Distribution-Free Certificates for Top-K Exploitation in Training-Free NAS

Liz Lemma Future Detective

January 20, 2026

Abstract

Training-free NAS methods rank architectures using zero-cost proxies and typically select the top-scoring model, leaving an 'estimation gap' between proxy rankings and true performance. RoBoT (ICLR 2024) addresses this gap by learning a robust proxy ensemble and then greedily searching within the proxy top set, but its key expected-rank analysis relies on a uniform-rank assumption within the relevant set. We remove this assumption. For a fixed candidate pool and a fixed proxy ranking, we derive an exact, distribution-free expression for the expected true rank achieved by randomized top-K exploitation (sampling B candidates uniformly from the proxy top-K). This expression depends only on the empirical conditional rank CDF inside the proxy top-K. We then design a practical certificate that estimates this CDF under limited objective queries, controls adaptive sampling error, and outputs high-probability upper bounds and stopping rules. Our bounds are tight given the identifiable statistics, and we prove near-matching sample-complexity lower bounds. Experiments on NAS-Bench-201 and TransNAS-Bench-101 show that (i) rank correlation can be misleading while our certificate tracks top-K utility, and (ii) the certificate enables reliable decide-to-exploit vs decide-to-stop behavior across tasks.

Table of Contents

1. 1. Introduction: estimation gap in training-free NAS; why RoBoT's uniform-rank assumption is problematic; goals—distribution-free certification and stopping rules for top-K exploitation.

2. 2. Setup and Definitions: finite pool, proxy scores, objective queries, true ranks; define proxy top-K set and empirical conditional rank CDF; define the randomized exploitation primitive and what is being certified.

3. 3. Exact Distribution-Free Exploitation Identity: derive the hypergeometric survival-function formula for expected best true rank after

sampling B from proxy top-K; interpret as an integral/sum over a survival function; discuss special cases (B=K, small B).

4. 4. What Can and Cannot Be Certified: identifiability limits; why knowing only Precision@K is insufficient; construct counterexamples where Spearman improves but conditional rank CDF (and exploitation utility) worsens.

5. 5. Certificate Construction: estimating the empirical conditional rank CDF under limited queries—sampling from proxy top-K; calibrating approximate ranks via a uniform sample from the whole pool; confidence intervals and union bounds.

6. 6. Certified Stopping and Adaptive Budgeting: online update of confidence bounds as more queries arrive; stopping criterion based on certified expected-rank upper bound; optional selection of K and B to meet a target guarantee.

7. 7. Sample Complexity and Lower Bounds: upper bounds for achieving ε-accurate certificates; minimax lower bounds via reduction to Bernoulli mean estimation / hypothesis testing; discussion of tightness.

8. 8. Empirical Evaluation Plan: NAS-Bench-201 and TransNAS-Bench-101; synthetic stress tests; metrics—certificate calibration, coverage, and decision quality; ablations on sampling strategy and calibration budget.

9. 9. Related Work and Positioning: RoBoT; zero-cost proxies; partial monitoring and top-k feedback; selective inference / adaptive concentration; certified ranking and retrieval.

10. 10. Discussion: extensions to deterministic greedy exploitation, multi-objective constraints, implicit search spaces; limitations and next steps.

1 1. Introduction: estimation gap in training-free NAS; why RoBoT's uniform-rank assumption is problematic; goals—distribution-free certification and stopping rules for top-K exploitation.

In training-free neural architecture search (NAS) and related model-selection problems, we often face an *estimation gap*: we possess a cheaply computed proxy score $s(a)$ for every candidate architecture a (e.g., a zero-cost score, a weight-sharing surrogate, or an ensemble predictor), yet the actual objective $f(a)$ (e.g., validation accuracy after full training) is expensive and can be queried only a small number of times. A common operational pattern is therefore *top-K exploitation*: we sort candidates by s, restrict attention to the proxy top set, query f for a small subset, and return the best observed objective value. This pattern is attractive because it is simple, parallelizable, and empirically effective when the proxy is informative.

The difficulty is that, on any fixed finite pool, the quality of top-K exploitation is governed not by a global correlation statistic between s and f, but by the fine-grained behavior of the true elite ranks inside the proxy-filtered set. In particular, for a given K and a given number B of objective queries spent within the proxy top set, the outcome depends on *how many truly top-ranked candidates are present among those* K and how these counts evolve with the true rank threshold. Two proxy scores may look similar under Spearman or Kendall metrics while inducing markedly different membership of the true top-r items inside the proxy top-K, and it is this membership that controls the success probability of sampling an elite candidate when we query only B items. Consequently, global rank correlation is not, by itself, a certificate of exploitation performance.

Existing certificate-style methods in this setting have frequently relied, implicitly or explicitly, on a *uniform-rank* assumption of the following flavor: conditional on being in the proxy top-K, candidates are treated as if their true ranks behave like (approximately) uniform draws from some range, or as if the proxy-induced ordering noise is exchangeable in a way that makes unobserved ranks "typical." Such assumptions are problematic for two reasons. First, they are rarely verifiable from the limited objective queries available: if we only observe a handful of queried f-values in the proxy top set, many distinct rank configurations remain consistent with the data. Second, these assumptions can be badly violated in precisely the regimes in which certification is most valuable: the proxy top-K may contain a mixture of a few genuinely elite architectures and many mediocre ones (or, conversely, may systematically exclude the global elites due to a proxy failure mode). In either case, a uniform-rank heuristic can yield overly optimistic estimates and premature stopping decisions, even though the underlying instance is fixed and adversarially unfavorable.

Our goal is to replace such assumptions with *distribution-free* guarantees on the performance of randomized top-K exploitation on a fixed finite pool. Concretely, we seek procedures that, using only a limited number of objective queries, return (i) a selected candidate $\hat{a}$ produced by an explicit exploitation primitive, and simultaneously (ii) a *high-probability certificate* upper bounding the expected true rank $\mathbb{E}[R_f(\hat{a})]$, where the expectation is taken solely over the algorithm's internal randomization (the instance $(\mathcal{A}, s, f)$ is treated as deterministic). This expectation is the relevant quantity for randomized exploitation because the only uncertainty is which subset of the proxy top set we happen to query. A certificate in this sense enables principled *stopping rules*: we may terminate exploitation once the certified bound crosses a user-specified target rank, or allocate additional objective queries only when the current certificate remains inconclusive.

The technical obstruction is that a certificate for $\mathbb{E}[R_f(\hat{a})]$ cannot be obtained from proxy information alone; it depends on the unknown true ranks within the proxy top set. Our approach is therefore to (a) identify an exact, assumption-free expression for the expected rank of the exploitation output in terms of simple combinatorial quantities, and (b) estimate conservative lower bounds on these quantities using a small number of carefully structured objective queries. The key point is that the certificate must hold *uniformly over all fixed instances* consistent with the observed objective values, rather than in expectation over an assumed data-generating distribution.

At a high level, we proceed in three steps. First, we analyze the randomized exploitation primitive that samples B distinct candidates uniformly from the proxy top-K, queries f on them, and returns the best observed candidate. For a fixed instance, the expected output rank admits an exact hypergeometric survival-function identity expressed through the cumulative counts of truly top-ranked items inside the proxy top set. Second, we design a calibration mechanism that uses a small number of objective queries both globally (to translate observed objective values into rank information) and conditionally within the proxy top set (to estimate, conservatively, how much true top-r mass is present there). Third, we combine these ingredients into a computable, data-dependent upper bound on $\mathbb{E}[R_f(\hat{a})]$ that holds with probability at least $1 - \delta$ over the algorithm's sampling.

