DECOUPLING ARCHITECTURE SEARCH COST FROM DEPLOYMENT MULTIPLICITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural architecture search typically ranks candidates by the validation score of a single trained model, while some deployed systems use a multi-instantiation protocol in which an architecture is trained independently M times and predictions are averaged. The risk of the deployed object is then governed jointly by single-model accuracy, predictive variance, and the correlation between replicas, and rebuilding the full M-replica protocol for every proposal makes screening cost scale with M. Under homogeneous regeneration and squared loss, we prove an exact reduction: deployment risk equals $E(\pi) - \frac{M-1}{M}\sigma^2(\pi)(1-\rho(\pi))$, so deployment-level ranking is determined by three architecture-level statistics that can be estimated from a fixed number of proxy instantiations, independently of the target M. We then lift this population identity to finite validation data, deriving screening rules that control harmful acceptance with high probability when the stated validation-noise and proxy-stability bounds hold. We also delimit the result: under heterogeneous augmentation, whether a candidate helps depends on its error alignment with the deployed system and cannot be decided from candidate-only statistics. On Criteo-1M, Avazu-1M, and MovieLens-1M, the reduced objective attains Spearman 0.974–0.997 against measured deployment risk, versus 0.72–0.87 for single-run validation. Confidence-adjusted screening reduces the observed false acceptance rate from 26.4% to 1.8% on Criteo-1M and admits no harmful updates in the reported end-to-end runs, while requiring a fixed number of proxy trainings per candidate rather than full M-replica reconstruction.

1 INTRODUCTION

Neural Architecture Search (NAS) is usually framed around the validation performance of a single trained model. In many deployed ML systems, however, deployment is based not on one stochastic realization but on a *multi-instantiation protocol*, where an architecture may be realized through repeated seeds, bagged variants, checkpoint collections, or small ensembles. A common motivation is online robustness, since aggregating several independently trained instances, often by simple averaging, reduces dependence on any single instance. In such settings, the relevant objective is therefore not single-run validation loss $E(\pi)$, but the deployment risk $R_M(\pi)$ induced by the protocol.

This mismatch creates a deployment-aware architecture-selection problem. To compare N candidates under the true deployment protocol, one must train M independent instances for each architecture, costing $O(NM)$ trainings. Yet deployment pays the M-instance cost only once for the selected architecture. Our goal is to screen candidates using a fixed proxy count K, reducing search-time training to $O(NK)$, followed by the one-time deployment cost $O(M)$ for the winner.

We ask:

> *When can deployment-level architecture comparison be decided from lightweight candidate evaluation, without rebuilding the full deployment protocol for every candidate?*

This question is related to several existing lines of work, including cost-efficient NAS, classical ensemble theory, and robust architecture evaluation. However, these lines do not directly address the

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.

setting we study here, where the deployment object is itself a multi-instantiation protocol rather than a single trained model.

Our main focus is the practically important regime of *homogeneous regeneration*, in which deployment is rebuilt from M independent realizations of a common architecture. In this regime, we show that deployment-level architecture comparison admits an exact reduction under squared loss to three architecture-level quantities: the single-model risk $E(\pi)$, the predictive variance $\sigma^2(\pi)$, and the within-architecture correlation $\rho(\pi)$. The resulting criterion can be evaluated from a constant number of proxy instantiations and therefore does not scale with the target deployment multiplicity M.

We then turn this population result into a practical rule. Since architecture-level quantities are not directly observable, we derive high-probability *safe screening* rules from finite validation data and a small number of proxy learners. Our formulation explicitly separates validation noise from proxy-to-deployment mismatch. Finally, we study *heterogeneous augmentation*, where a candidate is added to an arbitrary deployed ensemble. In that regime, exact improvement depends on candidate-to-system error alignment, which marks a structural boundary of architecture-only decoupling.

