb{E}_{\pi^\star_\vartheta(\cdot|x_j)} f(x_j, Y). \tag{125}$$

The summands in $\widehat{N}$ have mean μ, variance at most Γ, and absolute centered value at most $\Lambda + 1 \le 2\Lambda$. Bernstein's inequality with $u = \mathfrak{L}_\mathcal{F}(\rho, \delta)$ gives

$$|\widehat{N}_{j,\vartheta,f} - \mu_{j,\vartheta,f}| \le r_\mathcal{F} \tag{126}$$

except on an event of probability at most $2e^{-u}$. The same radius controls $|\widehat{R}_{j,\vartheta} - 1|$, because its summands have variance at most Γ and centered envelope at most Λ. A union bound over all $nN_\rho(|\mathcal{F}| + 1)$ numerator and denominator events gives simultaneous control with probability at least $1 - \delta$.

On this event, $\widehat{R}_{j,\vartheta} \geq 1 - r_{\mathcal{F}} \geq 1/2$, and

$$\left| \frac{\widehat{N}_{j,\vartheta,f}}{\widehat{R}_{j,\vartheta}} - \mu_{j,\vartheta,f} \right| \leq \frac{|\widehat{N}_{j,\vartheta,f} - \mu_{j,\vartheta,f}| + |\mu_{j,\vartheta,f}||1 - \widehat{R}_{j,\vartheta}|}{\widehat{R}_{j,\vartheta}} \leq 4r_{\mathcal{F}}, \tag{127}$$

because $|\mu_{j,\vartheta,f}| \leq 1$.

For an arbitrary θ, choose a net point ϑ within ρ. Apply Lemma 21 first to the reference policy $\pi_0(\cdot \mid x_j)$ and then to the base distribution that places weight

$$\frac{R_j(Y_{j,i})}{\sum_{i'=1}^{M} R_j(Y_{j,i'})} \tag{128}$$

on each stored candidate $Y_{j,i}$, which leaves the tilted policy (27) unchanged because the lemma is invariant to rescaling the base weights. Indeed, $\widehat{R}_{j,\vartheta} \geq 1/2$ and the score exponential is strictly positive, so the raw weights $R_j(Y_{j,i})$ do not all vanish and this normalization is well defined; the same fact makes every off-net denominator in (27) positive. In both cases, Assumption 8 gives $c = L_s\rho$. Adding the two perturbation errors to (127) proves (89). If Θ_N is finite, apply the union bound directly and omit the perturbation step. □

F.2 Main Envelope-Free Upper Bound

Proof of Lemma 9. Write

$$\mu = \mathbb{E}[Uf], \qquad S = \frac{1}{b}\sum_{i=1}^{b} U_i, \qquad X = \frac{1}{b}\sum_{i=1}^{b} U_i\big(f(Y_i) - \mu\big). \tag{129}$$

Then $\mathbb{E}X = 0$, $\mathbb{E}X^2 \leq 4\Gamma/b$, and $\mathbb{E}(S-1)^2 \leq \Gamma/b$. Let $R = P_b f - \mu$ and $G = \{S \geq 1/2\}$. On G, $R = X/S$, so

$$|R - X| \leq 2|X||S - 1|. \tag{130}$$

On G^c, $|R| \leq 2$ for both the normalized policy and the declared fallback, while $|X| \leq 2S < 1$. Chebyshev's inequality gives $\mathbb{P}(G^c) \leq 4\Gamma/b$. Therefore Cauchy–Schwarz yields

$$|\mathbb{E}R| = |\mathbb{E}(R - X)| \leq 2\sqrt{\mathbb{E}X^2\,\mathbb{E}(S-1)^2} + 3\mathbb{P}(G^c) \leq \frac{16\Gamma}{b}. \tag{131}$$

For the second moment, $R^2 \leq 4X^2$ on G and $R^2 \leq 4$ on G^c. Consequently,

$$\mathbb{E}R^2 \leq 4\mathbb{E}X^2 + 4\mathbb{P}(G^c) \leq \frac{32\Gamma}{b}. \tag{132}$$

The low-denominator event includes $S = 0$, so the argument also covers the zero-weight fallback. □

The proof of Theorem 10, the formal version of Theorem 2, has four steps. For a net preference, a prompt and an $f \in \mathcal{F}$, a ratio decomposition against U_θ bounds the within-block conditional bias of $\widehat{\pi}^{(k)}_{j,\theta} f$ by $O(\Gamma/b)$ (Lemma 9). The K block functionals are independent and lie in $[-1, 1]$ by construction, so Bernstein's inequality (Boucheron et al., 2013) applies to their average with variance proxy of order Γ/b; this is the step at which a one-block analysis would instead need Γ_∞. A union bound over prompts, Θ_ρ and $\mathcal{F}$ yields one event, and a Gibbs perturbation bound (Lemma 21) extends it to all of Θ, preferences at distance ρ inducing targets whose functional expectations differ by at most $2(e^{2L_s\rho/\beta} - 1)$. The value branch replaces the second step by median-of-means concentration.

Proof of Theorem 10. Let $\mathcal{T}_\rho$ be a fixed ρ-net with cardinality N_ρ. We first control the MoM partition estimate on this net. For a prompt j, net point ϑ, and value block $\mathcal{B}^Z_{j,r}$, define

$$S_{j,\vartheta,r} := \frac{1}{|\mathcal{B}^Z_{j,r}|} \sum_{i \in \mathcal{B}^Z_{j,r}} U_\vartheta(x_j, Y_{j,i}). \tag{133}$$

It has mean one and variance at most Γ/b_Z. Hence, for $\kappa_Z = 2\sqrt{\Gamma/b_Z}$,

$$\mathbb{P}(|S_{j,\vartheta,r} - 1| > \kappa_Z) \le \frac{1}{4}. \tag{134}$$

The value blocks are independent. Hoeffding's inequality for the number of bad blocks shows that their median leaves $[1-\kappa_Z, 1+\kappa_Z]$ with probability at most $e^{-K_Z/8}$. The condition on K_Z and a union bound over nN_ρ cells therefore give, except on an event of probability at most $\delta/4$,

$$\max_{j\in[n]} \max_{\vartheta\in\mathcal{T}_\rho} \left| \frac{\widetilde{Z}^{\mathrm{MoM}}_{j,\vartheta}}{Z_\vartheta(x_j)} - 1 \right| \le \kappa_Z. \tag{135}$$

Since $\kappa_Z \le 1/2$, all these estimates are positive and their scaled-log errors are at most $2\beta\kappa_Z$.

For the policy, fix (j,ϑ,f) on the same net and write $X_k = \mathbb{E}_{\widehat{\pi}^{(k)}_{j,\vartheta}} f(x_j, Y)$. Each policy, including its zero-weight fallback, is a fixed function only of the candidates in its own block. The K variables are therefore independent and lie in $[-1,1]$. By Lemma 9 and the fact that every block has at least b candidates,

$$\left| \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}X_k - \mathbb{E}_{\pi^\star_\vartheta} f \right| \le \frac{16\Gamma}{b}, \qquad \max_k \mathrm{Var}(X_k) \le \frac{32\Gamma}{b}. \tag{136}$$

Using centered envelope $c_{\mathrm{Ber}} = 2$ and conservatively doubling the exact linear term, Bernstein's inequality with $u_{\mathrm{P}} := \mathfrak{L}_{\mathrm{P}}(\rho,\delta)$ gives

$$\left| \frac{1}{K}\sum_{k=1}^{K} (X_k - \mathbb{E}X_k) \right| \le 8\sqrt{\frac{\Gamma u_{\mathrm{P}}}{bK}} + \frac{4u_{\mathrm{P}}}{3K} \tag{137}$$

except on an event of probability at most $2e^{-u_{\mathrm{P}}}$. A union bound over $nN_\rho|\mathcal{F}|$ cells costs at most $\delta/2$. Combining the bias with (137) proves the net-point metric bound r_{blk}.

It remains to extend both results off the net. Put $c = L_s\rho/\beta$. If $d_\Theta(\theta,\vartheta) \le \rho$, then the raw weights satisfy, candidate by candidate,

$$e^{-c} A_\vartheta(x_j, Y_{j,i}) \le A_\theta(x_j, Y_{j,i}) \le e^{c} A_\vartheta(x_j, Y_{j,i}). \tag{138}$$

Every value-block mean inherits this coordinatewise sandwich. Fixed order statistics are coordinatewise monotone, so the multiplicative sandwich also passes to their median. Therefore the MoM scaled log changes by at most $L_s\rho$; the true scaled log changes by the same amount by (118). This proves (57).

For a positive-denominator policy block, (138) and the corresponding denominator sandwich bound the density ratio between the two empirical block policies by e^{2c}. A zero-denominator block is the same for every preference and invokes the same fallback. Hence, for every $|f| \le 1$,

$$|\widehat{\pi}^{(k)}_{j,\theta} f - \widehat{\pi}^{(k)}_{j,\vartheta} f| \le e^{2c} - 1. \tag{139}$$

The inequality is preserved by averaging the block policies. The true Gibbs policies obey the same bound by Lemma 21. Adding these two extension terms to the net-point event proves (58). The MoM value event fails with probability at most $\delta/4$, and the policy–metric net event fails with probability at most $\delta/2$. The deterministic Lipschitz extension adds no failure probability, so the total is at most $3\delta/4 < \delta$. If Θ_N is finite, take $\mathcal{T}_\rho = \Theta_N$, replace N_ρ by N in both union bounds, and stop at the net-point value and metric bounds; no Lipschitz extension is used. □

Proof of Corollary 11. Let $u_{\mathrm{P}} = \mathfrak{L}_{\mathrm{P}}(\rho,\delta)$ and $u_{\mathrm{V}} = \mathfrak{L}_{\mathrm{V}}(\rho,\delta)$. For balanced policy blocks and $M \ge C\Gamma u_{\mathrm{P}}$, the choice of K in (60) can be rounded so that

$$b \asymp \sqrt{\frac{\Gamma M}{u_{\mathrm{P}}}}, \qquad bK \asymp M. \tag{140}$$

For balanced MoM blocks and $M \ge C\Gamma u_{\mathrm{V}}$, the choice of K_Z can be rounded so that $b_Z \asymp M/u_{\mathrm{V}}$ and (56) holds. Substitution into (54)–(55) gives, respectively,

$$r_{\mathrm{blk}} \le C_1\sqrt{\frac{\Gamma u_{\mathrm{P}}}{M}}, \qquad \kappa_Z \le C_2\sqrt{\frac{\Gamma u_{\mathrm{V}}}{M}} \tag{141}$$

for universal constants.