Two features are essential. The first is *monotonicity*: the exploitation expected rank improves as the number of true elites within the proxy top set increases, so lower bounds on elite counts yield upper bounds on expected rank. The second is *finite-population validity*: because we operate on a fixed pool and sample without replacement, our bounds must be phrased in terms of finite-sample concentration and combinatorial sampling, rather than asymptotics or parametric noise models. The resulting certificate procedure yields nonvacuous guarantees whenever the proxy top set contains a detectable fraction of high-rank candidates, and its query complexity matches the $\Omega(1/\varepsilon^2)$ barriers inherent to distribution-free certification up to logarith-

mic factors.

We now formalize the setting and introduce the definitions needed to state the exploitation identity and the certificate construction.

2 Setup and Definitions

We work in a finite-pool query model. Let $\mathcal{A} = \{a_1, \dots, a_N\}$ be a fixed set of N candidates (architectures). For each $a \in \mathcal{A}$ we are given a *proxy* score $s(a) \in \mathbb{R}$, which is assumed to be available for all candidates at negligible cost relative to objective evaluation. In contrast, the *objective* value $f(a) \in \mathbb{R}$ (e.g., accuracy after full training) is unknown a priori and may be revealed only by issuing an objective query for a, at unit cost. Throughout, we adopt the convention that larger objective values are better.

The objective values induce a global ranking on $\mathcal{A}$. We write $R_f(a) \in \{1, \dots, N\}$ for the (descending) true rank of a under f: $R_f(a) = 1$ means that a is globally best, and $R_f(a) \le r$ means that a lies in the true top-r set. When objective values tie, we fix an arbitrary deterministic tie-breaking rule once and for all; this makes $R_f(\cdot)$ a well-defined function on $\mathcal{A}$ for a fixed instance $(\mathcal{A}, s, f)$.

Given a budget-sensitive setting, we focus on the standard *proxy top-K* restriction. Fix $K \in \{1, \dots, N\}$. Let $S_K(s) \subseteq \mathcal{A}$ denote the set of the K candidates with largest proxy score s, with deterministic tie-breaking. Intuitively, $S_K(s)$ is the exploitation pool suggested by the proxy. The central quantities governing exploitation performance are the counts of truly elite candidates *inside* this proxy-filtered set. For each threshold $r \in \{0, 1, \dots, N\}$, define the empirical conditional elite count

$$m_r \;:=\; \big|\{a \in S_K(s) : R_f(a) \le r\}\big|, \qquad m_0 := 0,$$

and the associated empirical conditional rank CDF

$$F_K(r) \;:=\; \frac{m_r}{K}.$$

Thus m_r records how many members of the proxy top-K are also in the true top-r globally, and $F_K(r)$ is the corresponding fraction. The sequence $r \mapsto m_r$ is nondecreasing, satisfies $0 = m_0 \le m_r \le K$, and stabilizes at $m_N = K$. Our results will treat $(m_r)_{r=0}^{N}$ as an instance-dependent summary of how well the proxy top set captures the true elite ranks.

We now formalize the randomized exploitation primitive whose performance we seek to certify. Fix an integer $B \in \{1, \dots, K\}$, interpreted as the number of objective queries allocated *within* $S_K(s)$ for exploitation. The exploitation algorithm draws a subset $E \subseteq S_K(s)$ uniformly at random among all size-B subsets (sampling *without replacement*), queries $f(a)$ for all $a \in E$, and returns

$$\hat{a}_B \;\in\; \arg\max_{a \in E} f(a),$$

with deterministic tie-breaking on the argmax if needed. All randomness is therefore internal to the algorithm and arises only from the random choice of E. The induced output rank $R_f(\hat{a}_B)$ is a random variable on $\{1, \dots, N\}$ even though the instance $(\mathcal{A}, s, f)$ is fixed.

The quantity we will certify is the *expected* true rank of the exploitation output,

$$\mathbb{E}\big[R_f(\hat{a}_B)\big],$$

where the expectation is taken solely over the algorithm's sampling of E (equivalently, over all $\binom{K}{B}$ equally likely subsets of $S_K(s)$). This is the relevant notion of performance for randomized exploitation: once the instance is fixed, the only uncertainty is which B proxy-top candidates happen to be queried, and the expectation captures the average quality of the returned candidate under that randomization. Importantly, we make *no* distributional assumptions on how f relates to s; the ranks can be adversarially arranged subject only to being a fixed total order on $\mathcal{A}$.

A *certificate* is a computable, data-dependent upper bound U_B satisfying a high-probability validity requirement of the form

$$\Pr\big(\mathbb{E}\big[R_f(\hat{a}_B)\big] \le U_B\big) \;\ge\; 1 - \delta,$$

for a user-chosen failure probability $\delta \in (0, 1)$. Here the probability is over all random choices made by the certificate procedure (including any calibration sampling and, if desired, the exploitation sampling), while the instance $(\mathcal{A}, s, f)$ is treated as fixed. In particular, the certificate must remain valid uniformly over all possible objective assignments f on unqueried candidates that are consistent with the observed queried objective values and the deterministic tie-breaking rules. In the sequel, we will construct such certificates using a limited number of objective queries split between (i) global calibration queries (to translate observed f-values into rank information) and (ii) conditional queries within $S_K(s)$ (to conservatively lower bound m_r and hence upper bound the exploitation expectation). The next section establishes that, for fixed K and B, the expected rank $\mathbb{E}[R_f(\hat{a}_B)]$ admits an exact, distribution-free expression in terms of the finite-population quantities m_r, thereby reducing certification to estimating these counts conservatively.

3 Exact Distribution-Free Exploitation Identity

We first isolate an exact finite-population identity for the randomized exploitation primitive. The point of the identity is that, once $\mathcal{A}$, s, and f are fixed, the distribution of the output rank $R_f(\hat{a}_B)$ under uniform sampling within $S_K(s)$ is determined entirely by the sequence of conditional elite counts $(m_r)_{r=0}^{N}$. In particular, no stochastic relationship between s and f is required: the identity holds instance-by-instance.

Recall that $\hat{a}_B$ is obtained by drawing a size-B subset $E \subset S_K(s)$ uniformly without replacement and returning the queried element with maximum objective value f (equivalently, minimum true rank). Thus,

$$R_f(\hat{a}_B) \;=\; \min_{a \in E} R_f(a).$$

For any positive-integer-valued random variable X, we have the elementary survival-function representation

$$\mathbb{E}[X] \;=\; \sum_{r \geq 0} \Pr(X > r),$$

which is the discrete analogue of $\mathbb{E}[X] = \int_0^\infty \Pr(X > t)\, dt$. Applying this to $X = R_f(\hat{a}_B) \in \{1, \ldots, N\}$ yields

$$\mathbb{E}[R_f(\hat{a}_B)] \;=\; \sum_{r=0}^{N-1} \Pr(R_f(\hat{a}_B) > r)\,.$$

It remains to compute the event probabilities in terms of m_r. Fix $r \in \{0, 1, \ldots, N-1\}$. The event $\{R_f(\hat{a}_B) > r\}$ is equivalent to the sampled subset E containing no globally top-r elements. Among the K elements of $S_K(s)$, exactly m_r have true rank at most r, and hence $K - m_r$ have true rank strictly larger than r. Because E is a uniform size-B subset of $S_K(s)$, we obtain the hypergeometric survival probability

$$\Pr(R_f(\hat{a}_B) > r) \;=\; \frac{\binom{K-m_r}{B}}{\binom{K}{B}}.$$

Substituting into the survival-function representation yields the promised expression.