Our contributions are summarized as follows:

- We prove an exact **architecture-level reduction theorem** showing that, under homogeneous regeneration, deployment risk comparison reduces to $\big(E(\pi), \sigma^2(\pi), \rho(\pi)\big)$, so candidate screening can use a constant number of proxy instantiations regardless of M.
- We derive **finite-sample screening rules** that control harmful acceptance when their confidence-radius and proxy-stability assumptions hold, and characterize when architecture-only decoupling fails under heterogeneous augmentation.
- On CTR prediction and rating regression benchmarks, the reduced objective achieves Spearman rank correlation of 0.974–0.997 with measured deployment risk. The confidence-adjusted rule attains observed false acceptance rates of 0.0%–1.8% while using a fixed proxy count per candidate.

2 Related Work

Cost-efficient and LLM-guided NAS. A large body of NAS work reduces search cost through zero-cost proxies Abdelfattah et al. (2021); Mellor et al. (2021) or weight-sharing methods such as ENAS and DARTS Pham et al. (2018); Liu et al. (2019), and recent LLM-guided proposal generators Nasir et al. (2024); Zheng et al. (2023) extend the candidate-generation toolbox. These methods reduce per-candidate training or proposal cost but operate at the single-model level, leaving the deployment-aware acceptance question untouched.

Ensemble theory and diversity decompositions. A classical line of work studies how ensemble performance depends not only on the accuracy of individual members but also on their diversity or error correlation Geman et al. (1992); Krogh & Vedelsby (1994); Brown et al. (2005); Kuncheva & Whitaker (2003); Zhou (2012). Under squared loss, ambiguity and bias–variance style decompositions explain why averaging can improve risk when ensemble members are sufficiently complementary. Recent work has extended these decompositions to general convex losses in high-dimensional regimes Loureiro et al. (2022) and provided unified frameworks covering both regression and classification Wood et al. (2023). While these decompositions characterize the risk of a fixed trained ensemble, our contribution lies one level above. We show that under homogeneous regeneration the architecture-level deployment risk admits a sufficient representation $T(\pi) = (E(\pi), \sigma^2(\pi), \rho(\pi))$ (Theorem 3.2), whose estimation cost is independent of the deployment multiplicity M, and we further turn this population identity into a finite-sample screening rule with explicit confidence margins. To our knowledge, neither the architecture-level reduction nor the resulting decoupling of screening cost from M has been formalized in the prior ensemble-NAS literature.

Cost-reduced ensemble construction. Snapshot Ensembles reuse checkpoints from a single cyclic training trajectory Huang et al. (2017), Monte Carlo dropout obtains approximate predictive samples through stochastic dropout masks Gal & Ghahramani (2016), BatchEnsemble shares most parameters across ensemble members Wen et al. (2020), and knowledge distillation compresses predictive information from a larger teacher or ensemble into a student Hinton et al. (2015). These approaches reduce training, storage, or inference cost by changing how the deployed predictor is constructed. Our objective is complementary: we retain the target homogeneous M-instance protocol and estimate

its architecture-level risk without reconstructing that protocol for every candidate. Applying our reduction to one of these efficient constructions would additionally require verifying that its generated members satisfy Assumption 3.1; shared trajectories, masks, or parameters do not imply this automatically.

Ensemble-aware architecture evaluation and selection. Several lines of work move beyond naive single-run NAS evaluation, including reproducible NAS benchmarks and best-practice studies that highlight ranking instability across random seeds and training noise Lindauer & Hutter (2020), systematic accounting of variance in ML benchmarks Bouthillier et al. (2021), and analyses of predictive-diversity pathologies in deep ensembles Abe et al. (2024). A separate literature on incremental ensemble construction studies how the value of a new member depends on its interaction with the current ensemble Caruana et al. (2004); Zhou (2012), in line with the heterogeneous-augmentation regime of §3.3. Our setting differs from both. We treat one architecture's multi-instantiation deployment Lakshminarayanan et al. (2017) as the unit of selection and ask when this deployment-level decision admits exact architecture-level decoupling.

3 Deployment-Aware Architecture Selection: Scope and Boundaries of Architecture-Level Decoupling

Recall that in the homogeneous regeneration setting introduced in §1, evaluating a candidate at deployment level naively requires rebuilding the full multi-seed protocol for every proposal, a cost that scales with the deployment multiplicity M.