For the value, take $\rho = \epsilon_\psi/(4L_s)$. The extension term is $\epsilon_\psi/2$, while $M \geq C\beta^2\Gamma u_{\mathrm{V}}/\epsilon_\psi^2$ makes the first term at most $\epsilon_\psi/2$. The restriction $\epsilon_\psi \leq 2\beta$ also absorbs the feasibility condition $M \geq C\Gamma u_{\mathrm{V}}$. For the metrics, take

$$\rho = \frac{\beta}{2L_s} \log\left(1 + \frac{\epsilon_F}{4}\right), \tag{142}$$

so the extension term is exactly $\epsilon_F/2$; with $M \geq C\Gamma u_{\mathrm{P}}/\epsilon_F^2$ bounding the first term by $\epsilon_F/2$ and absorbing the feasibility condition $M \geq C\Gamma u_{\mathrm{P}}$, the second conclusion follows from (141). Substitution of the assumed entropy bound into u_{V} for the value and into u_{P} for the metrics gives the displayed complexity laws. □

F.3 Supporting Partial Coverage

Proof of Theorem 23. For a fixed (j, ℓ), abbreviate $m = m_{j\ell}$ and define

$$\widehat{R}^{\mathcal{R}}_{j\ell} := \frac{\widehat{Z}^{\mathcal{R}}_{j\ell}}{Z_{\theta_\ell}(x_j)} = \frac{1}{M}\sum_{i=1}^{M} \overline{U}_{j\ell}(Y_{j,i}). \tag{143}$$

Its expectation is $a_{j\ell} := 1 - m \geq 1 - \tau$. For $f \in \mathcal{F}$, also let

$$\widehat{N}^{\mathcal{R}}_{j\ell f} := \frac{1}{M}\sum_{i=1}^{M} \overline{U}_{j\ell}(Y_{j,i}) f(x_j, Y_{j,i}), \qquad \nu_{j\ell f} := \mathbb{E}\widehat{N}^{\mathcal{R}}_{j\ell f}. \tag{144}$$

Both centered summands have second moment at most $\Gamma_2(\mathcal{R})$: since $|f| \leq 1$, $\mathrm{Var}(\overline{U}_{j\ell} f) \leq \mathbb{E}_{q_j}[\overline{U}^2_{j\ell} f^2] \leq \mathbb{E}_{q_j}\overline{U}^2_{j\ell} \leq \Gamma_2(\mathcal{R})$, and the denominator case is the special case $f \equiv 1$. Because (92) floors the truncated envelope at one, $\Gamma_\infty(\mathcal{R}) \geq 1 \geq a_{j\ell}$, so the centered denominator summand is bounded by $\Gamma_\infty(\mathcal{R})$ and the centered numerator summand by $2\Gamma_\infty(\mathcal{R})$. Bernstein's inequality and a union bound over $nN(|\mathcal{F}| + 1)$ quantities therefore give, with probability at least $1 - \delta$,

$$|\widehat{R}^{\mathcal{R}}_{j\ell} - a_{j\ell}| \leq r_{\mathcal{R}}, \qquad |\widehat{N}^{\mathcal{R}}_{j\ell f} - \nu_{j\ell f}| \leq r_{\mathcal{R}} \tag{145}$$

simultaneously for all indices. Since $r_{\mathcal{R}} \leq (1 - \tau)/2 \leq a_{j\ell}/2$, every truncated denominator is positive. Moreover,

$$|\log \widehat{R}^{\mathcal{R}}_{j\ell}| \leq \left|\log \frac{\widehat{R}^{\mathcal{R}}_{j\ell}}{a_{j\ell}}\right| + |\log a_{j\ell}| \leq \frac{2r_{\mathcal{R}}}{a_{j\ell}} + \log\frac{1}{1-m} \leq \frac{2r_{\mathcal{R}}}{1-\tau} + \log\frac{1}{1-\tau}, \tag{146}$$

which proves (96) after multiplying by β.

Let $P^{\mathcal{R}}_{j\ell}$ be the target conditioned on $\mathcal{R}_{j\ell}$. Then $\nu_{j\ell f} = a_{j\ell}\mathbb{E}_{P^{\mathcal{R}}_{j\ell}} f$, so

$$\left|\frac{\widehat{N}^{\mathcal{R}}_{j\ell f}}{\widehat{R}^{\mathcal{R}}_{j\ell}} - \mathbb{E}_{P^{\mathcal{R}}_{j\ell}} f\right| \leq \frac{2r_{\mathcal{R}}}{\widehat{R}^{\mathcal{R}}_{j\ell}} \leq \frac{4r_{\mathcal{R}}}{1-\tau}. \tag{147}$$

Finally, if $m > 0$, decompose $\pi^\star_{\theta_\ell} = (1 - m)P^{\mathcal{R}}_{j\ell} + mP^{\mathcal{R}^c}_{j\ell}$ using the conditional law on the complement. If $m = 0$, the truncation bias is zero. In either case, the expectation of any $[-1, 1]$-valued f changes by at most $2m \leq 2\tau$ upon conditioning. The triangle inequality proves (97). □

Proof of Proposition 24. Use the densities of (43) and put $X_{j\ell} := p_{j\ell}/q_{\mathrm{pil},j}$. For $Y \sim q_{\mathrm{pil}}(\cdot \mid x_j)$ we have $\mathbb{E}_{q_{\mathrm{pil}}} X_{j\ell} = 1$, since $\pi^\star_{\theta_\ell}$ is a distribution and $q_{\mathrm{pil}} \geq a_{\mathrm{pilot}}\pi_0$ makes q_{pil} cover $\pi^\star_{\theta_\ell}$. Writing $p_{j\ell} = p_{0j}e^{s_{\theta_\ell}/\beta}/Z_{\theta_\ell}(x_j)$ and bounding the score oscillation as in (40) gives $0 \leq X_{j\ell} \leq e^{B/\beta}/a_{\mathrm{pilot}} = L_Z$, hence also $\mathrm{Var}_{q_{\mathrm{pil}}}(X_{j\ell}) \leq \mathbb{E}_{q_{\mathrm{pil}}} X^2_{j\ell} \leq L_Z$. These are the same three facts the proof of Theorem 13 establishes for this variable. Put $X^c_{j\ell} := X_{j\ell}\mathbf{1}\{\cdot \notin \mathcal{R}_{j\ell}\}$. Since $\mathcal{R}$ is fixed independently of the pilot responses, $X^c_{j\ell}$ obeys the same envelope and variance bounds, and

$$\mathbb{E}_{q_{\mathrm{pil}}} X^c_{j\ell} = \int_{\mathcal{R}^c_{j\ell}} \frac{p_{j\ell}}{q_{\mathrm{pil},j}}\, q_{\mathrm{pil},j}(y)\, \mathrm{d}y = \pi^\star_{\theta_\ell}(\mathcal{R}^c_{j\ell} \mid x_j) = m_{j\ell}, \tag{148}$$

which is the identity that makes this target-side mass estimable from pilot data. Writing $\overline{X}^{c,(v)}_{j\ell}$ for the fold average of $X^c_{j\ell}$, the Gibbs density identity gives $\widehat{Z}^{c,(v)}_{j\ell}/Z_{\theta_\ell}(x_j) = \overline{X}^{c,(v)}_{j\ell}$ alongside $\widehat{Z}^{(v)}_{j\ell}/Z_{\theta_\ell}(x_j) = \overline{X}^{(v)}_{j\ell}$.

Apply (113) on each fold of size at least h, once to $X_{j\ell}$ with tolerance η_Z and once to $X^c_{j\ell}$ with the additive tolerance η_τ. Both give $2\exp(-3h\eta^2/(8L_Z))$ with the relevant η, so (99) makes each at most $\delta_\tau/(4nN)$. There are $2nN$ deviation events of each of the two kinds, hence $4nN$ in total, and a union bound gives an event E_τ of probability at least $1-\delta_\tau$ on which, for all j, ℓ and both folds,

$$\left|\overline{X}^{(v)}_{j\ell} - 1\right| \le \eta_Z, \qquad \left|\overline{X}^{c,(v)}_{j\ell} - m_{j\ell}\right| \le \eta_\tau. \tag{149}$$

On E_τ we have $\overline{X}^{(v)}_{j\ell} \ge 1-\eta_Z > 0$, so $\widehat{Z}^{(v)}_{j\ell} > 0$ and each ratio $\widehat{Z}^{c,(v)}_{j\ell}/\widehat{Z}^{(v')}_{j\ell} = \overline{X}^{c,(v)}_{j\ell}/\overline{X}^{(v')}_{j\ell}$ is well defined; the common factor $Z_{\theta_\ell}(x_j)$ cancels, which is why no anchor normalizer is needed.

For validity, fix j, ℓ and distinct v, v'. On E_τ,

$$\frac{\overline{X}^{c,(v)}_{j\ell}}{\overline{X}^{(v')}_{j\ell}} \ge \frac{m_{j\ell}-\eta_\tau}{1+\eta_Z}, \tag{150}$$

so $(1+\eta_Z)\widehat{m}_{j\ell} + \eta_\tau \ge m_{j\ell}$, where the inequality also holds trivially when $m_{j\ell} \le \eta_\tau$ because $\widehat{m}_{j\ell} \ge 0$. Taking the maximum over (j,ℓ) yields $\max_{j,\ell} m_{j\ell} \le \widehat{\tau}$.

For the oracle direction, the same event gives

$$\frac{\overline{X}^{c,(v)}_{j\ell}}{\overline{X}^{(v')}_{j\ell}} \le \frac{m_{j\ell}+\eta_\tau}{1-\eta_Z}, \tag{151}$$

uniformly in $v \ne v'$, hence $(1+\eta_Z)\widehat{m}_{j\ell} \le \frac{1+\eta_Z}{1-\eta_Z}(m_{j\ell}+\eta_\tau)$. Adding η_τ and maximizing over (j,ℓ) gives (102). Since $\widehat{\tau}$ depends on the pilot responses alone, it is admissible as Proposition 32 requires. □

F.4 Supporting Anchor Approximation and Pilot Transfer

Proof of Corollary 20. For every θ, let $\theta_{\mathrm{anc}(\theta)}$ be a nearest anchor. The true value is L_s-Lipschitz by (118). Hence

$$|\widehat{\psi}_{\theta_{\mathrm{anc}(\theta)},M}(x_j) - \psi_\theta(x_j)| \le |\widehat{\psi}_{\theta_{\mathrm{anc}(\theta)},M}(x_j) - \psi_{\theta_{\mathrm{anc}(\theta)}}(x_j)| + |\psi_{\theta_{\mathrm{anc}(\theta)}}(x_j) - \psi_\theta(x_j)| \tag{152}$$

$$\le \epsilon_M + L_s h_N(\Theta_N). \tag{153}$$

Taking the maximum over prompts and supremum over preferences proves the result. □

Proof of Proposition 12. Let $c = L_s\rho$. The partition sandwich (118) will be used below. If q_j fails to cover $\pi^\star_{\theta_\ell}(\cdot \mid x_j)$, the right-hand side of (63) is $+\infty$ and the claim is immediate. Otherwise, the strictly positive Gibbs density gives $\pi^\star_{\theta_\ell}(\cdot \mid x_j)$ and $\pi_0(\cdot \mid x_j)$ the same support, so q_j covers $\pi_0(\cdot \mid x_j)$ as well. The Lipschitz condition holds on $\{y : q_j(y) > 0\}$, which carries every response the target can produce. On that set, the partition sandwich and score Lipschitz condition imply

$$\frac{\pi^\star_\theta(y \mid x_j)}{\pi^\star_{\theta_\ell}(y \mid x_j)} \le e^{2c/\beta}. \tag{154}$$

The Gibbs construction and the chain rule for densities now give

$$\frac{\pi^\star_\theta}{q_j} = \frac{\pi^\star_\theta}{\pi^\star_{\theta_\ell}}\frac{\pi^\star_{\theta_\ell}}{q_j} \qquad \text{for every } y \text{ with } q_j(y) > 0. \tag{155}$$

Squaring this identity, applying the ratio bound, and integrating with respect to q_j proves (63). □

F.5 Certified Proposal Selection

Proof of Theorem 13. Use the densities in (43). For a pilot response $Y \sim q_{\mathrm{pil}}(\cdot \mid x_j)$, define

$$X_{j\ell}(Y) := \frac{p_{j\ell}(Y)}{q_{\mathrm{pil},j}(Y)}, \qquad W_{j\ell\alpha}(Y) := \frac{p_{j\ell}(Y)^2}{q_{\mathrm{pil},j}(Y)q_{\alpha,j}(Y)}. \tag{156}$$

The score-oscillation and defensive-mixture conditions imply

$$0 \le X_{j\ell} \le L_Z, \qquad 0 \le W_{j\ell\alpha} \le L_G. \tag{157}$$

Moreover, $q_{\mathrm{pil}} \ge a_{\mathrm{pilot}}\pi_0$ makes q_{pil} cover $\pi^\star_{\theta_\ell}$, so $\mathbb{E}_{q_{\mathrm{pil}}}X_{j\ell} = 1$ and $\mathbb{E}_{q_{\mathrm{pil}}}W_{j\ell\alpha} = G_{j\ell}(\alpha)$. Cauchy–Schwarz gives $G_{j\ell}(\alpha) \ge 1$. Hence

$$\mathrm{Var}_{q_{\mathrm{pil}}}(X_{j\ell}) \le L_Z, \qquad \mathrm{Var}_{q_{\mathrm{pil}}}(W_{j\ell\alpha}) \le L_G G_{j\ell}(\alpha). \tag{158}$$