Theorem 3.1 (Exact exploitation identity)**.** *Fix any instance $(\mathcal{A}, s, f)$, and let $S_K(s)$ be the proxy top-K set. Let $E \subset S_K(s)$ be a uniformly random subset of size $B \leq K$ sampled without replacement, and let $\hat{a}_B \in \arg\max_{a \in E} f(a)$ (with deterministic tie-breaking). Then, with $m_r = |\{a \in S_K(s) : R_f(a) \leq r\}|$ and $m_0 = 0$,*

$$\mathbb{E}[R_f(\hat{a}_B)] \;=\; \sum_{r=0}^{N-1} \frac{\binom{K-m_r}{B}}{\binom{K}{B}}.$$

Theorem 3.1 makes explicit which aspects of the proxy matter for randomized exploitation. The expected rank is a functional of the entire conditional rank CDF $F_K(r) = m_r/K$, not merely a single summary such as m_K. Moreover, the dependence is monotone in the natural direction: for each fixed r, the term $\binom{K-m_r}{B}/\binom{K}{B}$ is nonincreasing in m_r, reflecting the fact that having

more true top-r items inside $S_K(s)$ can only improve the chance that a size-B sample hits at least one such item.

The identity also admits a useful interpretation as an *integral over missed-elite probabilities*. For each rank threshold r, the hypergeometric fraction is exactly the probability that B draws from $S_K(s)$ miss the subset $\{a \in S_K(s) : R_f(a) \le r\}$. Summing these miss-probabilities over r aggregates the entire tail behavior of the output rank: large contributions come precisely from thresholds r for which the proxy top set contains few truly elite candidates.

Several special cases sharpen intuition.

Case $B = K$ (query all proxy top-K). Here $E = S_K(s)$ deterministically, so $\hat{a}_B$ is simply the best element in $S_K(s)$. Let $r^\star := \min\{r : m_r \ge 1\}$, i.e., the true rank of the best element in $S_K(s)$. Since $\binom{K-m_r}{K} = 1$ iff $m_r = 0$ and 0 otherwise, the identity reduces to

$$\mathbb{E}[R_f(\hat{a}_K)] \;=\; \sum_{r=0}^{N-1} \mathbb{1}\{m_r = 0\} \;=\; r^\star,$$

as expected.

Case $B = 1$ (one random query in S_K). Then $\hat{a}_1$ is a uniformly random element of $S_K(s)$, and the identity recovers the mean rank within the proxy top set. Indeed, $\binom{K-m_r}{1}/\binom{K}{1} = (K - m_r)/K$, so

$$\mathbb{E}[R_f(\hat{a}_1)] \;=\; \sum_{r=0}^{N-1} \left(1 - \frac{m_r}{K}\right),$$

which is equivalent to $\frac{1}{K}\sum_{a \in S_K(s)} R_f(a)$ by the standard counting identity $\sum_{r \ge 0} \mathbb{1}\{R_f(a) > r\} = R_f(a)$.

Small B and the role of early ranks. When $B \ll K$, each term $\binom{K-m_r}{B}/\binom{K}{B}$ is close to $\left(1 - \frac{m_r}{K}\right)^B$ (the corresponding with-replacement probability), making transparent that the dominant contributions to $\mathbb{E}[R_f(\hat{a}_B)]$ arise from small r where m_r/K is small. Consequently, certification efforts should focus on lower bounding m_r for relatively small r, since improving knowledge of m_r near $r = N$ has negligible effect on the sum.

In the next section we use Theorem 3.1 as the deterministic backbone for certification: if we can produce high-probability lower bounds $\underline{m}_r \le m_r$ for a range of thresholds r, then monotonicity of the hypergeometric survival terms yields an immediate, computable upper bound on $\mathbb{E}[R_f(\hat{a}_B)]$ without any assumption linking s and f.

4 What Can and Cannot Be Certified

Theorem 3.1 reduces the analysis of randomized exploitation within $S_K(s)$ to the finite sequence $(m_r)_{r=0}^N$. Consequently, any *instance-wise* (distribution-free) performance statement about $\mathbb{E}[R_f(\hat{a}_B)]$ must, implicitly or explicitly, control m_r for a range of thresholds r. This creates an identifiability issue: if the available information does not determine (or at least constrain) the conditional rank CDF $F_K(r) = m_r/K$ at the relevant low ranks, then no nonvacuous certificate for $\mathbb{E}[R_f(\hat{a}_B)]$ can be valid uniformly over all fixed instances $(\mathcal{A}, s, f)$ consistent with that information.

Why knowing only Precision@K **is insufficient.** A common summary of proxy quality is Precision@$K = m_K/K$, i.e., the fraction of truly top-K items that appear in $S_K(s)$. However, m_K is only a single coordinate of the full vector (m_r), and Theorem 3.1 shows that $\mathbb{E}[R_f(\hat{a}_B)]$ depends on the entire tail $\{\binom{K-m_r}{B}/\binom{K}{B}\}_{r=0}^{N-1}$. In particular, for small B, the dominating contribution arises from small r, where m_r can vary widely even when m_K is fixed.

Formally, fix $N \gg K$ and $B \le K$. Consider two fixed instances that share the *same* pool $\mathcal{A}$ and the *same* proxy top set $S_K(s)$ (so the proxy scores may even be identical), and satisfy the *same* value of m_K, but differ in the placement of true ranks within $S_K(s)$. Let $t \in \{1, \dots, K\}$ and enforce $m_K = t$. In both instances, we place $K - t$ elements of $S_K(s)$ at very poor global ranks, say

$$R_f(a) \in \{N - (K - t) + 1, \dots, N\} \qquad \text{for } K - t \text{ elements of } S_K(s).$$

The remaining t elements of $S_K(s)$ are placed within the true top-K, ensuring $m_K = t$, but we choose two extreme configurations:

- (*Front-loaded*) the t in-top-K elements have ranks $1, 2, \dots, t$, hence $m_r = t$ for all $r \ge t$.
- (*Back-loaded*) the t in-top-K elements have ranks $K - t + 1, \dots, K$, hence $m_r = 0$ for all $r \le K - t$.

In both cases $m_K = t$ holds, yet the exploitation behavior differs sharply. Indeed, with back-loading, the probability that B uniform draws from $S_K(s)$ miss all t acceptable elements is

$$\Pr(\text{miss all } t) = \frac{\binom{K-t}{B}}{\binom{K}{B}},$$

and on this miss event the returned rank is at least $N-(K-t)+1$. Therefore,

$$\mathbb{E}[R_f(\hat{a}_B)] \;\ge\; \frac{\binom{K-t}{B}}{\binom{K}{B}} \cdot \big(N - (K - t) + 1\big). \tag{1}$$

For t fixed and $B \ll K$, the factor $\binom{K-t}{B}/\binom{K}{B} \approx (1-t/K)^B$ is close to 1, so the lower bound in (1) is $\Theta(N)$. Thus, any purported certificate that depends only on m_K (or Precision@K) must remain valid on both configurations and is forced to be of order N in the worst case, i.e., essentially vacuous in normalized rank. This is an identifiability barrier, not an artifact of a particular analysis.