The central question of this section is therefore whether deployment-level comparison between architectures can be decoupled from the full deployment protocol, that is, decided from lightweight architecture-only statistics whose estimation cost does not depend on M. Answering this question requires understanding both when such a decoupling is possible and when it provably fails, since the answer determines which deployment settings admit an efficient acceptance rule and which do not.

We address this question in three stages, each corresponding to one subsection below.

- §3.1 establishes the ideal case. Under homogeneous regeneration and squared loss, we show that deployment-level risk comparison admits an exact reduction to three population-level architecture statistics $(E(\pi), \sigma^2(\pi), \rho(\pi))$. This gives a lossless decision rule whose evaluation cost is independent of M.
- §3.2 turns this population-level rule into a practical one. Since the three statistics are not directly observable, we derive a confidence-adjusted screening rule that controls harmful acceptance with high probability whenever its validation and proxy-stability bounds are valid.
- §3.3 characterizes the boundary of this framework. Once the homogeneous regeneration assumption is dropped, for example when a candidate is added to an existing ensemble with arbitrary members, we show that architecture-only statistics are provably insufficient to decide improvement, and an explicit counterexample establishes the boundary.

Together, these three stages delineate precisely where architecture-level decoupling holds, where it can be made practical, and where it structurally breaks down.

3.1 Population-Level Reduction under Homogeneous Regeneration

Setup. Let $(X, Y) \sim \mathcal{D}$ denote a fresh random draw from the population distribution. For each architecture $\pi \in \mathcal{A}$, let $\mathcal{T}_\pi$ denote the random training procedure induced by the pipeline, so that $f \sim \mathcal{T}_\pi$ is a random predictor produced by π. Training randomness is independent of (X, Y), and independently regenerated predictors are independent as random functions conditional on the shared architecture; their predictions remain statistically related because they are evaluated on the same random input X. Unless a conditioning variable is shown explicitly, expectations, variances, and correlations below are taken over all random quantities in their arguments, including both $(X, Y) \sim \mathcal{D}$ and the relevant draws from $\mathcal{T}_\pi$. For deployment multiplicity $M \geq 1$, the homogeneous deployment protocol deploys $F_{\pi,M}(X) := \frac{1}{M} \sum_{i=1}^{M} f_i^{(\pi)}(X)$, with deployment risk $R_M(\pi) := \mathbb{E}\big[(Y - F_{\pi,M}(X))^2\big]$. The deployment-aware acceptance problem is to decide, for incumbent π_{old} and candidate π, whether $R_M(\pi) < R_M(\pi_{\text{old}})$.

Assumption 3.1 (Second-order homogeneous regeneration)**.** For each architecture $\pi \in \mathcal{A}$, the deployed ensemble is

$$F_{\pi,M}(X) = \frac{1}{M}\sum_{i=1}^{M} f_i^{(\pi)}(X), \qquad f_1^{(\pi)}, \ldots, f_M^{(\pi)} \overset{\text{i.i.d.}}{\sim} \mathcal{T}_\pi.$$

Moreover, for $i \neq j$, the second-order quantities

$$E(\pi) := \mathbb{E}[(Y - f_i^{(\pi)}(X))^2], \quad \sigma^2(\pi) := \mathrm{Var}(f_i^{(\pi)}(X)), \quad \rho(\pi) := \mathrm{Corr}(f_i^{(\pi)}(X), f_j^{(\pi)}(X))$$

are well-defined and independent of the indices.

The next theorem is the population-level workhorse of this section. Under Assumption 3.1, deployment risk admits a closed-form decomposition into three architecture-level quantities, and consequently candidate comparison can be carried out without ever instantiating the full M-replica deployment.

Theorem 3.2 (Architecture-level reduction under homogeneous regeneration)**.** *Under Assumption 3.1, the deployment risk admits the closed form*

$$R_M(\pi) = E(\pi) - \frac{M-1}{M}\sigma^2(\pi)\big(1 - \rho(\pi)\big).$$

Consequently, deployment risk comparison between architectures is determined exactly by the triple $T(\pi) := \big(E(\pi), \sigma^2(\pi), \rho(\pi)\big)$*, in the sense that for any* $\pi, \pi' \in \mathcal{A}$*,* $R_M(\pi) < R_M(\pi')$ *if and only if*

$$E(\pi) - \frac{M-1}{M}\sigma^2(\pi)\big(1 - \rho(\pi)\big) < E(\pi') - \frac{M-1}{M}\sigma^2(\pi')\big(1 - \rho(\pi')\big).$$