For each fold $v \in \{1,2\}$, define the sample averages

$$\overline{X}^{(v)}_{j\ell} := \frac{1}{|I_{j,v}|}\sum_{i\in I_{j,v}} X_{j\ell}(\widetilde{Y}_{j,i}), \qquad \overline{W}^{(v)}_{j\ell\alpha} := \frac{1}{|I_{j,v}|}\sum_{i\in I_{j,v}} W_{j\ell\alpha}(\widetilde{Y}_{j,i}), \tag{159}$$

$$\widehat{Z}^{(v)}_{j\ell} := \frac{1}{|I_{j,v}|}\sum_{i\in I_{j,v}} \frac{p_{0j}(\widetilde{Y}_{j,i})e^{s_{\theta_\ell}(x_j,\widetilde{Y}_{j,i})/\beta}}{q_{\mathrm{pil},j}(\widetilde{Y}_{j,i})}, \quad \widehat{T}^{(v)}_{j\ell}(\alpha) := \frac{1}{|I_{j,v}|}\sum_{i\in I_{j,v}} \frac{p_{0j}(\widetilde{Y}_{j,i})^2 e^{2s_{\theta_\ell}(x_j,\widetilde{Y}_{j,i})/\beta}}{q_{\mathrm{pil},j}(\widetilde{Y}_{j,i})q_{\alpha,j}(\widetilde{Y}_{j,i})}. \tag{160}$$

The Gibbs density identity then gives

$$\frac{\widehat{Z}^{(v)}_{j\ell}}{Z_{\theta_\ell}(x_j)} = \overline{X}^{(v)}_{j\ell}, \qquad \frac{\widehat{T}^{(v)}_{j\ell}(\alpha)}{Z_{\theta_\ell}(x_j)^2} = \overline{W}^{(v)}_{j\ell\alpha}. \tag{161}$$

For distinct $v, v' \in \{1,2\}$, the corresponding cross-fitted ratio is

$$\widehat{G}^{(v|v')}_{j\ell}(\alpha) := \frac{\widehat{T}^{(v)}_{j\ell}(\alpha)}{(\widehat{Z}^{(v')}_{j\ell})^2}. \tag{162}$$

For a fold containing at least h observations, Bernstein's inequality applied with either tolerance therefore gives

$$\mathbb{P}\left(\left|\overline{X}^{(v)}_{j\ell} - 1\right| > \eta_Z\right) \le 2\exp\left(-\frac{3h\eta_Z^2}{8L_Z}\right), \tag{163}$$

$$\mathbb{P}\left(\left|\overline{W}^{(v)}_{j\ell\alpha} - G_{j\ell}(\alpha)\right| > \eta_G G_{j\ell}(\alpha)\right) \le 2\exp\left(-\frac{3h\eta_G^2}{8L_G}\right). \tag{164}$$

The sharper second exponent contains the additional factor $G_{j\ell}(\alpha) \ge 1$; dropping it gives the displayed uniform catalogue-wide bound. Under (65), each right-hand side is at most $2e^{-u_{\mathrm{p}}}$. There are two folds for each of the nN partition means and two folds for each of the $nN|\mathcal{A}|$ second-moment means. A union bound and the definition of u_{p} show that, with probability at least $1-\delta_{\mathrm{p}}$, simultaneously

$$1-\eta_Z \le \frac{\widehat{Z}^{(v)}_{j\ell}}{Z_{\theta_\ell}(x_j)} \le 1+\eta_Z, \qquad (1-\eta_G)G_{j\ell}(\alpha) \le \frac{\widehat{T}^{(v)}_{j\ell}(\alpha)}{Z_{\theta_\ell}(x_j)^2} \le (1+\eta_G)G_{j\ell}(\alpha), \tag{165}$$

for both $v \in \{1,2\}$. In particular all partition estimates are positive. Dividing a second-moment mean from one fold by the squared partition estimate from the other gives (66); averaging the two fold-swapped ratios preserves the same interval.

The left inequality in (66) shows $G_{j\ell}(\alpha) \le U_{j\ell}(\alpha)$. Its right inequality gives

$$U_{j\ell}(\alpha) \le \kappa_{\mathrm{p}} G_{j\ell}(\alpha). \tag{166}$$

Because the multiplier defining U is independent of (j,ℓ,α), $\widehat{\alpha}$ minimizes $\max_{j,\ell} U_{j\ell}(\alpha)$ over $\mathcal{A}$. Comparing it with the best catalogue member proves (68). Finally, Proposition 12 multiplies the anchor UCB by at most $e^{4L_s\rho/\beta}$, proving (70). □

Proof of Corollary 14. Let E_{p} be the pilot event in Theorem 13, so $\mathbb{P}(E_{\mathrm{p}}) \geq 1-\delta_{\mathrm{p}}$. Conditional on the pilot responses and on E_{p}, the proposal, certificate, and block counts are fixed, while the fresh serving responses satisfy the independence and coverage conditions of Theorem 10. Its conditional failure probability is at most δ_{e}. Thus

$$\mathbb{P}(E_{\mathrm{p}} \cap E_{\mathrm{e}}) \geq 1 - \mathbb{P}(E_{\mathrm{p}}^c) - \mathbb{E}[\mathbf{1}_{E_{\mathrm{p}}} \mathbb{P}(E_{\mathrm{e}}^c \mid \text{pilot})] \geq 1 - \delta_{\mathrm{p}} - \delta_{\mathrm{e}}. \tag{167}$$

The same conditioning argument applies to either one-block theorem when its envelope assumptions hold. □

F.6 Supporting Proposal-Design Geometry

Proof of Proposition 25. Let $A(y) = \sum_{\ell=1}^{N} \omega_\ell p_\ell(y)^2$. For any feasible density q, Cauchy–Schwarz gives

$$\left(\int \sqrt{A}\, \mathrm{d}y \right)^2 = \left(\int \sqrt{\frac{A}{q}} \sqrt{q}\, \mathrm{d}y \right)^2 \leq \int \frac{A}{q}\, \mathrm{d}y. \tag{168}$$

Equality holds exactly when $\sqrt{A/q}$ and $\sqrt{q}$ are proportional, that is, when $q \propto \sqrt{A}$ on $\{q > 0\}$. Since $\int \sqrt{A}\, \mathrm{d}y < \infty$, the density (104) is well defined and attains the bound, which proves the claim. On the stated union support, $A > 0$ throughout. If the optimization is instead posed on a larger ambient space, assigning positive mass to $\{A = 0\}$ leaves less than unit mass for $\{A > 0\}$ and makes the Cauchy–Schwarz inequality strict; hence an optimizer assigns no mass there. □

Proof of Proposition 26. Write the primal in epigraph form, using the extended-value convention in the proposition: minimize t over $q \in \Delta(\mathcal{S})$ subject to

$$I_\ell(q) := \sum_y \frac{p_\ell(y)^2}{q(y)} \leq t, \qquad \ell \in [N]. \tag{169}$$

The uniform q together with any $t > \max_\ell I_\ell(q)$ is strictly feasible, so Slater's condition holds and $V_2 \leq \max_\ell I_\ell(q) < \infty$. Moreover, because $\mathcal{S}$ is the union support, every finite feasible point has $q(y) > 0$ for all $y \in \mathcal{S}$. Thus all Lagrangian minimizations may be restricted to the relative interior $\Delta^\circ(\mathcal{S})$, where every $I_\ell(q)$ is finite. The Lagrangian is

$$t \left(1 - \sum_{\ell=1}^{N} \lambda_\ell \right) + \sum_{\ell=1}^{N} \lambda_\ell I_\ell(q), \qquad \lambda_\ell \geq 0. \tag{170}$$

Minimization over t restricts the dual variable to Δ_N. For fixed $\lambda \in \Delta_N$, let $A_\lambda(y) = \sum_{\ell=1}^{N} \lambda_\ell p_\ell(y)^2$. Cauchy–Schwarz gives

$$\inf_{q \in \Delta^\circ(\mathcal{S})} \sum_{\ell=1}^{N} \lambda_\ell I_\ell(q) = \left(\sum_{y \in \mathcal{S}} \sqrt{\sum_{\ell=1}^{N} \lambda_\ell p_\ell(y)^2} \right)^2. \tag{171}$$

If A_λ vanishes at some points, this is an infimum: assign an arbitrarily small positive mass to those points and let it tend to zero. Thus the formula does not rely on the strictly positive weights assumed in Proposition 25. Strong duality proves (105).

It remains to justify attainment at a dual optimum. If a maximizer $\lambda^\star$ had $A_{\lambda^\star}(y_0) = 0$ at a union-support point, choose k with $p_k(y_0) > 0$ and replace $\lambda^\star$ by $(1-\epsilon)\lambda^\star + \epsilon e_k$. The y_0 term in the sum of square roots increases by order $\sqrt{\epsilon}$, whereas the total change over points where $A_{\lambda^\star} > 0$ is only order ϵ. For small ϵ this strictly increases the dual objective, a contradiction. Hence $A_{\lambda^\star} > 0$ on $\mathcal{S}$ and (106) is well defined.

Write $S = \sum_y \sqrt{A_{\lambda^\star}(y)}$, so that $q_2^\star = \sqrt{A_{\lambda^\star}}/S$ and $V_2 = S^2$. Since A_λ is linear in λ and $A_{\lambda^\star} > 0$ on $\mathcal{S}$, the dual objective is differentiable at $\lambda^\star$ with

$$\frac{\partial}{\partial \lambda_\ell} S^2 = 2S \cdot \frac{1}{2} \sum_y \frac{p_\ell(y)^2}{\sqrt{A_{\lambda^\star}(y)}} = \sum_y \frac{p_\ell(y)^2}{q_2^\star(y)} = I_\ell(q_2^\star). \tag{172}$$

No separate concavity property of S^2 is needed. Instead, note the exact identity

$$\sum_{\ell=1}^{N} \lambda_\ell^\star I_\ell(q_2^\star) = S \sum_y \frac{A_{\lambda^\star}(y)}{\sqrt{A_{\lambda^\star}(y)}} = S^2 = V_2, \tag{173}$$

which is the Euler relation for the 1-homogeneous map $\lambda \mapsto S^2$. Because $\lambda^\star$ maximizes the dual objective over Δ_N, the directional derivative along $e_\ell - \lambda^\star$ is nonpositive for every ℓ, so $I_\ell(q_2^\star) \le \sum_k \lambda_k^\star I_k(q_2^\star) = V_2$ by (173). All N terms in the $\lambda^\star$-weighted average therefore lie at or below its value V_2, which forces equality whenever $\lambda_\ell^\star > 0$. Hence $q_2^\star$ is feasible for the primal with objective at most V_2, and so is primal optimal. □

Proof of Proposition 27. For the general sandwich, fix any q and any ℓ. Cauchy–Schwarz gives $\big(\sum_y p_\ell(y)\big)^2 \le \big(\sum_y p_\ell(y)^2/q(y)\big)\big(\sum_y q(y)\big)$, i.e. $\sum_y p_\ell(y)^2/q(y) \ge 1$, so $V_2 \ge 1$. For the upper bound take $q = N^{-1}\sum_m p_m \ge N^{-1} p_\ell$, which gives $\sum_y p_\ell(y)^2/q(y) \le N \sum_y p_\ell(y) = N$ for every ℓ, hence $V_2 \le N$. The two extremal cases are as follows. If every target equals p, Cauchy–Schwarz gives $\sum_y p(y)^2/q(y) \ge (\sum_y p(y))^2 = 1$, with equality at $q = p$. For pairwise disjoint supports $S_1, \dots, S_N$, let $r_\ell = q(S_\ell)$. Cauchy–Schwarz on S_ℓ yields

$$\sum_{y \in S_\ell} \frac{p_\ell(y)^2}{q(y)} \ge \frac{1}{r_\ell}. \tag{174}$$

Since $\sum_\ell r_\ell \le 1$, some $r_\ell \le 1/N$ and the worst risk is at least N. For $q = N^{-1}\sum_\ell p_\ell$, the ratio p_ℓ/q equals N on S_ℓ, so every second moment equals N. □

Proof of Proposition 28. Under the uniform reference, (107) gives

$$p_\ell(\ell) = \frac{e^{B/\beta}}{e^{B/\beta} + N - 1} = 1 - \eta, \qquad p_\ell(y) = \frac{1}{e^{B/\beta} + N - 1} = \frac{\eta}{N-1} \tag{175}$$

for $y \ne \ell$. Let $c_\eta = (1-\eta)^2 + \eta^2/(N-1)$. For any proposal q,

$$\max_\ell \sum_y \frac{p_\ell(y)^2}{q(y)} \ge \frac{1}{N}\sum_{\ell=1}^{N}\sum_{y=1}^{N} \frac{p_\ell(y)^2}{q(y)} \tag{176}$$

$$= \frac{c_\eta}{N}\sum_{y=1}^{N} \frac{1}{q(y)} \ge N c_\eta, \tag{177}$$

where the last inequality is Cauchy–Schwarz. The uniform proposal makes every target second moment exactly $N c_\eta$, proving (109). The stated limit and score scaling follow from (107). □