Rank correlation is not a certificate. A related pitfall is to treat global rank correlation between s and f (e.g., Spearman ρ_{Sp}) as a surrogate for exploitation utility. Theorem 3.1 implies that exploitation depends on whether $S_K(s)$ contains truly elite items (and how many), i.e., on m_r for small r, whereas ρ_{Sp} averages squared displacements over *all* N items and can be dominated by the behavior on the non-elite majority.

We can construct explicit proxies s and s' (for the same fixed f) such that $\rho_{\mathrm{Sp}}(s,f) > \rho_{\mathrm{Sp}}(s',f)$ while $\mathbb{E}[R_f(\hat{a}_B(s))] \gg \mathbb{E}[R_f(\hat{a}_B(s'))]$. Identify f with the identity ranking $R_f(a_i) = i$. Let s induce the permutation that swaps the top block $\{1,\dots,K\}$ with the bottom block $\{N-K+1,\dots,N\}$ and fixes all middle ranks. Then $S_K(s)$ consists of the K worst items under f, so $m_r = 0$ for all $r \le N-K$, and in particular $\mathbb{E}[R_f(\hat{a}_B(s))] \ge N-K+1$ (indeed, every queried item has rank at least $N-K+1$). Yet the Spearman degradation of this block swap is small when $K \ll N$: the squared displacement is $\Theta(K(N-K)^2)$, hence

$$\rho_{\mathrm{Sp}}(s,f) = 1 - \frac{6\sum_{i=1}^{N}(i-\pi(i))^2}{N(N^2-1)} \;\ge\; 1 - O\!\left(\frac{K}{N}\right),$$

which can be arbitrarily close to 1 for fixed K and large N.

Conversely, let s' be a proxy that places the true top-K items in $S_K(s')$ but permutes the remaining $N-K$ items nearly arbitrarily. Then m_r for small r is maximized ($m_r = r$ for $r \le K$), implying much smaller $\mathbb{E}[R_f(\hat{a}_B(s'))]$, while the global Spearman correlation can be made close to 0 by sufficiently scrambling the bottom $N-K$ ranks. Hence, higher ρ_{Sp} does not imply better exploitation, and it cannot serve as a distribution-free certificate for $\mathbb{E}[R_f(\hat{a}_B)]$.

These examples isolate the correct target for certification: we must estimate (or lower bound) the conditional elite counts m_r for a range of r, particularly at small r where the hypergeometric survival terms are sensitive. This motivates the next section, where we show how to obtain high-probability lower bounds $\underline{m}_r \le m_r$ from a limited number of objective queries via conditional sampling within $S_K(s)$ and rank calibration from a uniform sample of $\mathcal{A}$.

5 Certified Stopping and Adaptive Budgeting

Having reduced certification to lower bounds on the conditional elite counts m_r and having described how to obtain $\underline{m}_r \le m_r$ with high probability from limited objective queries, we now describe how to use the resulting certificate online. The central point is that the certificate

$$U_B \;=\; \sum_{r=0}^{N-1} \frac{\binom{K-\underline{m}_r}{B}}{\binom{K}{B}}$$

is (i) computable from the queried data, (ii) monotone nonincreasing in each $\underline{m}_r$, and (iii) monotone nonincreasing in B (for fixed $\underline{m}_r$). These monotonicities allow us to maintain a valid expected-rank upper bound as more queries arrive, and to implement a stopping rule and a budget-allocation policy that are themselves certified.

Online maintenance of confidence bounds. Suppose we run in rounds $t = 1, 2, \ldots$, and at round t we have queried $n_U(t)$ items from $\mathcal{A}$ (global calibration) and $n_S(t)$ items from $S_K(s)$ (conditional sampling). From these data we compute updated rank intervals $[\underline{R}_t(\cdot), \overline{R}_t(\cdot)]$ for the conditional samples and updated one-sided bounds $\underline{m}_{r,t}$ for m_r over $r \in \mathcal{R}$.

To ensure validity under optional stopping, we must guarantee that the event

$$\mathcal{E} \;:=\; \bigcap_{t\ge 1} \bigcap_{r\in\mathcal{R}} \{\underline{m}_{r,t} \le m_r\}$$

holds with probability at least $1-\delta$. There are two standard ways to enforce this in our finite-pool setting. First, we may commit to a maximal horizon $t_{\max}$ (equivalently a maximal number of calibration queries) and apply a union bound over $|\mathcal{R}|t_{\max}$ events, using one-sided confidence intervals at level $\delta/(|\mathcal{R}|t_{\max})$. Second, we may avoid committing to $t_{\max}$ by using a summable error schedule $(\delta_t)_{t\ge1}$ (e.g. $\delta_t = \delta\cdot 2^{-t}$) and enforce at round t that all intervals are valid at level $\delta_t/|\mathcal{R}|$; then $\sum_t \delta_t = \delta$ implies $\Pr(\mathcal{E}) \ge 1-\delta$. Either construction yields an *anytime* certificate process $(U_{B,t})_{t\ge1}$, where

$$U_{B,t} \;:=\; \sum_{r=0}^{N-1} \frac{\binom{K-\underline{m}_{r,t}}{B}}{\binom{K}{B}}, \qquad \text{and} \qquad \Pr(\forall t:\ \mathbb{E}[R_f(\hat{a}_B)] \le U_{B,t}) \ge 1-\delta.$$

In particular, since $\underline{m}_{r,t}$ is nondecreasing in t (additional evidence can only certify more elite items), the sequence $U_{B,t}$ is nonincreasing in t.

A certified stopping criterion. Fix a target expected-rank guarantee $U^\star \in \{1, \ldots, N\}$ (or a normalized target $u^\star \in (0,1)$ with $U^\star = \lceil u^\star N \rceil$). We define the stopping time

$$\tau \;:=\; \inf\{t \ge 1:\ U_{B,t} \le U^\star\},$$

with the convention $\tau = \infty$ if the set is empty. On the event $\mathcal{E}$ we have $\mathbb{E}[R_f(\hat{a}_B)] \leq U_{B,t}$ for all t, hence whenever $\tau < \infty$ we obtain the desired guarantee $\mathbb{E}[R_f(\hat{a}_B)] \leq U^\star$. Importantly, this statement is instance-wise and does not rely on any stochastic model for $(\mathcal{A}, s, f)$: only the algorithm's randomness is controlled.

When τ triggers, we may execute exploitation (sample B items from $S_K(s)$, query f, return the best) using the remaining budget, or we may have already interleaved exploitation queries with calibration. In either case the stopping rule is purely certificate-based: we stop once the certificate becomes strong enough, and not merely when the observed best f-value appears large.

Adaptive choice of B given a remaining budget. Because $U_{B,t}$ is monotone in B, once we have computed $\underline{m}_{r,t}$ we can select B adaptively to meet a target with minimal exploitation cost. Concretely, at round t we may define

$$B_t^\star \;:=\; \min\{B \in \{1, \ldots, K\} :\; U_{B,t} \leq U^\star\},$$

provided the set is nonempty. If we have a remaining objective-query budget $Q_{\mathrm{rem}}(t)$, we may impose $B \leq Q_{\mathrm{rem}}(t)$ and either (i) pick $B = B_t^\star$ to satisfy the target as soon as it is feasible, or (ii) pick the largest feasible B to strengthen the certificate further (since increasing B can only help), then execute exploitation. This turns the certificate into a control knob: we translate desired performance into a concrete number of exploitation queries.