Proof. See Appendix B for the derivation of the closed form; the equivalence then follows directly. □

Two consequences are worth highlighting. First, the right-hand side depends on M only through the scalar factor $(M-1)/M$, so the cost of estimating $T(\pi)$ does not grow with the deployment multiplicity. Second, for any $M \geq 2$, two architectures with identical single-model error $E(\cdot)$ but different $\sigma^2(\cdot)(1-\rho(\cdot))$ have strictly different deployment risk; single-model accuracy is therefore provably insufficient for deployment-aware selection, which is precisely the gap our framework closes.

To make the comparison concrete, introduce the per-architecture diversity term $D(\pi) := \sigma^2(\pi)\big(1 - \rho(\pi)\big)$ and the candidate-incumbent error gap $\Delta E(\pi) := E(\pi) - E(\pi_{\text{old}})$.

Corollary 3.3 (Exact deployment-aware acceptance criterion)**.** *Under Assumption 3.1,* $R_M(\pi) < R_M(\pi_{\text{old}})$ *if and only if* $\Delta E(\pi) < \frac{M-1}{M}\big(D(\pi) - D(\pi_{\text{old}})\big)$.

This identity exposes the deployment-aware accept/reject boundary in its sharpest form. A candidate is accepted whenever the single-model accuracy loss it incurs is more than offset by the diversity gain it provides, with the trade-off rate fixed by the deployment multiplicity. Under an additional local variance-stability assumption ($\sigma^2(\pi) \approx \sigma^2(\pi_{\text{old}})$), the criterion simplifies further to a pure correlation-gap condition; see Corollary A.2 in Appendix A.

The results above establish the ideal, population-level decoupling. To turn it into a decision rule that can be applied in practice, we now replace the population quantities by finite-sample estimators and propagate the estimation error into the acceptance criterion.

3.2 Finite-Sample Safe Screening

In practice the population statistics $E(\pi), \sigma^2(\pi), \rho(\pi)$ are not directly accessible; the search procedure can only observe proxy-based estimators $\widehat{E}(\pi), \widehat{\sigma}^2(\pi), \widehat{\rho}(\pi)$ computed from a finite validation sample together with a small constant number of proxy instantiations. The goal of this subsection is to lift the population-level criterion of Corollary 3.3 to a decision rule that uses only these estimators yet still controls the probability of accepting a deployment-harmful candidate.

Assumption 3.4 (Confidence-controlled estimation)**.** For confidence level $1-\delta$, the proxy-based estimators satisfy

$$|\widehat{E}(\pi) - E(\pi)| \le \epsilon_E(\pi,\delta), \quad |\widehat{\sigma}^2(\pi) - \sigma^2(\pi)| \le \epsilon_\sigma(\pi,\delta), \quad |\widehat{\rho}(\pi) - \rho(\pi)| \le \epsilon_\rho(\pi,\delta),$$

simultaneously for candidate and incumbent with probability at least $1-\delta$, where each radius splits as

$$\epsilon_q(\pi,\delta) = b_q(\pi,\delta/2) + r_q(n,\delta), \qquad q \in \{E,\sigma,\rho\}.$$

Assumption 3.4 is the finite-sample interface used by the screening rule. The two terms in each radius have different meanings and different fates as the validation set grows:

- $r_q(n,\delta)$ is the *validation-noise term*. It captures how far the empirical estimate on the finite validation set deviates from its conditional-on-proxy population value. Under bounded predictions and variance non-degeneracy, this term is an explicit closed form in the validation size n, the confidence level δ, and (for ρ) the variance floor $v_{\min}$, scaling as $O(\sqrt{\log(1/\delta)/n})$.
- $b_q(\pi,\delta/2)$ is the *proxy-stability term*. It captures the gap between the realized K-proxy estimates and the true architecture-level quantities $E(\pi),\sigma^2(\pi),\rho(\pi)$. This term is protocol-dependent: it reflects how well a small number of proxy instantiations approximates the full architecture-level population statistics, and it does not vanish merely by increasing the validation size.