Proof of Proposition 29. Since $q_{\mathrm{env}} = \bar{p}/C$ and $p_\ell \le \bar{p}$, $p_\ell/q_{\mathrm{env}} \le C$ for every ℓ. Because N is finite, $\bar{p}$ is a pointwise maximum of finitely many densities, so $\bigcup_\ell \{p_\ell = \bar{p}\}$ exhausts the union support and some ℓ attains $p_\ell/q_{\mathrm{env}} = C$; the maximum over ℓ therefore equals C. Conversely, a proposal q whose worst-target envelope is at most C' has $p_\ell(y) \le C' q(y)$ at every y and every ℓ, so $\bar{p} \le C' q$ and integration gives $C \le C'$. Finally, $C \le \sum_\ell \int p_\ell(y)\, \mathrm{d}y = N$. □

Proof of Proposition 30. Fix j and θ, and let $s_{j,\theta}^-$ and $s_{j,\theta}^+$ be the infimum and supremum of the score on the common support S_j from (40). By that equation, $s_{j,\theta}^+ - s_{j,\theta}^- \le B$, while

$$Z_\theta(x_j) \ge e^{s_{j,\theta}^-/\beta}. \tag{178}$$

Therefore

$$\frac{\pi_\theta^\star(y \mid x_j)}{\pi_0(y \mid x_j)} = \frac{e^{s_\theta(x_j,y)/\beta}}{Z_\theta(x_j)} \le e^{B/\beta}. \tag{179}$$

The defensive inequality $q_\alpha \ge a\pi_0$ now yields $\pi_\theta^\star/q_\alpha \le e^{B/\beta}/a$. Because the same S_j works for all θ and $\pi_0/q_\alpha = 0$ off S_j, the envelope holds at every y with $q_\alpha(y) > 0$. For the second moment, note that $q_\alpha \ge a\pi_0$ with $a > 0$ makes q_α cover both $\pi_0(\cdot \mid x_j)$ and $\pi_\theta^\star(\cdot \mid x_j)$, so U_θ is a genuine density ratio with $\mathbb{E}_{q_\alpha} U_\theta = 1$. Writing $\Lambda_\theta := \sup_{y:\, q_\alpha(y|x_j)>0} U_\theta(x_j, y)$, the conclusion then follows from $\mathbb{E}_{q_\alpha} U_\theta^2 \le \Lambda_\theta\, \mathbb{E}_{q_\alpha} U_\theta = \Lambda_\theta$. □

Proof of Proposition 31. At a response with $\widetilde{q}(y) = 0 < p(y)$ both sides are $+\infty$ by the convention of Section B. At every other response, $q(y) \geq c\widetilde{q}(y)$ gives $p(y)/q(y) \leq c^{-1}p(y)/\widetilde{q}(y)$. Taking the supremum proves the second claim, and multiplying by $p(y)$ and integrating proves the first. □

Proof of Proposition 32. Condition on the pilot responses. The selected proposal is then fixed, while the serving candidates retain the independence and distribution required by Assumption 6. The proof of the invoked theorem can therefore be applied under the conditional law of the fresh serving pool, including the block theorem when only a conditional second-moment bound is available. Its conditional failure probability is at most δ; taking expectation over the pilot data gives the same unconditional failure probability. □

F.7 Main Support Obstruction and Fixed-Proposal Lower Bound

Proof of Proposition 15. Let E be the event that none of the M candidates lies in A. Independence gives $\mathbb{P}(E) = (1 - r_\star)^M$. On E, every candidate-supported policy assigns zero mass to A, whereas the target assigns mass $m_\star$. The expectation error for $f = \mathbf{1}_A$ is therefore $m_\star$. Requiring $(1 - r_\star)^M \leq \delta$ and solving for M gives (72).

For the divergence relation, the claim is immediate if q fails to cover $\pi^\star_\theta$, under the convention in the proposition. Otherwise, Cauchy–Schwarz on A gives

$$m_\star^2 = \left(\int_A \frac{\pi^\star_\theta(y)}{q(y)}\, q(y)\, \mathrm{d}y \right)^2 \leq q(A) \int_A \left(\frac{\pi^\star_\theta(y)}{q(y)} \right)^2 q(y)\, \mathrm{d}y \leq r_\star\{1 + \chi^2(\pi^\star_\theta \| q)\}. \tag{180}$$

Rearranging proves (73). □

Proof of Theorem 16. Let $m = 2\epsilon$, let $r = 4\epsilon^2/(\overline{\Gamma} - 1)$, and take $\mathcal{Y} = \{0, 1\}^N$ with

$$q = \pi_0 = \bigotimes_{k=1}^{N} \mathrm{Bern}(r). \tag{181}$$

The assumptions imply $0 < r \leq 1/4$ and $0 < m \leq 1/2$. For each $\ell \in [N]$, define the product target

$$p_\ell = \mathrm{Bern}(m)_\ell \otimes \bigotimes_{k \neq \ell} \mathrm{Bern}(r)_k, \qquad f_\ell(y) = y_\ell, \qquad \mathcal{F} = \{f_1, \ldots, f_N\}. \tag{182}$$

Its density ratio depends only on coordinate ℓ and equals

$$u_\ell(y) = \left(\frac{m}{r} \right)^{y_\ell} \left(\frac{1 - m}{1 - r} \right)^{1 - y_\ell}. \tag{183}$$

The finite score $s_\ell(y) = \beta \log u_\ell(y)$ has $Z_\ell = \mathbb{E}_q u_\ell = 1$, so its Gibbs tilt of the common full-support reference is exactly p_ℓ. Its two score values differ by

$$\beta \log \frac{m(1 - r)}{r(1 - m)} = \beta \log \frac{(\overline{\Gamma} - 1)(1 - r)}{2\epsilon(1 - 2\epsilon)}, \tag{184}$$

which proves (77). Direct calculation also gives

$$\mathbb{E}_q u_\ell^2 = \frac{m^2}{r} + \frac{(1 - m)^2}{1 - r} = \overline{\Gamma} - 1 + \frac{(1 - m)^2}{1 - r}. \tag{185}$$

Since $r \leq 4\epsilon^2 \leq 4\epsilon - 4\epsilon^2 = 1 - (1 - m)^2$, the last fraction lies in $(0, 1]$, proving (74).

Let E_ℓ be the event that coordinate ℓ is zero in every one of the M candidates. The product structure makes the events $E_1, \ldots, E_N$ independent and $\mathbb{P}(E_\ell) = (1 - r)^M$. On E_ℓ, every distribution supported on the realized candidates has expectation zero for f_ℓ, whereas $p_\ell f_\ell = m = 2\epsilon$. Randomization cannot alter this support obstruction. Therefore simultaneous error at most ϵ implies $\bigcap_{\ell=1}^N E_\ell^c$, which proves (75).

If the left-hand probability in (75) is at least $1 - \delta$, solving $[1 - (1 - r)^M]^N \geq 1 - \delta$ gives the first inequality in (76). For $\delta \leq 1/2$,

$$(1 - \delta)^{1/N} \geq e^{-2\delta/N} \geq 1 - \frac{2\delta}{N}, \tag{186}$$

so its logarithmic numerator is at least $\log(N/(2\delta))$. Since $r \le 1/4$, $-\log(1-r) \le r/(1-r) \le 4r/3$. Substituting the definition of r proves the second inequality; the third follows from $\overline{\Gamma} - 1 \ge \overline{\Gamma}/2$. □

G EXPERIMENTS

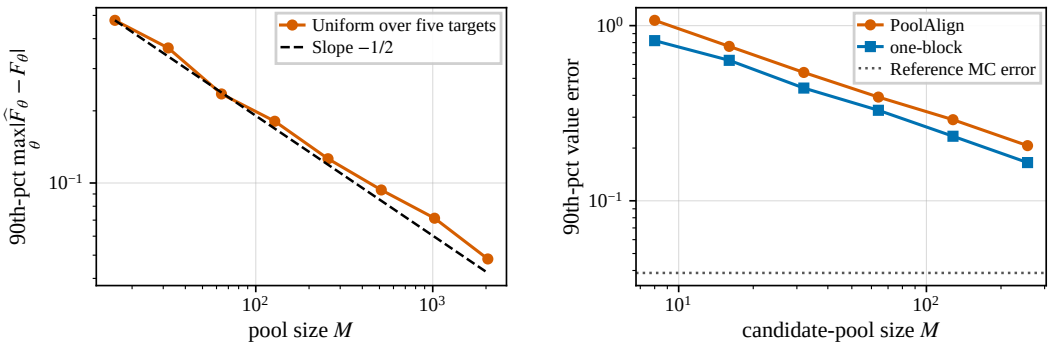
In this section, we detail the experiments summarized in Section 6, which test effective-budget scaling, robust upper-tail aggregation, independent proposal selection, and shared language-model candidate pools. The comparisons against clipped SNIS and BR-SNIS, the split-reference diagnostics, the block-count settings, and the downstream win-rate are reported here rather than in the main text. The controlled studies use exact target quantities, while the language-model study uses an independent large-pool proxy and reports its split-reference discrepancy.

Protocol and reporting conventions. Throughout, a *uniform* error is the deviation from the target maximized over the preference family and, for functional endpoints, over the evaluation class $\mathcal{F}$, as defined in Eqs. (2) and (3). One *repetition* draws a fresh candidate pool of size M for each prompt and reruns every estimator on that same pool, which keeps the pool fixed across methods and across preferences exactly as the post-generation protocol requires. The reported statistic is an upper quantile of the resulting error over repetitions, and two levels appear in this paper. The 90th percentile is a moderate-envelope quantile that summarizes typical behavior, whereas $\delta = 10^{-3}$ is the deep-tail level that Theorem 2 actually bounds through its $\log(1/\delta)$ dependence. The two levels separate precisely where the envelope Γ_∞ matters, which is why the aggregate curves of Section G.3 read the 90th percentile while Figures 2a and 2b read the deep tail. For the language-model study the exact target expectation is unavailable, and an independent reference pool of $M_{\mathrm{ref}} = 4096$ responses per prompt, formed from two disjoint halves of 2048, supplies the evaluation target. The same halves supply the split-reference discrepancy, which lower-bounds the error any estimator can resolve against this proxy. The three panels of Figure 2 differ in instance, repetition count, and reported statistic.

- Figure 2a uses Llama-3.1-8B-Instruct pools on HH-RLHF prompts, with seven ArmoRM helpfulness-safety scalarizations read over six tilt strata. One repetition resamples a serving pool from the stored candidate bank, 30,000 times per prompt–preference–budget cell over 60 prompts, and the panel reports the $\delta = 10^{-3}$ upper quantile of the uniform functional error (Section G.3).
- Figure 2b uses Pareto-tailed importance weights with tail index α and a single tail-aligned functional against an exact target. One repetition draws a fresh pool of size $M = 512$, 4×10^4 times per α, and the panel reports the ratio of $\delta = 10^{-3}$ upper quantiles with a 2,000-draw bootstrap interval (Section G.1).
- Figure 2c uses a single-prompt Gaussian tilt family with five preferences and exact targets. One repetition draws a fresh pool at each budget, 300 times, and the panel reports the 90th percentile of the functional error (Section G.1).

G.1 CANDIDATE COMPLEXITY AND ROBUST AGGREGATION

Candidate complexity. Let $q = \pi_0 = \mathcal{N}(0,1)$, $s_\theta(y) = \theta y$, and $\beta = 1$. Then $\pi_\theta^\star = \mathcal{N}(\theta, 1)$ and $\Gamma_2(\theta) = \exp(\theta^2)$. We estimate $F_\theta = \mathbb{E}_{\pi_\theta^\star}[\tanh(Y)]$ for five tilts from the same M candidates in each repetition. For confidence level 0.9, we set $\mathfrak{L}_{\mathrm{P}} = \log(800)$ and use the common block count $K = \mathrm{round}\,\sqrt{M\mathfrak{L}_{\mathrm{P}}/\max_\theta \Gamma_2(\theta)}$. Across 300 repetitions, the fitted log–log slopes of the marginal 90th-percentile errors are $-0.51, -0.51, -0.49, -0.48, -0.47$ for $\theta = 0.5, 0.75, 1, 1.25, 1.5$; the 90th percentile of the simultaneous maximum has slope -0.47. The uniform error over the five targets decays at the predicted square-root rate (Figure 3a), and plotting the marginal errors against M/Γ_2 aligns the five curves (Figure 2c).