Optional selection of K. If K is not fixed a priori, we may treat it as a tunable parameter and compute certificates over a discrete set $\mathcal{K} \subseteq \{1, \ldots, N\}$. For each $K \in \mathcal{K}$ we form $S_K(s)$, draw conditional samples from that set, and compute $U_{B,t}(K)$. The dependence on K is not monotone in general: larger K increases diversity (potentially increasing m_r) but also dilutes elite fraction m_r/K. Consequently, the principled choice is to select

$$(K_t^\star, B_t^\star) \;\in\; \arg\min_{K \in \mathcal{K},\, B \leq K} \; U_{B,t}(K) \quad \text{subject to budget constraints},$$

or, in the target-driven variant, the least-cost pair achieving $U_{B,t}(K) \leq U^\star$. Since all certificates are valid simultaneously under the same $\mathcal{E}$ event (via a union bound over $K \in \mathcal{K}$ as well), this model selection does not compromise correctness.

In summary, once we have an anytime-valid mechanism producing $\underline{m}_{r,t}$, we obtain an anytime-valid expected-rank upper bound $U_{B,t}$, a certified stopping rule based on $U_{B,t} \leq U^\star$, and an adaptive budgeting scheme that chooses B (and optionally K) to meet a user-specified guarantee within a finite query budget.

Making exploitation compatible with adaptive stopping. A subtlety is that the certificate $U_{B,t}$ is derived for the randomized exploitation rule that draws a *fresh* uniform subset $E \subset S_K(s)$ of size B. If we were to "reuse" previously queried conditional samples as exploitation candidates, then the final evaluated set would no longer be distributed as a uniform size-B subset, and Theorem 1 would not apply as stated. A convenient remedy is to couple calibration, stopping, and exploitation through a single randomization: at the beginning, draw a uniformly random permutation π of $S_K(s)$, and reveal/query its elements sequentially. For each $b \in \{1, \dots, K\}$, the prefix $\{\pi(1), \dots, \pi(b)\}$ is a uniform size-b subset of $S_K(s)$. Consequently, if we define a (possibly data-dependent) stopping time τ and then set $E := \{\pi(1), \dots, \pi(B_\tau)\}$, we still have that E is uniform conditional on B_τ, and the certified bound $\mathbb{E}[R_f(\hat{a}_{B_\tau})] \le U_{B_\tau,\tau}$ remains valid on $\mathcal{E}$. Operationally, we may use the early queried prefix elements both (i) to update $\underline{m}_{r,t}$ and (ii) as the eventual exploitation pool, while preserving the sampling model needed by the hypergeometric identity.

Anytime construction of $\underline{m}_{r,t}$ with a summable error schedule. When we update $\underline{m}_{r,t}$ online, we require the whole trajectory $\{\underline{m}_{r,t}\}_{t\ge1,r\in\mathcal{R}}$ to be simultaneously valid. In practice we implement the summable schedule described above by selecting $\delta_t = \delta \cdot 2^{-t}$ and, at round t, running (i) the global DKW band for rank calibration at confidence level $\delta_t/2$, and (ii) the one-sided binomial lower bounds for $\{\underline{p}_{r,t}\}_{r\in\mathcal{R}}$ at level $\delta_t/(2|\mathcal{R}|)$ each. By the union bound over $r \in \mathcal{R}$ and the summability of (δ_t), we ensure

$$\Pr(\forall t \ge 1,\ \forall r \in \mathcal{R} :\ \underline{m}_{r,t} \le m_r) \ \ge\ 1 - \delta,$$

without committing to a finite horizon. This design makes the certificate "anytime" in the usual optional-stopping sense: the algorithm may examine $U_{B,t}$ at each round and decide whether to stop, increase calibration, or increase exploitation, while maintaining the stated coverage.

Budget allocation via marginal certificate improvement. Given that objective queries are scarce, a natural control question is how to allocate the next query between global calibration (improving rank intervals) and conditional sampling (improving the estimated elite fractions within S_K). While the exact optimal allocation depends on the instance, we can base a simple policy on the observed bottleneck in the current certificate. Since

$$U_{B,t} = \sum_{r=0}^{N-1} \frac{\binom{K-\underline{m}_{r,t}}{B}}{\binom{K}{B}}$$

is monotone in each $\underline{m}_{r,t}$, we may inspect which ranks r contribute most to the sum (typically small r, where the survival probabilities are near 1)

and prioritize improving $\underline{m}_{r,t}$ at those thresholds. Concretely, if the DKW-induced rank intervals are wide, then additional global samples (increasing $n_U(t)$) shrink $[\underline{R}_t(\cdot), \overline{R}_t(\cdot)]$ and can convert previously "uncertified" conditional points into certified top-r points, increasing $I_r(\cdot)$ and thus $\underline{m}_{r,t}$. Conversely, if rank intervals are already tight but $\widehat{p}_{r,t}$ has large binomial uncertainty, then additional conditional samples (increasing $n_S(t)$) are more effective. A minimal implementation is a doubling schedule: increase n_U until the calibration band width is below a preset rank tolerance, then increase n_S until $U_{B,t}$ meets the target.

Choosing B under finite remaining budget. At any time t, we may compute $U_{b,t}$ for all $b \in \{1, \dots, \min\{K, Q_{\mathrm{rem}}(t)\}\}$ and select the smallest feasible b that meets the target. Since $b \mapsto U_{b,t}$ is nonincreasing, this can be done by binary search over b once we can evaluate the sum efficiently. In implementations we also truncate the sum over r: because the terms $\binom{K-\underline{m}_{r,t}}{b}/\binom{K}{b}$ become negligible once $\underline{m}_{r,t}$ is large relative to b, it is often sufficient to sum only up to a moderate $r_{\max}$ (e.g. the smallest r for which the term is below a numerical tolerance), yielding a fast approximate computation that remains conservative if we drop only provably nonnegative tail terms.

Simultaneous consideration of multiple K. When K is not predetermined, we may evaluate a finite set $\mathcal{K}$ of candidate proxy set sizes while maintaining correctness via an additional union bound over $K \in \mathcal{K}$. To reduce query overhead, it is helpful to exploit the nesting $S_K(s) \subseteq S_{K'}(s)$ for $K \le K'$. For example, fix $K_{\max} := \max \mathcal{K}$, sample a random permutation π of $S_{K_{\max}}(s)$, and query sequentially along π. For any $K \in \mathcal{K}$, the induced relative order of elements in $S_K(s)$ is uniform, so the first n queried items that fall in $S_K(s)$ form a uniform sample without replacement from $S_K(s)$. Thus a single stream of conditional queries can support certificates for all $K \in \mathcal{K}$, at the cost of bookkeeping for each K and an additional $\log|\mathcal{K}|$ factor in the confidence accounting. We may then select (K, B) either to minimize the certificate value $U_{B,t}(K)$ or to minimize exploitation cost subject to $U_{B,t}(K) \le U^\star$, while preserving the same instance-wise, distribution-free validity guarantee.