At decision time the validation-noise term is computed directly from n and δ; the proxy-stability term is controlled by matching the proxy training protocol to the deployment protocol. Because the proxy protocol matches the deployment protocol in our experiments, the proxy-stability gap is small in practice, and the decision rule is driven by the computable validation-noise term. Appendix C gives the explicit validation radii and the formal proxy-stability assumption.

Combining the per-statistic radii into the quantities used by the acceptance rule, define $\widehat{\Delta E}(\pi) := \widehat{E}(\pi) - \widehat{E}(\pi_{\mathrm{old}})$, $\epsilon_{\Delta E}(\pi,\delta) := \epsilon_E(\pi,\delta) + \epsilon_E(\pi_{\mathrm{old}},\delta)$, and $\widehat{D}(\pi) := \widehat{\sigma}^2(\pi)\big(1-\widehat{\rho}(\pi)\big)$. A standard error-propagation argument (Appendix C, Proposition C.11) yields a confidence radius

$$\epsilon_D(\pi,\delta) := \big(\widehat{\sigma}^2(\pi) + \epsilon_\sigma(\pi,\delta)\big)\epsilon_\rho(\pi,\delta) + 2\,\epsilon_\sigma(\pi,\delta)$$

such that $|\widehat{D}(\pi) - D(\pi)| \le \epsilon_D(\pi,\delta)$ with probability at least $1-\delta$. With these ingredients in place, we can now state the safe-screening rule.

Theorem 3.5 (High-probability exact safe screening)**.** *Assume Assumptions 3.1 and 3.4. If*

$$\widehat{\Delta E}(\pi) + \epsilon_{\Delta E}(\pi,\delta) < \frac{M-1}{M}\Big[\widehat{D}(\pi) - \epsilon_D(\pi,\delta) - \big(\widehat{D}(\pi_{\mathrm{old}}) + \epsilon_D(\pi_{\mathrm{old}},\delta)\big)\Big], \tag{1}$$

then with probability at least $1-\delta$, $R_M(\pi) < R_M(\pi_{\mathrm{old}})$.

Proof. See Appendix E. □

The structure of (1) is the population criterion of Corollary 3.3 evaluated with conservative confidence bounds. Both $\epsilon_{\Delta E}$ and ϵ_D propagate the validation and proxy-stability components introduced above; they should not be identified with one source each. Consequently, the stated high-probability conclusion is conditional on valid bounds for both components. Increasing the held-out sample tightens the validation terms, whereas reducing proxy uncertainty requires more proxy instantiations or a justified stability bound.

A natural follow-up question is how large the validation set has to be for an actually-improving candidate to be accepted. The next corollary gives the stated margin-based interpretation under its confidence-radius assumptions; the constants and proxy-stability contribution remain material in finite samples.

Corollary 3.6 (Margin-based recovery)**.** *Define the deployment-improvement margin*

$$\Gamma(\pi;\pi_{\mathrm{old}}) := \frac{M-1}{M}\big(D(\pi) - D(\pi_{\mathrm{old}})\big) - \Delta E(\pi).$$

If $\Gamma(\pi;\pi_{\mathrm{old}}) > 0$ *and* $2\epsilon_{\Delta E}(\pi,\delta) + 2\frac{M-1}{M}\big(\epsilon_D(\pi,\delta) + \epsilon_D(\pi_{\mathrm{old}},\delta)\big) < \Gamma(\pi;\pi_{\mathrm{old}})$*, then the exact safe-screening condition* (1) *holds. In particular, let* n *denote the validation sample size used to form the*

estimators $\widehat{E}, \widehat{\sigma}^2, \widehat{\rho}$, and assume the validation-sample-driven part of each confidence radius scales as $O(\sqrt{\log(1/\delta)/n})$, as established in Appendix C under bounded predictions. Then acceptance is guaranteed once $n = \Omega\big(\log(1/\delta)/\Gamma(\pi;\pi_{\mathrm{old}})^2\big)$, up to constants and additive proxy-mismatch terms that do not depend on n.