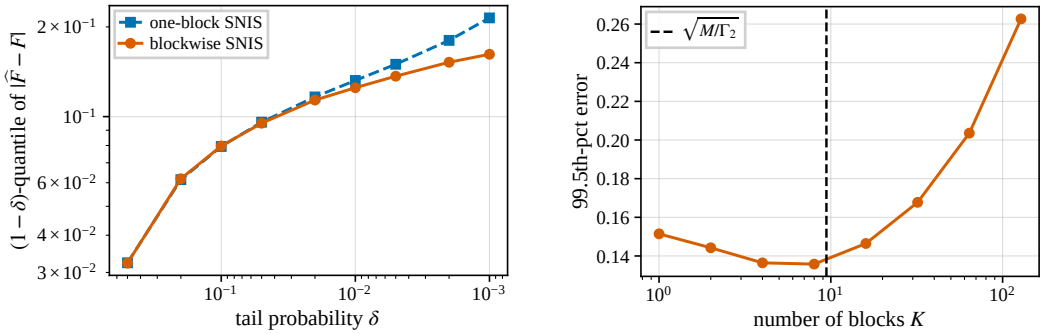
(a) Functional-error rate.

(b) Uniform value error on LM pools.

Figure 3: Exact-target functional-error rate and the uniform value error on language-model pools.

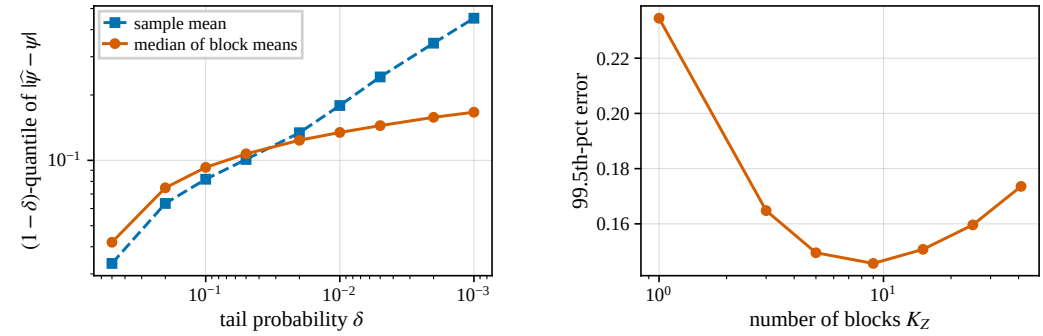
Upper-tail behavior. To separate second-moment control from a bounded-weight assumption, take $V \sim \mathrm{Unif}(0,1)$ and $U = ((\alpha-1)/\alpha)V^{-1/\alpha}$ with $\alpha = 2.1$. This defines a target density Uq with $\Gamma_2 \approx 5.76$, while Γ_∞ and the third moment are infinite. For functional estimation we use the tail-aligned $f(V) = \mathbf{1}\{V < 1/2\} - \mathbf{1}\{V \geq 1/2\}$; for partition estimation we use raw weights eU so that $\psi = 1$. At $M = 512$, 4×10^4 repetitions estimate each $(1-\delta)$-quantile in Figure 4, using $K = 8$ for PoolAlign and $K_Z = 9$ for the log partition. PoolAlign and one-block SNIS agree at the median, while the observed one-block-to-blockwise error ratio reaches 1.32 at $\delta = 10^{-3}$. For the log-partition, the sample-mean-to-MoM ratio changes from 0.80 at the median to an observed 2.71 at $\delta = 10^{-3}$. The block sweeps show the finite-sample sensitivity of K and K_Z; their minima occur near $K = 4$–8 and $K_Z = 9$ in this instance.

Tail-index sweep. Since the result above is one instance, we repeat the functional experiment over the tail index $\alpha \in \{2.05, 2.1, 2.25, 2.5, 3.0\}$ at $M = 512$, $K = 8$, with 4×10^4 repetitions each, and report the ratio of the one-block to blockwise errors at $\delta = 10^{-3}$ with a 2,000-draw bootstrap interval (Figure 2b). As α decreases toward 2, Γ_2 grows (1.33 at $\alpha = 3$ to 10.76 at $\alpha = 2.05$) and the ratio rises from 1.004 (95% interval $[0.98, 1.02]$, i.e., the two estimators coincide at the lighter-tailed end of the considered Pareto family) to 1.446 ($[1.29, 1.55]$) at $\alpha = 2.05$: the blockwise gain grows with the heaviness of the tail and approaches parity for the lighter-tailed members of the family. The bootstrap interval excludes 1 for every $\alpha \leq 2.5$ but not at $\alpha = 3.0$, so the heavy-tail advantage is significant precisely when Γ_2 is large.



(a) SNIS functional error.

(b) SNIS block-count sweep.



(c) Log-partition error.

(d) MoM block-count sweep.

Figure 4: Robust aggregation under finite second moment and unbounded importance weights. The gains appear in high-confidence quantiles rather than at the median.

G.2 Independent-Pilot Proposal Selection

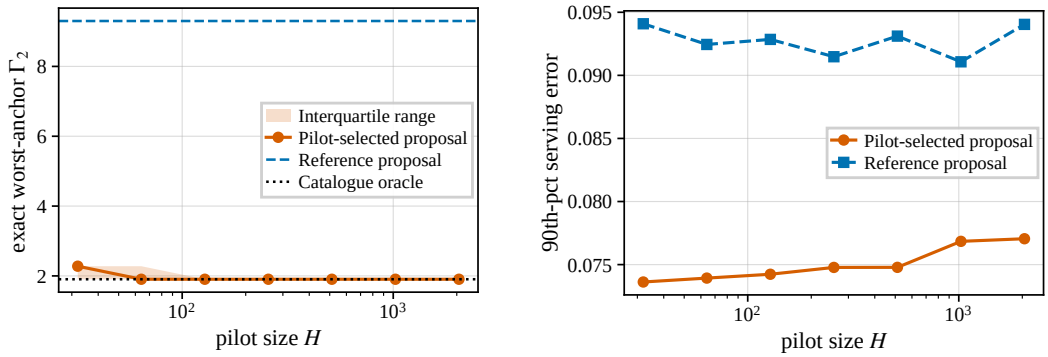
We next implement the cross-fitted selector in (48). The response space is a 101-point grid; π_0 is a discretized $\mathcal{N}(0,1)$ law, the seven anchors range from -1.5 to 1.5, and two auxiliary proposal components are discretized normals centered at -1.5 and 1.5. The fixed catalogue contains π_0 and seven defensive mixtures with reference mass 0.2. For each H, we run 600 independent pilot trials and evaluate the selected proposal using exact worst-anchor coverage and a fresh $M = 256$ serving pool.

The pilot law is the symmetric defensive mixture $q_{\mathrm{pil}} = 0.2\pi_0 + 0.4q_{\mathrm{left}} + 0.4q_{\mathrm{right}}$; in this symmetric instance it coincides with the catalogue oracle, isolating the finite-sample proposal-identification error.

The reference proposal has worst-anchor coverage 9.30, whereas the catalogue oracle has coverage 1.90. Over the 600 trials at each pilot budget, the oracle-selection frequency rises from 0.39 (95% Wilson interval $[0.35, 0.43]$) at $H = 32$ to 0.92 ($[0.90, 0.94]$) at $H = 256$ and 1.00 ($[0.99, 1.00]$) at $H = 2048$; the median selected coverage reaches the oracle value by $H = 64$ (Figure 5a). At $H = 256$, on independent serving pools, the selected proposal reduces the 90th-percentile worst-anchor functional error from 0.091 to 0.075, a reduction of 0.017 with 95% prompt-bootstrap interval $[0.011, 0.024]$ excluding zero (Figure 5b).

Since the certificate in Theorem 13 is an upper confidence bound on the worst-anchor coverage of the selected proposal, we also measure its empirical validity. The inflation factor $(1+\eta_Z)^2/(1-\eta_G)$ is constant across the catalogue, so it leaves the selector in (48) unchanged but scales the reported certificate. At the tolerance pair $(\eta_Z, \eta_G) = (0.1, 0.1)$, the inflated value $U_{j\ell}(\widehat{\alpha})$ exceeds the exact worst-anchor coverage of the selected proposal in all 600 trials at every pilot size, i.e., an empirical coverage frequency of 1.000 against the nominal $1 - \delta_{\mathrm{p}}$; at $(\eta_Z, \eta_G) = (0.05, 0.05)$ the frequency is 0.978 at $H = 32$ and reaches 1.000 from $H = 128$ onward. The certificate is thus not violated on this instance.

The certified budget (65) is stated through a worst-case envelope: on this instance the score oscillation is $B = 15$ on the grid with $\beta = 1$, so the envelope constants in (64) are $L_Z = e^{15}/0.2 \approx 1.6 \times 10^7$ and $L_G = e^{30}/0.04 \approx 2.7 \times 10^{14}$, placing the sufficient budget well above the pilot sizes $H \in \{32, \dots, 2048\}$ studied here. The empirical behavior is governed instead by the actual second-moment errors, which are far smaller than the worst-case envelope: the certificate holds and the oracle is recovered across every pilot size tested, so the selector reaches its target regime well before the conservative sufficient budget of (65) is met. Reducing B through reward normalization is the direct route to tightening the constant.



(a) Exact coverage of the selected proposal.

(b) Error on fresh serving candidates.

Figure 5: Finite-catalogue selection with independent pilot and serving samples. Increasing H recovers the catalogue oracle and the resulting coverage reduction transfers to serving error.

G.3 Language-Model Candidate Pools

The controlled studies above separate the theoretical mechanisms using exact targets. On model-generated responses the exact target expectation is unavailable, so we measure the candidate-budget behavior against a large-pool reference.

Setup. We sample responses from Llama-3.1-8B-Instruct (Grattafiori et al., 2024) on 200 prompts drawn from the HH-RLHF test split (Bai et al., 2022). From each selected `chosen` example, we take the first user turn, truncate it to 1,000 characters, and apply the Llama chat template. ArmoRM-Llama3-8B (Wang et al., 2024b) supplies helpfulness and safety reward coordinates. For every prompt, we independently generate a calibration pool of size $M_{\text{cal}} = 256$, two reference pools of size 2048, and three serving pools of size 256. All pools come from the same preference-agnostic proposal π_0 before the preference is revealed. The serving budgets $M \in \{8, 16, 32, 64, 128, 256\}$ are nested prefixes of each serving pool. Generation uses temperature 1, top-$p = 1$, no top-k truncation, and at most 96 new tokens.

The calibration pool fixes the mean and standard deviation of each reward coordinate; the resulting affine transformation is then held fixed for all reference and serving responses. Writing the standardized coordinates as r_{help} and r_{safe}, the seven preferences define

$$s_\theta(y) = \theta r_{\text{help}}(y) + (1-\theta) r_{\text{safe}}(y), \qquad \theta \in \left\{0, \frac{1}{6}, \frac{2}{6}, \dots, 1\right\}, \qquad \beta = 1.$$

Policy evaluation uses the three bounded functionals

$$\mathcal{F}_{\text{LM}} = \{\tanh(r_{\text{help}}),\ \tanh(r_{\text{safe}}),\ \text{clip}(2L/96 - 1, -1, 1)\},$$

where L is the tokenized response length.

Algorithms and endpoints. The language-model instantiation of PoolAlign uses the predeclared count $K = 4$ for every prompt, preference, and serving budget; its log-partition estimator uses $K_Z = 5$. The one-block comparator is ordinary SNIS for bounded functionals and the sample mean for the partition function. Best-of-n returns the highest-scoring serving response. We include it only as a target-mismatch diagnostic: it does not estimate the Gibbs target or its log partition.