Sample complexity for ε-accurate certificates. We quantify how many objective queries are sufficient for the certificate to approximate the "oracle" value obtained if the entire conditional rank profile $\{m_r\}_{r=1}^N$ were known. Fix a finite grid of thresholds $\mathcal{R} \subseteq \{1, \dots, N\}$ (typically logarithmically spaced), and let $\varepsilon, \delta \in (0, 1)$. We allocate confidence budget across (i) the global rank-calibration step and (ii) the conditional sampling step, via a union bound over $r \in \mathcal{R}$. Assuming (after a monotone transformation if needed) that

$f(a) \in [0,1]$, the Dvoretzky–Kiefer–Wolfowitz inequality implies that with $n_U = \Theta(\varepsilon^{-2}\log(|\mathcal{R}|/\delta))$ uniform queries from $\mathcal{A}$ we obtain a simultaneous (over all $z \in [0,1]$) CDF band

$$\sup_{z\in[0,1]} \left|\widehat{G}(z) - G(z)\right| \;\le\; c\sqrt{\frac{\log(|\mathcal{R}|/\delta)}{n_U}}$$

with probability at least $1-\delta/2$ (for a universal constant c), where G denotes the empirical CDF of $\{f(a)\}_{a\in\mathcal{A}}$. Interpreting a queried value $f(t)$ through this band yields a rank interval $[\underline{R}(t), \overline{R}(t)]$ whose width is $O(N\varepsilon)$ uniformly over all queried t. Consequently, for each $r \in \mathcal{R}$, the conservative indicator $I_r(t) = \mathbb{1}\{\overline{R}(t) \le r\}$ undercounts true membership in the global top-r set, but does so in a controlled manner.

Independently, with $n_S = \Theta(\varepsilon^{-2}\log(|\mathcal{R}|/\delta))$ uniform queries from $S_K(s)$, standard one-sided binomial concentration (e.g. Clopper–Pearson or Chernoff-type bounds) yields simultaneous lower confidence bounds $\underline{p}_r$ on the conditional elite fractions $p_r = m_r/K$:

$$\Pr\Big(\forall r \in \mathcal{R}:\ \underline{p}_r \le p_r \ \text{ and } \ p_r - \underline{p}_r \le c'\sqrt{\tfrac{\log(|\mathcal{R}|/\delta)}{n_S}}\Big) \;\ge\; 1-\delta/2,$$

for a universal $c' > 0$. Thus $\underline{m}_r := \lfloor K\underline{p}_r \rfloor$ satisfies $\underline{m}_r \le m_r$ and $m_r - \underline{m}_r = O(K\varepsilon)$ for all $r \in \mathcal{R}$ on the joint event of probability at least $1-\delta$.

To translate these inaccuracies in $\underline{m}_r$ into inaccuracies in the certificate

$$U_B = \sum_{r=0}^{N-1} \frac{\binom{K-\underline{m}_r}{B}}{\binom{K}{B}},$$

we use that, for fixed K, B, the map $m \mapsto \binom{K-m}{B}/\binom{K}{B}$ is nonincreasing and has bounded discrete slope:

$$\frac{\binom{K-(m+1)}{B}}{\binom{K}{B}} - \frac{\binom{K-m}{B}}{\binom{K}{B}} \;=\; -\frac{B}{K-m}\cdot\frac{\binom{K-m}{B}}{\binom{K}{B}} \;\in\; [-1,0].$$

Hence replacing m_r by $m_r - \Delta_r$ perturbs each summand by at most Δ_r, and summing over r yields a crude but distribution-free bound of order $\sum_r \Delta_r$. With a grid $\mathcal{R}$ and monotone interpolation of $\underline{m}_r$ between grid points, one obtains an additive certificate error of order $O(\varepsilon N)$ in rank units (and $O(\varepsilon)$ in normalized ranks after dividing by N), consistent with Theorem 3. In applications we often care only about small ranks (e.g. a target $U^\star \ll N$); then truncating the sum at a data-dependent $r_{\max}$ improves constants while remaining conservative because all omitted terms are nonnegative.

Minimax lower bounds via Bernoulli mean testing. We next justify that the ε^{-2} dependence is unavoidable for distribution-free certificates. The core obstacle is identifiability of m_r from finitely many objective queries: the certificate must upper bound the expected exploitation rank for *any fixed* instance consistent with observed queries, and thus must distinguish instances whose conditional elite fractions differ by $\Theta(\varepsilon)$. A standard reduction uses two families of instances that share the same proxy scores (hence the same $S_K(s)$) and differ only in which elements of $S_K(s)$ fall into the true global top-r. Concretely, we set $f(a) \in \{0, 1\}$ with deterministic tie-breaking, and choose r so that membership in the global top-r coincides with having label 1. Within $S_K(s)$, we generate labels with mean p versus $p + \Delta$ where $\Delta = \Theta(\varepsilon)$, and arrange the remaining pool $\mathcal{A} \setminus S_K(s)$ so that global ranks are consistent with these labels. Any valid certificate with additive error at most εN in the expected rank (or equivalently, which non-trivially separates the two resulting values of $\sum_r \binom{K-m_r}{B}/\binom{K}{B}$) yields a test distinguishing the two Bernoulli means with probability at least $1 - \delta$. By classical information-theoretic bounds (e.g. Le Cam's method), this requires $\Omega(\Delta^{-2} \log(1/\delta)) = \Omega(\varepsilon^{-2} \log(1/\delta))$ objective queries in the worst case, regardless of adaptivity and regardless of whether the queries are drawn from $S_K(s)$ or from $\mathcal{A}$ for calibration. This establishes Theorem 4.

Tightness and interpretation. Taken together, the upper and lower bounds show that our certificate procedure is minimax-rate optimal up to logarithmic factors in $|\mathcal{R}|$ (and any additional factors arising from anytime schedules). Importantly, this tightness holds without any stochastic assumptions on $(\mathcal{A}, s, f)$: the role of randomness is only the algorithm's sampling. The bounds also clarify what is and is not learnable under finite budgets. Estimating only a single number such as Precision@K (i.e. m_K/K) is insufficient to control $\mathbb{E}[R_f(\hat{a}_B)]$, because Theorem 1 depends on the entire profile $r \mapsto m_r$; conversely, obtaining a coarse approximation of this profile on a modest grid is already enough to yield nonvacuous, quantitatively meaningful certificates for practical values of B. This motivates the empirical evaluation plan that follows, where we measure both coverage (validity) and the decision quality induced by stopping based on U_B.

Empirical evaluation plan. We evaluate the practical behavior of CertiTopK along three axes: (i) validity of the certificate (coverage), (ii) tightness/calibration of the certificate relative to the instance-specific oracle value from Theorem 1, and (iii) decision quality when the certificate is used to drive a stopping rule under a fixed objective-query budget. All experiments are performed in the finite-pool query model of the paper: the pool $\mathcal{A}$ and proxy scores $s(a)$ are fixed and fully observed; randomness arises only from our uniform sampling steps and from any optional anytime schedule.

Benchmarks and instances. We instantiate $(\mathcal{A}, s, f)$ using two standard NAS tabular benchmarks. For NAS-Bench-201, we take $\mathcal{A}$ to be the full set of $N = 15625$ cell architectures and define $f(a)$ as the reported validation accuracy (or a monotone transformation to $[0, 1]$) on each dataset (CIFAR-10, CIFAR-100, ImageNet16-120), yielding three fixed instances per choice of proxy. For TransNAS-Bench-101, we treat each task (and any prescribed search space variant) as a separate instance $(\mathcal{A}, s, f)$, with $f(a)$ given by the benchmark's task-specific score. In both benchmarks, proxy scores $s(a)$ are computed without using objective queries (e.g. zero-cost proxies, training-free scores, or a learned proxy built from auxiliary data); we treat s as fixed input and do not charge its computation to Q. For each instance we sweep (K, B) over a range reflecting realistic exploitation regimes (e.g. $K \in \{50, 100, 200, 500, 1000\}$ with $B \leq K$), and we evaluate multiple total budgets Q with specified splits $Q = n_U + n_S + B$.