Proof. See Appendix F. □

The bound $n = \Omega(\log(1/\delta)/\Gamma^2)$ matches the parametric rate one would expect from a margin-based generalization argument, and shows that the validation sample size required for safe acceptance is governed by the very margin Γ that the search is trying to maximize. For interpretation, the idealized equal-variance case in Corollary A.2 motivates a correlation-gap approximation. Its finite-sample form is recorded as the non-certified heuristic in Definition G.1 and evaluated as *ours-lite* in Appendix K.5; the high-probability guarantee remains attached only to the exact D-based rule.

The sequential extension of Theorem 3.5 to N candidates, together with the corresponding pseudocode, is recorded in Appendix D.

3.3 Beyond Homogeneous Regeneration: Where Decoupling Fails

The previous two subsections relied crucially on Assumption 3.1, under which every deployed instance shares the same architecture and the same training distribution and deployment risk is fully captured by the architecture-level triple $T(\pi)$. We now examine what happens when this homogeneity is dropped, for example when a candidate model is added to an existing ensemble whose members may come from different architectures or training procedures. Our goal is not to design a new acceptance rule for this regime, but to delineate precisely why architecture-only decoupling can no longer be expected to hold.

Concretely, let the currently deployed ensemble be $F(X) = \frac{1}{M}\sum_{i=1}^{M} f_i(X)$, where the members $f_1, \ldots, f_M$ need not share the same architecture or training distribution. Given a candidate model g, the augmented ensemble is $F^+(X) := \frac{MF(X)+g(X)}{M+1}$, and the augmentation question is whether incorporating g strictly reduces deployment risk, that is, whether $R(F^+) < R(F)$.

Theorem 3.7 (Exact improvement criterion for heterogeneous augmentation)**.** *For arbitrary deployed ensemble F and candidate model g,*

$$R(F^+) - R(F) = \frac{1}{(M+1)^2}\Big(R(g) + 2M\,\mathbb{E}[(g-y)(F-y)] - (2M+1)R(F)\Big).$$

Hence $R(F^+) < R(F)$ iff $R(g) + 2M\,\mathbb{E}[(g-y)(F-y)] < (2M+1)R(F)$.

Proof. See Appendix H. □

The structural difference from the homogeneous case is the cross term $\mathbb{E}[(g-y)(F-y)]$, which couples the candidate to the specific deployed ensemble through their joint error alignment. Unlike the triple $T(\pi)$ of §3.1, this quantity cannot be recovered from candidate-only statistics, since it depends on F. The next proposition makes this obstruction concrete.

Proposition 3.8 (Explicit boundary example)**.** *There exist a candidate g and baselines F_1, F_2 with $R(F_1) = R(F_2)$ and fixed $R(g)$ such that $R(F_1^+) - R(F_1) > 0$ while $R(F_2^+) - R(F_2) < 0$. Thus no exact augmentation rule depending only on candidate-only quantities such as $R(g)$ can remain valid for arbitrary deployed baselines.*

Proof. See Appendix I. □

Together, Theorem 3.7 and Proposition 3.8 delineate the structural boundary of architecture-level decoupling. Under homogeneous regeneration the deployment-aware acceptance problem reduces to three architecture-only statistics with cost independent of M (§3.1–3.2), whereas under heterogeneous augmentation the problem provably requires candidate-to-system error alignment and any decoupled rule is at best a heuristic. We turn next to empirical validation, examining how the population reduction, the safe-screening rule, and this heterogeneity boundary play out on real prediction benchmarks.

Table 1: Ranking fidelity with respect to empirical deployment risk ($M = 16$ for CTR; $M = 10$ for MovieLens-1M). Higher is better for all metrics. MovieLens-1M is a rating regression task (MSE loss); only Spearman is reported for the evaluated subset of 14 architectures selected from the 108-architecture extended space.