The two independent reference halves are concatenated before computing the primary $M_{\mathrm{ref}} = 4096$ SNIS and log-partition quantities, using one pooled normalizer. We reserve the halves for the finite-reference diagnostic $\frac{1}{2}|\widehat{T}_{\mathrm{ref},a} - \widehat{T}_{\mathrm{ref},b}|$. Thus the reported deviations are relative to the pooled large-sample proxy, rather than an exact target.

For method a, prompt j, and serving replicate r, we form

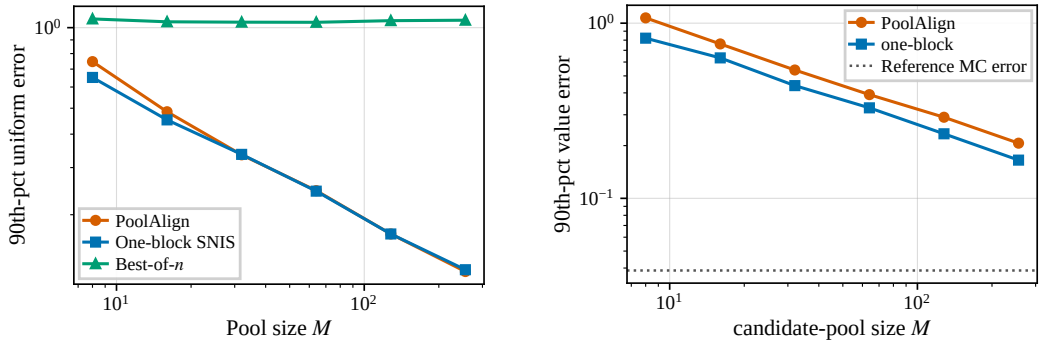
$$E^{(a)}_{F,jr}(M) = \max_{\theta\in\Theta,\, f\in\mathcal{F}_{\mathrm{LM}}} \left|\widehat{\pi}^{(a)}_{jr,\theta,M} f - \widehat{F}^{\mathrm{ref}}_{j,\theta,f}\right|, \qquad E^{(a)}_{\psi,jr}(M) = \max_{\theta\in\Theta} \left|\widetilde{\psi}^{(a)}_{jr,\theta,M} - \widehat{\psi}^{\mathrm{ref}}_{j,\theta}\right|.$$

We summarize the 200×3 prompt–serving-pool cells. Figure bands are pointwise 95% percentile intervals from 2,000 prompt-cluster bootstrap draws; each draw resamples prompts and retains all three serving replicates.

Results. For PoolAlign, the 90th-percentile uniform functional deviation decreases from 0.747 at $M = 8$ to 0.123 at $M = 256$; the one-block values are 0.652 and 0.125 (Figure 2a), so the two track one curve at the moderate Γ_2 these pools realize, consistent with the shared square-root rate in Theorem 2. The Best-of-n diagnostic remains near 1.05–1.08. The corresponding PoolAlign log-partition deviations decrease from 1.072 to 0.206 and the sample-mean values from 0.821 to 0.165, again on a common rate (Figure 6b); the median $\Gamma_2 \approx 3.13$ here places both in the shared-rate regime, and the median-of-means advantage is the deep-tail effect the controlled study of Section G.1 isolates.

On the predeclared fitting range $M \in \{16, 32, 64, 128\}$, the PoolAlign functional slope is -0.500 with 95% prompt-cluster bootstrap interval $[-0.532, -0.474]$; its log-partition slope is -0.464 with interval $[-0.513, -0.420]$. Both intervals contain $-1/2$ and lie below zero. The one-block slopes are -0.470 ($[-0.497, -0.433]$) and -0.474 ($[-0.508, -0.419]$), respectively, on the same bootstrap. The 90th-percentile split-reference discrepancies are 0.033 ($[0.031, 0.037]$) for the functional endpoint and 0.039 ($[0.036, 0.046]$) for the log partition, below every reported method curve. Across prompts, the pooled-reference estimate of the worst-preference Γ_2 has median 3.13, 90th percentile 5.14, and maximum 50.98. No policy or value block has zero total weight.

Figure 6a reports the functional curve and Figure 6b the value curve; Figure 2a in Section 6 reads the same pools at the deep-tail confidence level.



(a) Uniform functional error.

(b) Uniform value error.

Figure 6: **Language-model candidate pools.**

Heavy-tail stratification and the clipped-SNIS baseline. The aggregate curves above pool prompts of every difficulty. To locate where blockwise normalization acts and to benchmark it against a standard heavy-tail remedy, we generate one preference-agnostic pool per prompt and sweep the tilt temperature over a β ladder $\{1.0, 0.7, 0.55, 0.45, 0.375, 0.3\}$, recomputing the Gibbs weights on the *same* stored candidates at each β. A smaller β sharpens the tilt and raises the realized second moment $\mathbb{E}_{y\sim q_j}[U_\theta(x_j, y)^2]$, giving six operating points whose stratum medians over prompt–preference pairs are $1.8, 3.1, 5.4, 9.7, 16.1, 32.1$. At each we compare three estimators of the upper-tail functional error: one-block (naive), clipped (winsorized) self-normalized importance sampling at the 95% weight quantile, and blockwise (ours). Clipping stands in for the weight-modification

family as a whole; Pareto-smoothed importance sampling (Vehtari et al., 2024) replaces the upper weight tail by a fitted generalized Pareto quantile rather than truncating it, and like clipping returns a modified weight vector, so it sits on the same side of the comparison. A cell whose effective sample size collapses toward a single dominant candidate is degenerate for every weighting scheme, so we condition on cells clearing a fixed effective-sample floor of eight, applied identically to all three estimators; this retains $91{,}446$ paired functional cells (prompt $\times$ preference $\times$ budget $\times$ replicate $\times$ β). The floor is a common well-definedness requirement, not a filter tuned to any method, and we report its effect across floors below.

Against clipped SNIS, a standard heavy-tail remedy, blockwise normalization attains a lower 90th-percentile error at every Γ_2 stratum. The clipped-to-blockwise tail-error ratio is 1.078 (95% prompt-cluster bootstrap interval $[1.04, 1.14]$) at $\Gamma_2 = 1.8$ and stays between 1.08 and 1.10 across the ladder, reaching 1.083 ($[1.05, 1.12]$) at the heaviest stratum $\Gamma_2 = 32.1$; all six intervals exclude one (Figure 7). At this moderate-envelope quantile one-block normalization instead holds a small but significant constant edge over blockwise: its tail-error ratio to blockwise runs from 0.961 ($[0.94, 0.98]$) at $\Gamma_2 = 1.8$ to 0.911 ($[0.89, 0.95]$) at $\Gamma_2 = 32.1$, all six prompt-cluster intervals excluding one—the shared $\sqrt{\Gamma_2/M}$ rate with a lighter constant that Theorem 2 predicts when Γ_∞ stays finite; the regime where the two separate in blockwise's favor is the deep tail, which the controlled Pareto sweep isolates (Figure 2b) by driving Γ_2 high directly through the tilt until the $\delta = 10^{-3}$ ratio reaches 1.446. Reading the same real pools at that deep-tail confidence level confirms the picture directly on the data (Table 3): the clipped-to-blockwise $\delta = 10^{-3}$ tail-error ratio stays between 1.05 and 1.10 across the reward functionals at every stratum, with all six prompt-cluster intervals excluding one, so the advantage the theorem attributes to the deep tail is present on the real geometry and not an artifact of the synthetic family. Clipping trades bias for a variance cap and pays the bias on these real pools, while blockwise normalization caps no weights and needs no such tuning.

Comparison to BR-SNIS, the nearest prior method. BR-SNIS (Cardoso et al., 2022) is the estimator we place closest to ours in Table 1, and it optimizes a different axis: it reduces the $O(M^{-1})$ bias of self-normalization through iterated sampling-importance-resampling, whereas blockwise normalization targets the tail of the pool-draw error. Because the serving pool is fixed before preferences are revealed, we run the fixed-pool bootstrap i-SIR variant of Cardoso et al. (2022)—resampling candidates with replacement from the frozen bank—at the budget-matched setting of $T = K$ iterations, so that the two methods spend the same M candidate evaluations. Read at $\delta = 10^{-3}$ on the same above-floor cells, blockwise normalization has the lower deep-tail error at every stratum: the BR-SNIS-to-blockwise ratio rises from 1.01 ($[1.01, 1.02]$) at $\Gamma_2 = 2.1$ to 1.04 ($[1.02, 1.07]$) at $\Gamma_2 = 70.5$, all six prompt-cluster intervals excluding one (Table 3). The two axes are visibly orthogonal: BR-SNIS does reduce the central bias at light tails (its bias is 0.99 that of one-block SNIS at $\Gamma_2 = 2.1$), but the reduction reverses as the tilt sharpens (bias ratio 1.29 at $\Gamma_2 = 70.5$), because its guarantee, like clipped SNIS, rests on the bounded-weight envelope (Table 1, Assumption A1 of Cardoso et al. (2022)) that fails precisely in the deep tail—the assumption blockwise normalization removes. The gap is also monotone in the i-SIR iteration count under budget matching: sweeping $K = T \in \{2, 4, 8, 16\}$, the median BR-SNIS-to-blockwise ratio grows from 0.99 ($[0.99, 1.00]$) at K=2, where a two-step chain is close to one-block SNIS, through 1.03 ($[1.02, 1.04]$) at the declared K=4 and 1.08 ($[1.07, 1.11]$) at K=8 to 1.18 ($[1.15, 1.21]$) at K=16, the four prompt-cluster intervals ordered without overlap, as a longer frozen chain lets a single tail candidate persist across iterations rather than being averaged out across disjoint blocks. Blockwise normalization wins the deep tail against the nearest prior method without inheriting its envelope condition.

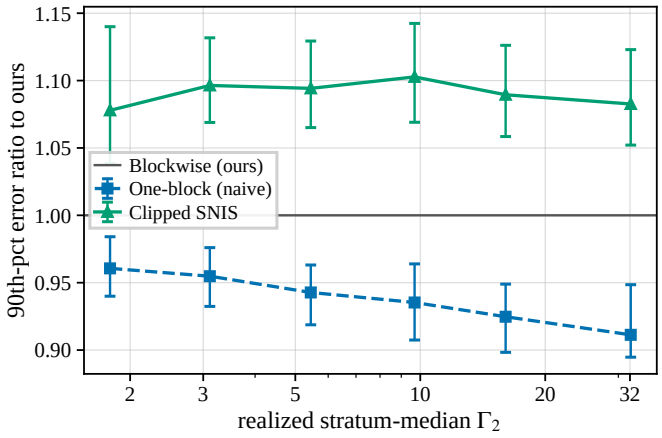


Figure 7: **Real-pool heavy-tail sweep against the clipped-SNIS baseline.** 90th-percentile functional error of one-block (naive), clipped (winsorized) self-normalized importance sampling, and blockwise (ours) estimators on the Llama-3.1/ArmoRM pools, as the tilt temperature drives the stratum-median second moment from 1.8 to 32.1; error bars are 95% prompt-cluster bootstrap intervals. Blockwise normalization is below clipped SNIS at every stratum (clipped-to-blockwise ratio 1.08–1.10, all intervals excluding one); blockwise and one-block track a common rate at this moderate-envelope quantile, their separation being the deep-tail effect isolated in Figure 2b. Figure 2a reads the same ladder at $\delta = 10^{-3}$ against the two envelope-based remedies.