Ground-truth quantities enabled by tabular access. Because the benchmarks provide $f(a)$ for all $a \in \mathcal{A}$, we can compute the true ranks $R_f(a)$ exactly and hence compute the full conditional profile $m_r = |\{a \in S_K(s) : R_f(a) \leq r\}|$ for all $r \in \{0, \ldots, N\}$. This enables two forms of ground truth. First, we compute the *oracle expected exploitation rank* for any (K, B) using the exact identity

$$\mathbb{E}[R_f(\hat{a}_B)] \;=\; \sum_{r=0}^{N-1} \frac{\binom{K-m_r}{B}}{\binom{K}{B}}.$$

Second, we can simulate the randomized exploitation procedure itself (sampling B elements from $S_K(s)$ and selecting the best by f) across many random seeds to estimate the realized output-rank distribution; this distributional view is not required for validity (which concerns the expectation), but it diagnoses how conservative certificates translate into realized decision outcomes.

Certificate validity (coverage) and tightness (calibration). For each instance, budget split, and random seed, we run CertiTopK to produce a certificate U_B and record whether it covers the oracle expectation:

$$\mathbf{1}\Big\{\mathbb{E}[R_f(\hat{a}_B)] \leq U_B\Big\}.$$

Empirical coverage is the average of this indicator over independent runs; we report it as a function of δ, (n_U, n_S), and (K, B). To assess tightness, we report additive and multiplicative gaps, e.g.

$$\mathrm{gap}(B) = U_B - \mathbb{E}[R_f(\hat{a}_B)], \qquad \mathrm{ratio}(B) = \frac{U_B}{\mathbb{E}[R_f(\hat{a}_B)]},$$

as well as normalized variants obtained by dividing ranks by N. We additionally evaluate the monotone dependence of U_B on B (and on any anytime schedule) to verify that the certificate yields a sensible tradeoff curve for exploitation size under fixed K.

Decision quality under certificate-driven stopping. To convert certificates into decisions, we fix a target rank level $U^\star$ (or a target normalized rank $U^\star/N$) and run an anytime variant that increases B (and optionally increases n_S) until the certificate satisfies $U_B \leq U^\star$ or the query budget Q is exhausted. We then execute exploitation with the final B and output $\hat{a}_B$. We report decision quality in terms of (i) true rank $R_f(\hat{a}_B)$, (ii) objective value $f(\hat{a}_B)$, and (iii) objective regret relative to the best achievable within $S_K(s)$ or within $\mathcal{A}$, depending on the comparison. As baselines that use the same total query budget, we include: querying B architectures chosen uniformly from $S_K(s)$ without calibration (no certificate), proxy-only selection (returning $\arg\max s$), and fixed-split heuristics (varying $n_U : n_S : B$ without adaptivity). The primary comparison is not in runtime but in how certificate-driven stopping reallocates queries between calibration and exploitation to achieve a desired bound on expected rank.

Ablations and synthetic stress tests. We ablate (i) the calibration split (n_U, n_S) at fixed Q, (ii) the grid choice $\mathcal{R}$ (dense vs. logarithmic vs. truncated to small ranks), and (iii) the conditional sampling policy inside $S_K(s)$. For (iii) we compare uniform without replacement (our default) to with-replacement sampling, and to alternative designs such as stratification by proxy score quantiles; when considering non-uniform designs, we evaluate whether conservative reweighting can preserve validity. We also include synthetic stress tests in which we construct instances by prescribing the placement of true top-r items inside $S_K(s)$ (thus controlling m_r) while varying global proxy–objective rank correlation. These tests isolate the dependence of $\mathbb{E}[R_f(\hat{a}_B)]$ on the conditional profile $r \mapsto m_r$ and allow us to probe regimes where correlation-based summaries are misleading, while still measuring certificate coverage and tightness against the exact oracle value computed from the constructed m_r.

Related work and positioning. Our setting combines two ingredients that are often treated separately: (i) a *finite* candidate pool with a fully observed proxy score for every candidate, and (ii) a *small* budget of expensive objective evaluations used both for exploitation and for producing a high-probability certificate on the *expected* true rank of the returned item (expectation over the algorithm's internal sampling, with the instance fixed). This places our work between NAS heuristics driven by cheap proxies and classical sequential-design/bandit formulations that assume repeated sam-

pling from an underlying distribution. The central distinction is that our guarantees are *instance-wise* and *distribution-free*: we do not assume any generative model for (s, f), but instead certify what randomized top-K exploitation achieves on the realized finite pool.

RoBoT-style bounds and uniform-rank assumptions. The closest conceptual precursor is RoBoT and related proxy-guided selection methods that attempt to bound the performance of proxy-based exploitation under assumptions on the relationship between proxy rank and true rank **?**. A recurring simplification in this line is a *uniformity* (or approximate uniformity) assumption on the true ranks within the proxy top set, which allows closed-form expectations but is generally unverifiable on a fixed instance. Our contribution is to replace such assumptions by a *measurable* instance quantity: the conditional rank profile $m_r = |\{a \in S_K(s) : R_f(a) \leq r\}|$. We then (a) express the expected exploitation rank exactly as a hypergeometric survival sum determined by m_r, and (b) provide a data-dependent, finite-sample certificate by lower-bounding m_r using a small number of objective queries. Thus, instead of assuming a rank model to obtain a bound, we infer a conservative surrogate of the relevant instance statistics and propagate it through an exact identity. In this sense, we view CertiTopK as a *de-assumption* of RoBoT: the bound is no longer contingent on an untestable rank-uniformity premise, but on explicit confidence events controlled by δ.

Zero-cost proxies and NAS heuristics. Zero-cost proxies and training-free predictors are widely used in NAS and related architecture screening pipelines because they enable large-scale ranking without objective training **??**. The empirical literature largely evaluates such proxies by correlation metrics (Spearman/Kendall) or by downstream performance when selecting the proxy top-k. Our results explain why correlation summaries can be insufficient for predicting exploitation outcomes: the expected rank of a randomized top-K exploitation strategy depends on the *entire* conditional profile $r \mapsto m_r$, not on a single global monotonicity score. Accordingly, we view our certificate as a practical complement to proxy design: it provides a principled way to answer, on a given pool and proxy, how many objective evaluations are required to certify that exploiting the proxy top-K will (in expectation) return an architecture within a target true-rank range. This aligns with the operational use of zero-cost proxies (screening under a strict budget) while introducing a missing reliability layer.

Partial monitoring, top-k feedback, and identifiability. Our query model is closer to *finite-population* inference than to stochastic bandits: the pool is fixed, and objective values are revealed only for queried items. This connects to partial monitoring and top-k feedback models where only lim-

ited information about the global ranking is observed **??**. A key point is identifiability: if one only ever observes the single proxy argmax (or a deterministically chosen subset), then the placement of elite items inside $S_K(s)$ is generally unidentifiable without further structure. Our approach makes the minimal exploration access explicit by requiring uniform sampling within $S_K(s)$ for conditional estimation; this is precisely what allows us to estimate (conservatively) the conditional inclusion counts m_r. We regard this as a concrete articulation of a partial-monitoring barrier in NAS-style pipelines: certificates require not only a proxy but also an exploration mechanism that reveals information about the proxy top set beyond the single best predicted architecture.