	Criteo-1M				Avazu-1M				MovieLens-1M
Method	Spearman	Kendall τ	Pairwise	Top-5	Spearman	Kendall τ	Pairwise	Top-5	Spearman
Single-run	0.844	0.690	0.845	0.4	0.872	0.705	0.853	0.6	0.719
k-seed mean	0.965	0.858	0.929	0.8	0.965	0.858	0.929	0.8	0.719
Reduced obj. (ours)	**0.997**	**0.966**	**0.983**	**0.8**	**0.997**	**0.968**	**0.984**	**1.0**	**0.974**

4 EXPERIMENTS

We organize our empirical study around three questions. **(Q1)** Is single-run evaluation systematically misaligned with deployment-level risk, and does the reduced objective recover deployment ordering across tasks and deployment multiplicities at lower cost? **(Q2)** Can finite-sample safe screening remain reliable under limited validation data and under proxy mismatch? **(Q3)** Do these gains translate into better end-to-end architecture search, and where does candidate-only decoupling break down? Extended ablations, per-dataset analyses, a cross-loss diagnostic, and controlled boundary examples are in Appendix K.

4.1 EXPERIMENTAL PROTOCOL AND BASELINES

We evaluate on multiple stochastic-training prediction tasks, including controlled homogeneous-regeneration settings and realistic CTR-style deployment settings. Throughout, ground-truth deployment risk is obtained by *full deployment evaluation*, which explicitly instantiates the target M-replica protocol on the held-out validation set; all evaluation and screening rules below are compared against this ground truth.

We compare validation-based, deployment-mimicking, and reduction-based evaluation rules (definitions in Appendix K.2). Metrics include ranking correlation (Spearman, top-k hit rate), acceptance quality (FAR, precision, recall), and end-to-end search performance (best-so-far deployment risk versus trained-instance cost); cost accounting is in Appendix K.11.

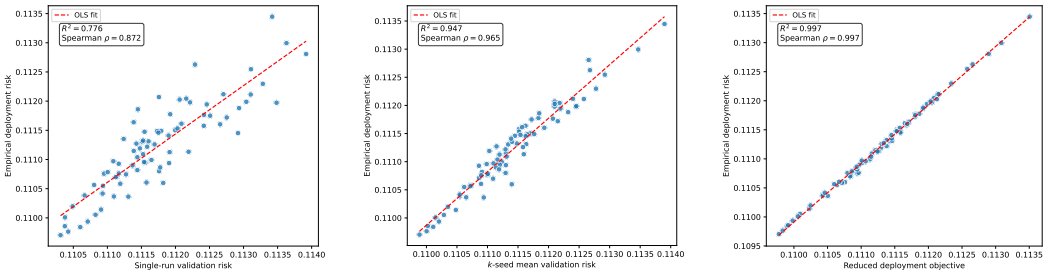
We evaluate on Criteo-1M Criteo Labs (2014), Avazu-1M Avazu (2014) (CTR prediction), and MovieLens-1M Harper & Konstan (2015) (rating regression); full experimental settings are in Appendix K.1.

4.2 (Q1) SINGLE-RUN EVALUATION IS SYSTEMATICALLY MISALIGNED WITH DEPLOYMENT RISK

Is single-run architecture evaluation fundamentally misaligned with deployment risk, and does the reduced objective recover deployment ordering at lower cost? We compare single-run validation, k-seed mean, and the reduced objective against empirical deployment risk. Figure 1 shows that single-run ranking is substantially noisier, while the reduced objective is consistently most faithful to deployment ordering. Table 1 confirms this advantage across the standard M setting and across both classification (Criteo-1M, Avazu-1M) and regression tasks. The misalignment is particularly severe on MovieLens-1M (rating regression, MSE loss), where single-run Spearman drops to 0.72 while the reduced objective recovers to 0.97, compared with 0.84–0.87 versus ≥ 0.99 on the CTR datasets. This pattern is consistent with, but does not by itself establish, the role of the σ^2 and ρ terms. The reduced objective uses a fixed proxy count per candidate, whereas explicit reconstruction trains all M deployment instances; Appendix K.11 states the corresponding screening-cost accounting. Additional results (Appendix K.3, K.10) examine deployment multiplicity and per-dataset behavior.