The deep tail read directly on real pools. The 90th-percentile comparison above is a moderate-envelope quantile; the object Theorem 2 actually bounds is the high-confidence tail, whose $\log(1/\delta)$ dependence controls the δ-upper quantile of the estimator error over pool draws. We therefore read the real pools at $\delta = 10^{-3}$: holding the reference target fixed from the independent $M_{\mathrm{ref}} = 4096$ slice, we bootstrap 30,000 serving pools for each prompt–preference–budget cell over the 60 prompts of this sweep and record $q_{1-\delta}(|\hat{\mu} - \mu_{\mathrm{ref}}|)$ for each estimator, sweeping the same β ladder. Resampling from the stored bank, rather than regenerating responses, is what makes 30,000 repetitions per cell affordable at this confidence level. A finite candidate bank caps its own realized tail, so this read is conservative relative to the fresh-sampling Pareto sweep. One qualification is required and is the exact analogue of the effective-sample floor of Table 4: a cell whose blockwise deep-tail error does not clear its own independent split-reference gap is degenerate for *every* estimator—the reference expectation is then buried in its finite-M_{ref} noise—so we condition on cells clearing that gap, a well-definedness test computed before any comparison and applied identically to all three estimators (it retains 5,266 of 6,480 cells, and all 4,304 reward-functional cells but only the near-constant length functional heavily). On this set the deep-tail blockwise advantage over clipped SNIS is present at every stratum (Table 3): the reward-functional clipped-to-blockwise $\delta = 10^{-3}$ ratio rises from 1.05 at $\Gamma_2 = 2.1$ to a peak of 1.098 ($[1.079, 1.114]$) at $\Gamma_2 = 15.8$ and holds 1.061 ($[1.042, 1.084]$) at the heaviest stratum, all six prompt-cluster intervals excluding one; pooling all functionals gives the same conclusion at five of six strata. One-block normalization, by contrast, stays at or below blockwise throughout ($R_{\mathrm{naive}} \leq 0.99$), the finite-bank shadow of the same tail effect that Figure 2b drives above one only under fresh heavy-tailed sampling.

Table 3: Deep-tail ($\delta = 10^{-3}$) real-pool comparison across Γ_2 strata, on cells clearing the split-reference well-definedness gap (Table 4's floor analogue). Each entry is a ratio of δ-upper-quantile functional error with a 95% prompt-cluster bootstrap interval; $R > 1$ favors blockwise. Reward functionals (helpfulness, harmlessness) are the regime in which the deep-tail signal clears the reference floor; the clipped-to-blockwise and BR-SNIS-to-blockwise ratios both exclude one at every stratum. BR-SNIS runs its fixed-pool bootstrap i-SIR variant at the budget-matched setting $T = K$ iterations.

stratum-median $\mathbb{E}_q[U_\theta^2]$	2.1	4.2	8.7	15.8	30.8	70.5
clipped / blockwise (ours)	1.05	1.08	1.09	1.10	1.08	1.06
95% interval	[1.04, 1.06]	[1.06, 1.09]	[1.08, 1.11]	[1.08, 1.11]	[1.07, 1.10]	[1.04, 1.08]
BR-SNIS / blockwise (ours)	1.01	1.03	1.03	1.03	1.02	1.04
95% interval	[1.01, 1.02]	[1.02, 1.04]	[1.02, 1.05]	[1.02, 1.05]	[1.01, 1.05]	[1.02, 1.07]
one-block / blockwise (ours)	0.99	0.99	0.98	0.97	0.97	0.96
95% interval	[0.98, 1.00]	[0.98, 0.99]	[0.97, 0.99]	[0.96, 0.98]	[0.95, 0.97]	[0.95, 0.98]

Invariance to the effective-sample-size floor. Since the well-definedness floor also fixes how well-conditioned the retained cells are, we verify that the comparison does not depend on its value by repeating the entire sweep at floors $\{4, 8, 16, 32\}$ (Table 4). The conclusion is invariant: blockwise normalization is below clipped SNIS at every stratum and every floor, with all 95% prompt-cluster bootstrap intervals excluding one. The clipped-to-blockwise margin only widens as the floor rises, from 1.07–1.11 at floors 4–8 to 1.08–1.28 at floor 32: a higher floor retains lighter-tailed cells, on which the winsorization in clipped SNIS caps almost nothing and acts as pure bias, while blockwise normalization caps no weights and pays none. The floor of eight reported above is thus the conservative end of the observed margin.

Table 4: Heavy-tail stratified comparison across effective-sample-size floors. Ranges span the six Γ_2 strata; each entry is a ratio of 90th-percentile functional error. Blockwise normalization stays below clipped SNIS at every floor (every clipped-to-blockwise stratum interval excludes one), its margin widening as the floor retains lighter cells. At this moderate envelope one-block SNIS carries a small constant edge over blockwise (4–9%), consistent with the shared $\sqrt{\Gamma_2/M}$ rate in Theorem 2 rather than identical constants; the blockwise advantage is confined to the deep tail that the Pareto family of Figure 2b isolates.

effective-sample floor	4	8	16	32
paired cells retained	103,926	91,446	71,685	45,753
clipped / blockwise (ours)	1.07–1.11	1.08–1.10	1.07–1.15	1.08–1.28
one-block / blockwise (ours)	0.95–0.96	0.91–0.96	0.91–0.96	0.93–0.97

Sensitivity to the clipping quantile. Clipped SNIS has one free parameter, the weight quantile at which it winsorizes, and the comparison above fixes it at 95%. Because that choice governs how much of the tail the baseline discards, we sweep it over $q \in \{0.80, 0.90, 0.95, 0.99\}$ and recompute the full stratified comparison at each, holding blockwise normalization fixed (Figure 8). Each (q, β) point aggregates 18,900 functional cells. The clipped-to-blockwise tail-error ratio rises monotonically as the clip tightens: at $q = 0.80$ it is 1.55–1.87 across the β ladder, at $q = 0.90$ it is 1.23–1.38, and at $q = 0.95$ it is 1.05–1.12 with every stratum interval excluding one, the operating point of Figure 7. The ratio falls below one only at $q = 0.99$, where winsorization touches at most the single largest weight, so the baseline reduces to one-block SNIS and the comparison no longer isolates clipping. The blockwise advantage over clipped SNIS is therefore not an artifact of the reported clip level; it holds at every quantile that actually caps the tail and grows as the cap tightens, and the 95% setting reported above is the conservative end of that range.

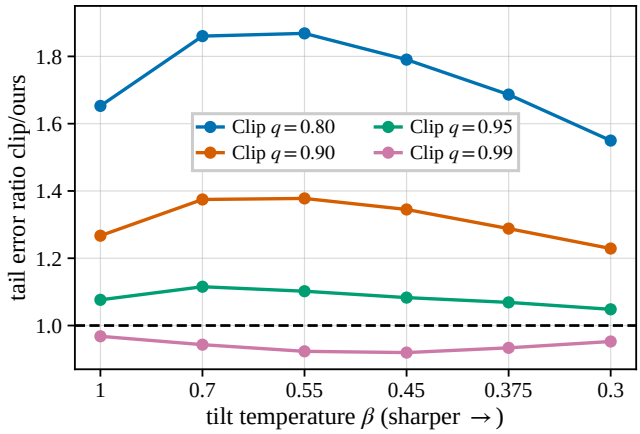


Figure 8: **Blockwise versus clipped SNIS across the clipping quantile.** Clipped-to-blockwise ratio of 90th-percentile functional error on the Llama-3.1/ArmoRM pools, as a function of the tilt temperature β, one curve per winsorization quantile q; error bars are 95% prompt-cluster bootstrap intervals over 18,900 cells per point. Blockwise normalization improves as the clip tightens (ratio above one for every $q \leq 0.95$), and only the near-inactive $q = 0.99$ clip, which reduces to one-block, dips below one.

Continuous preference families. The guarantees are uniform over the whole preference set Θ, while the studies above evaluate seven scalarizations. To check that the uniform statement does not conceal a dependence on how finely Θ is sampled, we form the serving law once and evaluate the worst-over-θ functional error on two grids of the coefficient $\theta \in [0, 1]$: the coarse seven-point grid used throughout and a dense forty-nine-point grid. The two sup-over-θ curves coincide to the reported precision at every budget ($0.242, 0.185, 0.137$ at $M \in \{64, 128, 256\}$ for both grids), and the dense-grid curve decays with fitted log–log slope -0.413 (Figure 9). Refining the grid sevenfold leaves the worst-case error unchanged, so the maximum over the sampled preferences already captures the supremum over the continuum: the uniform guarantee is not an artifact of a coarse evaluation grid.

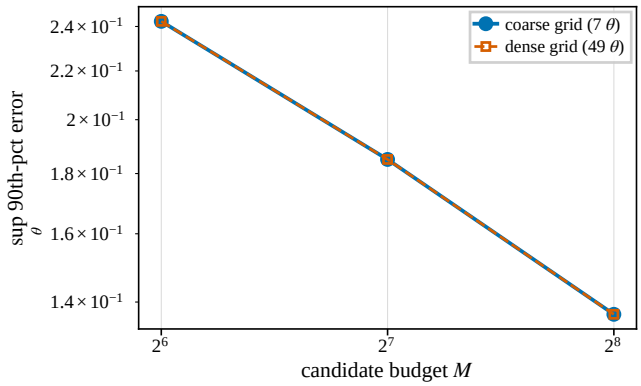


Figure 9: **Uniformity over the preference continuum.** Worst-over-θ 90th-percentile functional error of the blockwise served law on the Llama-3.1/ArmoRM pools, evaluated on a coarse seven-point and a dense forty-nine-point grid of the preference coefficient $\theta \in [0, 1]$, versus the candidate budget M. The two curves coincide, so densifying the grid does not raise the realized supremum.

Downstream serving win-rate. The deviations above are relative to the reference proxy; we complement them with a downstream comparison that scores served responses directly. Treating the reward scalarization s_θ as a judge, we draw a response from the served law and one from the preference-agnostic base π_0 and record which scores higher, averaging over the seven preferences at $M = 256$ and $\beta = 1$ (Figure 10). The blockwise served law wins against the base 0.707 of the

time (95% interval $[0.701, 0.713]$), matching both the one-block law (0.711, $[0.705, 0.717]$) and the reference target ceiling obtained from the M_{ref} pool (0.713, $[0.706, 0.720]$): serving the Gibbs target through the candidate pool recovers the improvement the target itself would deliver. Best-of-n wins against the base 0.998 of the time, but this measures the judge against itself (Gao et al., 2023) — it selects the response that maximizes s_θ and is scored by s_θ — and the same judge ranks the blockwise sample above the best-of-n response only 0.015 of the time. Best-of-n optimizes the scalar it is graded on rather than serving the tilted law, and the win-rate makes the target mismatch explicit rather than favorable.

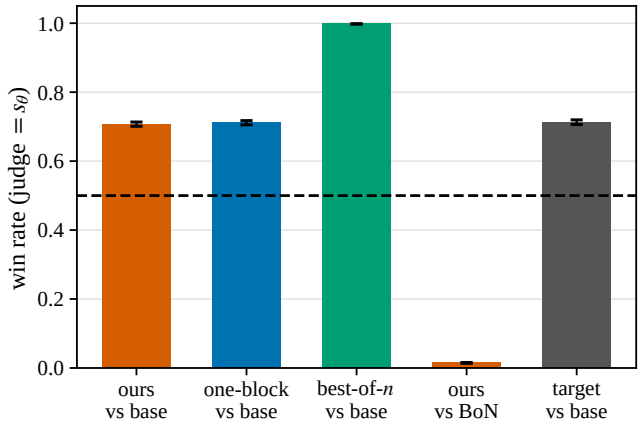


Figure 10: **Downstream serving win-rate under the reward scalarization.** Fraction of prompts and preferences ($M = 256$, $\beta = 1$) on which a response from each policy outscores a base draw under the judge s_θ; the final bar is the blockwise sample against the best-of-n response. Error bars are 95% prompt-cluster bootstrap intervals. The blockwise served law matches the one-block law and the reference target ceiling; the near-perfect best-of-n bar is the judge scoring its own argmax.

G.4 Explicit Constants and the Certified Regime

The rate $r_{\text{blk}} = \widetilde{\mathcal{O}}(\sqrt{\Gamma \log(1/\delta)/M})$ carries the explicit constants of (54), and evaluating them fixes the candidate budget at which the guarantee becomes numerically informative on the pools of Section G.3. There Γ_2 has median 3.13, and the log-factors realized by the shared serving family are $\mathfrak{L}_{\text{P}} = 12.32$ and $\mathfrak{L}_{\text{V}} = 11.63$; we hold Γ and both log-factors at these values and minimize r_{blk} over the block count K at each M, with $b = M/K$.

Table 5: Functional radius r_{blk} at the median real-pool constants ($\Gamma = 3.13$, $\mathfrak{L}_{\text{P}} = 12.32$), minimized over K, versus the per-prompt candidate budget M. The bounded-functional target is $|f| \le 1$, so a radius below the trivial value 2 is the point at which the certificate improves on reporting the range of f.