Selective inference and adaptive concentration. Because our certificate is intended to support stopping rules and anytime schedules (e.g. increasing B until a target bound is met), it interacts with adaptive data usage. There is a broad literature on selective inference and time-uniform concentration, including confidence sequences and always-valid p-values **??**. Our present development uses standard finite-sample tools (DKW bands and one-sided binomial confidence bounds) combined with an explicit union bound over a rank grid; this keeps the argument transparent and distribution-free in the fixed-pool model. We view sharper time-uniform or adaptively valid bounds as an orthogonal enhancement: replacing the union bound by a confidence sequence for the conditional membership rates would directly yield tighter certificates under repeated checks, while preserving the same basic monotonic propagation of lower bounds on m_r through the hypergeometric identity.

Certified ranking, retrieval, and top-k identification. Finally, our objective—certifying the quality of a returned item in rank units—is related to PAC-style ranking and best-arm identification **??**, as well as certified retrieval and auditing of search/ranking systems **?**. The main difference is structural: in best-arm identification one typically assumes repeated sampling with noise, whereas we assume noiseless but expensive queries on a fixed pool and focus on certifying the *expected* rank produced by a *randomized* exploitation policy constrained to $S_K(s)$. In retrieval auditing, one often certifies recall/precision of a top-k output; our certificate targets the expected minimum true rank achieved by random sub-sampling of the proxy top-K, which is the natural performance criterion for budgeted exploitation when one cannot afford to query all of $S_K(s)$. In this way, our work can be read as importing finite-population certification techniques into proxy-guided NAS exploitation, with an explicit bridge from conditional rank counts to an exact performance identity.

Discussion: beyond randomized top-K exploitation. Our development focuses on the randomized exploitation policy that samples a size-B subset uniformly from $S_K(s)$ and returns the best queried item. This particular randomization is not merely a technical convenience: it is the mechanism that makes the relevant instance statistic m_r *identifiable* from finitely many conditional queries, and it yields the exact hypergeometric identity in Theorem 1. Nevertheless, the same perspective—reduce performance to a finite-population inclusion profile, then lower-bound that profile from partial objective queries—extends to several common variants.

Deterministic greedy exploitation. A frequently used baseline is deterministic "greedy" exploitation: query the proxy top-B items (or, more generally, a fixed deterministic subset $D \subseteq S_K(s)$ of size B) and return the best among them. Here there is no internal randomness, hence no nontrivial $\mathbb{E}[\cdot]$ over the exploitation step. However, if we query all of D, then certifying the true rank of the returned item reduces to *global* rank calibration: from the global calibration sample U we can produce a high-probability rank interval for each queried $a \in D$, and hence for $\hat{a} = \arg\max_{a \in D} f(a)$. This yields an instance-wise certificate of the form $R_f(\hat{a}) \leq \overline{R}(\hat{a})$ with probability $\geq 1-\delta$, which is distinct from our expected-rank certificate but operationally similar when one wants a bound on the returned architecture itself.

If one insists on an *expected-rank* statement for greedy policies (e.g. because D is itself random due to tie-breaking, randomized proxy ensembling, or stochastic screening), then an analogue of Theorem 1 holds with K replaced by $|D|$ and m_r replaced by $m_r(D) := |\{a \in D : R_f(a) \leq r\}|$, provided the randomization makes D a uniform (or otherwise known) sampling design from a superset. In that case, the core requirement becomes a valid lower bound on $m_r(D)$ under the sampling design; uniform-without-replacement is the simplest case, while more complex designs suggest importance-weighted or Horvitz–Thompson-style conservative estimators for membership rates. We view this as a clean separation: the hypergeometric identity is specific to uniform sampling, but the "propagate a lower bound on an inclusion profile through a survival-sum identity" template is not.

Multi-objective constraints and feasibility filtering. Many NAS pipelines optimize accuracy subject to constraints (latency, memory, energy), or more generally a vector objective. One natural extension is constrained exploitation: restrict attention to the feasible subset $\mathcal{F} := \{a \in \mathcal{A} : g(a) \leq 0\}$ for some constraint function(s) g, and run our procedure on $\mathcal{A}' = \mathcal{F}$ with objective f. When constraints are also expensive to evaluate, we can treat feasibility as an additional queried attribute and certify *conditional* performance: for example, certify the expected rank among feasible candidates, $R_{f,\mathcal{F}}(\hat{a}_B)$, by calibrating ranks with respect to $\mathcal{F}$ using a uniform sample

from $\mathcal{F}$ (or a two-phase sampling scheme that first estimates $|\mathcal{F}|$ and then samples uniformly from it). If one instead wants a certificate in the *global* rank $R_f(\cdot)$ while guaranteeing feasibility with high probability, then the certificate naturally becomes two-part: (i) a lower bound on the feasible conditional profile $m_r^{\mathcal{F}} := |\{a \in S_K(s) \cap \mathcal{F} : R_f(a) \le r\}|$ and (ii) a bound on the probability that exploitation returns an infeasible item (which can be forced to zero by discarding infeasible queried items, at the cost of changing the sampling design). For genuinely multi-objective settings without a total order, one may replace rank by dominance depth or by rank induced by a fixed scalarization; our arguments remain valid as long as the performance criterion can be written as a minimum over a subset and admits a survival-sum representation with respect to a monotone threshold.

Implicit or extremely large search spaces. We have taken $\mathcal{A}$ to be a finite explicit pool with known proxy scores for all items, which matches the common "one-shot scoring then shortlist" use of zero-cost proxies. For implicit spaces where $\mathcal{A}$ is too large to enumerate, an immediate adaptation is to treat the proxy itself as defining a *proposal distribution* over candidates, and to define $S_K(s)$ implicitly as the top-K among a large proxy-sampled slate. Our certificates then apply to the realized finite slate, not to the underlying infinite space; this is appropriate when the actual decision is made from the slate. A more ambitious direction is to certify performance relative to the full implicit space; this would require additional assumptions linking the proxy-sampling mechanism to the unseen portion of the space, or an explicit exploration model that allows uniform (or otherwise controlled) sampling from progressively higher proxy quantiles.

Limitations. The present certificate is only as informative as (a) the ability to sample *uniformly* within $S_K(s)$ and (b) the tightness of global rank calibration from n_U samples. When uniform access to $S_K(s)$ is unavailable (e.g. due to determinism or constrained generation), the conditional profile m_r can fail to be identifiable, and any distribution-free guarantee becomes correspondingly weaker. Moreover, our conservative construction intentionally undercounts top-r membership; when n_U is small or f has heavy ties/noise, the resulting $\underline{m}_r$ may be too small to yield a non-vacuous U_B. Finally, our union-bound treatment over a rank grid is simple but can be loose when one checks many r values or adapts B repeatedly.

Next steps. Several improvements are conceptually straightforward within our framework: replacing the union bound by confidence sequences to support repeated stopping decisions; using stratified or diversity-aware sampling within $S_K(s)$ while retaining valid finite-population lower bounds; and exploiting monotonicity of $r \mapsto m_r$ more aggressively (e.g. isotonic tightening

of the lower confidence envelope). On the calibration side, one can replace DKW by finite-population variants (e.g. Serfling-type bounds) and incorporate variance-sensitive empirical Bernstein bounds when f is bounded. We view these as refinements rather than changes of principle: the core object remains a certified lower envelope for the conditional inclusion counts, which is then propagated through an exact or design-specific survival identity to yield an instance-wise certificate for budgeted exploitation.