M	256	512	1024	2048	4096	8192
$\min_K r_{\text{blk}}$	6.69	4.73	3.35	2.37	1.67	1.18
minimizing K	9	13	18	26	37	52

The radius crosses the trivial bound 2 at $M = 2865$ candidates per prompt and reaches 1.67 at $M = 4096$; the value branch needs $K_Z \ge 8\mathfrak{L}_{\text{V}} = 95$ odd blocks and, for $\kappa_Z \le \frac{1}{2}$, block size $b_Z \ge 16\Gamma = 51$, hence $M \ge K_Z b_Z = 4845$. The $M \in \{16, \dots, 128\}$ pools of Section G.3 therefore lie below this crossover: at those budgets the explicit constants leave the radius above 2, and the measured $M^{-1/2}$ decay (slope -0.500) is the operative statement, the certificate itself entering once M reaches a few thousand candidates — which we reach and verify on these same pools at $M = 4096$ below. The synthetic study of Section G.1 runs at those budgets and enters the certified regime directly, which is why its curves match the theory slope with radii below the trivial

bound. The constants are those of the analysis, not tuned; sharper block moments would move the crossover left without changing the $\sqrt{\Gamma/M}$ law.

Empirical validation of the certificate. On the theorem's high-probability event, the radius (54) upper-bounds the realized error. Thus two claims are separable: that evaluating its constants reproduces Table 5, and that the bound it reports at each M is not vacuous — the measured error stays below it. We verify both directly. Fixing $\Gamma = 3.13$ and $\mathfrak{L}_{\mathrm{P}} = 12.32$ and minimizing over K recovers Table 5 to the reported precision at every column ($r_{\mathrm{blk}} = 6.69, 4.73, 3.35, 2.37, 1.67, 1.18$ with $K = 9, 13, 18, 26, 37, 52$) and locates the crossover of the trivial bound at $M = 2865$. We then draw a Gaussian tilt family ($q = \pi_0 = \mathcal{N}(0,1)$, $s_\theta(y) = \theta y$, $\beta = 1$) whose worst-preference envelope matches the pools, $\Gamma_{\max} \approx 3.129$, and measure the joint 90th-percentile blockwise functional error across $\theta \in \{0.6, 0.75, 0.9, 1.0, 1.068\}$ over 300 replicates at each budget $M \in \{2^{10}, \ldots, 2^{17}\}$, with K set to the certificate minimizer at every M. The measured error rides the $M^{-1/2}$ law (fitted log–log slope -0.505) and the certified radius upper-bounds it at every budget: the radius falls below the trivial bound 2 at $M = 4096$ ($r_{\mathrm{blk}} = 1.67$) and reaches 0.296 at $M = 131{,}072$, while the realized error it dominates decays from 0.047 to 0.0037 (Figure 11). The certificate is thus an honest envelope over the whole range, tightest exactly where the budget is large enough to make it informative.

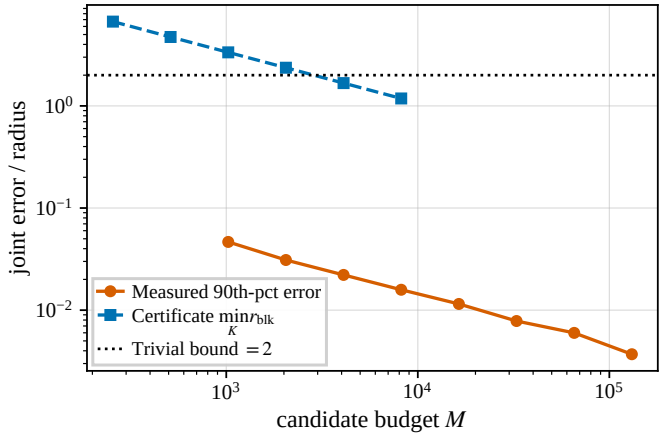


Figure 11: **The high-probability certified radius as a valid, non-vacuous envelope.** Joint 90th-percentile blockwise functional error (measured over 300 replicates of a Gaussian tilt family with $\Gamma_{\max} \approx 3.13$, matching the real-pool constants) against the certified radius $\min_K r_{\mathrm{blk}}$ of (54), versus the candidate budget M. The measured error tracks the $M^{-1/2}$ rate (fitted slope -0.505) and lies below the certified radius at every M; the radius crosses the trivial $|f| \le 1$ bound of 2 at $M = 4096$ and continues to 0.296 at $M = 131{,}072$. Evaluating the constants reproduces Table 5 exactly.

The certificate on the real LM pools. The study above validates the radius against a quadrature ground truth; we now measure it directly on the Llama-3.1/ArmoRM pools, closing the gap between the certified budget and the $M \le 128$ pools of Section G.3. We regenerate 24 HH prompts at a serving budget $M = 4096$ (past the crossover $M = 2865$) with an independent reference bank of 8192 candidates per prompt—drawn under a disjoint generation seed—as the ground-truth target, and measure the joint 90th-percentile blockwise functional error across the six tilt strata $\beta \in \{1.0, \ldots, 0.3\}$, three scalarizations θ, and two functionals (864 reward cells). Each cell carries its own measured Γ_2; we hold $\mathfrak{L}_{\mathrm{P}} = 12.32$ at the shipped value and minimize r_{blk} over K exactly as in Table 5.

The certificate is a valid envelope on every cell: the measured error lies below the radius at all 864 cells ($\mathrm{frac}_{\mathrm{valid}} = 1.0$; median measured error 0.030, median radius 2.88), and it is non-vacuous at the envelope the pools actually realize. At the light-tilt stratum ($\beta = 1.0$, median second moment 2.11, comparable to the $\Gamma_2 \approx 3.13$ of the full run in Section G.3) the radius has median $1.37 < 2$ and 93% of cells fall inside the certified regime; as the tilt deepens the measured concentrability climbs (Γ_2 median 6.9 at $\beta = 0.55$, 130 at $\beta = 0.3$) and the median radius crosses the trivial bound (2.48, then 10.8), so the heavier strata enter the certified regime only at larger M—exactly

the $\sqrt{\Gamma/M}$ crossover the theory predicts. Overall 34% of the cells are certified at $M = 4096$, and on that certified subset the radius bounds the measured error at every cell (largest certified error 0.059 against a radius of at most 1.99). At LLM scale the certificate is thus not vacuous: at the realistic envelope it is informative and empirically validated, and it degrades into the heavy-tail strata at precisely the rate the bound names, rather than failing.

The realized envelope and the one-block certificate. Section 5.1 contrasts a guarantee that needs only Γ_2 with one that needs the uniform envelope Γ_∞. On the same $M = 4096$ cells we can now put numbers on that contrast. Theorem 22 certifies the one-block policy at radius $4r_M(\mathfrak{L}_\mathcal{F}; \Gamma, \Lambda) = 4[\sqrt{2\Gamma\mathfrak{L}_\mathcal{F}/M} + 2\Lambda\mathfrak{L}_\mathcal{F}/(3M)]$ for any valid $\Lambda \geq \Gamma_\infty$, with $\mathfrak{L}_\mathcal{F} = \mathfrak{L}_\mathrm{P} - \log 2 = 11.63$ at the shipped constants. Equating it to $\min_K r_{\mathrm{blk}}$ gives the envelope $\Lambda^\star$ below which the one-block certificate would be the tighter of the two, and the comparison in Table 6 is unambiguous.

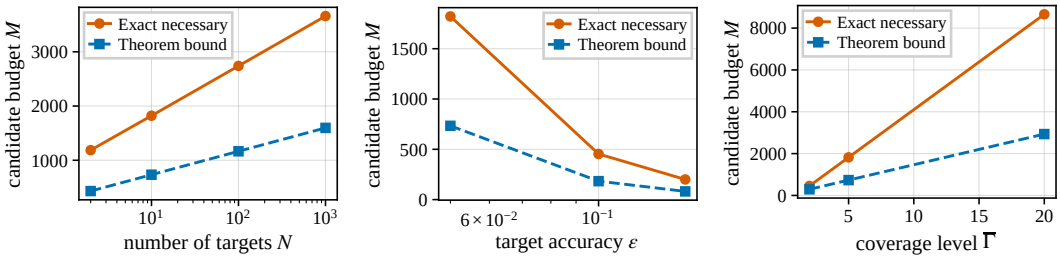
Table 6: Realized weight ratios against the envelope $\Lambda^\star$ at which the one-block certified radius of Theorem 22 would equal the blockwise radius, at $M = 4096$ and the constants of Table 5. $\widehat{R}_\infty := \max_i w_i/\bar{w}$ is the finite-pool weight ratio over the $8{,}192$-candidate bank, which estimates a *lower* bound on Γ_∞; $e^{\widehat{B}/\beta}$ uses the median realized score oscillation $\widehat{B} = 7.98$ and is the only *valid* upper bound on Γ_∞ available without further assumptions ($a = 1$ here, since $q = \pi_0$).

	$\beta = 1.0$	$\beta = 0.55$	$\beta = 0.3$
measured Γ_2 (median)	2.11	6.90	130
blockwise radius $\min_K r_{\mathrm{blk}}$	1.37	2.48	10.79
crossover envelope $\Lambda^\star$	123	223	971
$\widehat{R}_\infty$, median cell	14.8	81.8	540
$\widehat{R}_\infty$, worst cell	120	1,944	7,013
valid bound $e^{\widehat{B}/\beta}$	2.9×10^3	2.0×10^6	3.5×10^{11}
as a multiple of $\Lambda^\star$	$24\times$	8.9×10^3	3.6×10^8

Two readings, and they point the same way. Empirically, the worst cell a uniform guarantee must cover sits at the crossover at $\beta = 1.0$ (120 against 123) and passes it by $8.7\times$ and $7.2\times$ at $\beta = 0.55$ and $\beta = 0.3$; only the *median* cell sits below it, and the median is not the statistic a maximum over prompts, preferences and functionals is taken at. Structurally — and this is the operative point — $\widehat{R}_\infty$ is a finite-pool maximum, hence an estimate of a lower bound on the envelope Γ_∞, so no pool of any size can supply the *upper* bound $\Lambda \geq \Gamma_\infty$ that Theorem 22 requires. The only upper bound available without further assumptions is the deterministic $e^{B/\beta}$ of Proposition 30, which at the *median* realized oscillation — the reading most favorable to one-block, since (40) asks for a single B dominating every cell — already exceeds $\Lambda^\star$ by $24\times$ at $\beta = 1.0$ and by eight orders of magnitude at $\beta = 0.3$; at the largest realized oscillation $\widehat{B} = 13.73$ the three multiples are 7.5×10^3, 3.1×10^8 and 7.8×10^{16}. A second gap compounds this: Theorem 22 is stated under $r_\mathcal{F} \leq \frac{1}{2}$, which at $M = 4096$ admits $\Lambda \leq 206$ at $\beta = 1.0$, $\Lambda \leq 160$ at $\beta = 0.55$, and no Λ at all at $\beta = 0.3$, where the $\sqrt{2\Gamma\mathfrak{L}_\mathcal{F}/M}$ term alone is 0.86. At the two heavier strata the one-block certificate is therefore inapplicable at this budget. The one-block certificate is therefore not instantiable at a useful value on these pools even where the one-block *estimator* is the more accurate of the two (Table 4), whereas the blockwise radius needs only Γ_2, which Theorem 4 certifies before the pool is drawn. We report $\widehat{R}_\infty$ as evidence of scale, never as a certificate.

G.5 Numerical Illustration of the Lower-Bound Instance

Figure 12 evaluates the exact threshold of the necessary coordinate-coverage event in Theorem 16. The theorem's expression remains below this threshold throughout the grid (bound-to-threshold ratio 0.31–0.73) and displays the same near-linear dependence on $\overline{\Gamma} - 1$, inverse-square dependence on ϵ, and logarithmic dependence on N. It visualizes the analytic hard instance used by the proof.



(a) Varying N ($\epsilon = .1, \overline{\Gamma} = 5$). (b) Varying ϵ ($N = 100, \overline{\Gamma} = 5$). (c) Varying $\overline{\Gamma}$ ($N = 100, \epsilon = .1$).

Figure 12: Exact necessary threshold and theorem lower bound on the hard instance.

The controlled experiments use seed 20260828. The language-model run uses prompt seed 7 and generation seed 20260828, and aggregates four independent generation shards. The realized weight ratios and score oscillations of Table 6 are computed from the cached $M_{\mathrm{ref}} = 8192$ reference bank.