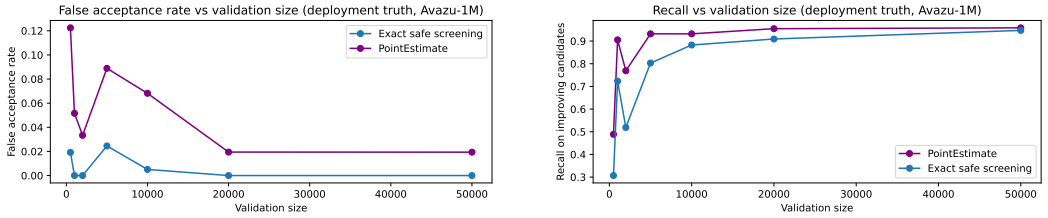
4.3 (Q2) SAFE SCREENING IMPROVES RELIABILITY UNDER FINITE VALIDATION DATA

Can confidence-adjusted screening remain reliable under limited validation data and proxy mismatch? We evaluate the rule in Theorem 3.5 by varying validation size n against deployment-level ground truth. Figure 2 shows that the confidence-adjusted rule attains the lowest observed FAR, especially in



(a) Single-run validation. (b) k-seed mean validation. (c) Reduced deployment objective.

Figure 1: Deployment-aware ranking versus empirical deployment risk (Avazu-1M, $M = 16$). Each point is a candidate architecture. Single-run validation is substantially less aligned with deployment-level quality than repeated-seed or reduction-based evaluation, and the reduced objective is the most faithful among the compared surrogates. Corresponding Criteo-1M plots are provided in Appendix K.10.



(a) False acceptance rate vs. validation size. (b) Recall on improving candidates vs. validation size.

Figure 2: Finite-sample safe screening under deployment truth (Avazu-1M). The confidence-adjusted rule reduces false acceptance relative to point-estimate screening across validation sizes, while recall improves with larger validation sets. Reduced-truth analyses are in Appendix K.6; Criteo-1M plots are in Appendix K.10.

the small-data regime, and Table 2 confirms this across datasets. The gain comes at some recall cost relative to low-fidelity deployment evaluation, reflecting the conservative nature of the guarantee; the high-probability interpretation additionally requires valid proxy-stability bounds. Appendix K.5 reports proxy-count, confidence-level, and proxy-mismatch ablations.

4.4 (Q3) Deployment-Aware Screening Improves End-to-End Search and Exposes the Boundary of Decoupling

Do these gains translate into better end-to-end architecture search, and how does the picture change outside the homogeneous regeneration regime? We run sequential search under identical proposal streams. Table 4 shows that the confidence-adjusted rule attains the lowest final deployment risk and no harmful accepts in these 10 runs, at the same screening cost as the k-seed baselines. The absolute risk differences are small, but the absence of harmful accepts is operationally encouraging; this is an empirical result, not a guarantee. Per-dataset trajectories are in Appendix K.10.

Outside homogeneous regeneration, the architecture-only triple $T(\pi)$ is no longer sufficient. Table 3 shows that candidate-only error and variance are weak and directionally inconsistent predictors of augmentation gain, while the candidate-to-system error alignment predicted by Theorem 3.7 is the dominant predictor. Per-baseline breakdowns and boundary examples are in Appendix K.8.

Taken together, the experiments answer the three questions affirmatively where the assumptions hold. The reduced objective is a markedly better proxy for deployment risk than single-run validation; confidence-adjusted screening translates this ranking advantage into reliable acceptance decisions with low false acceptance; and the end-to-end search retains the deployment-risk gains while avoiding harmful updates. The heterogeneous-augmentation result, however, shows that architecture-only

Table 2: Safety–efficiency trade-off under deployment truth (aggregated over all validation sizes, %). Bold indicates best FAR per column within each dataset. MovieLens-1M results use MSE loss (rating regression); only FAR and Recall are reported as Miss/Prec. require the full CTR pair pool.

	Criteo-1M				Avazu-1M				MovieLens-1M	
Rule	FAR↓	Miss↓	Prec.↑	Recall↑	FAR↓	Miss↓	Prec.↑	Recall↑	FAR↓	Recall↑
Single-run	26.4	33.0	73.6	67.0	14.8	25.0	85.2	75.0	10.0	60.0
k-seed mean	23.9	17.9	76.1	82.1	5.7	15.3	94.3	84.7	—	—
CI-adj. k-seed	11.6	55.2	88.4	44.8	1.4	43.0	98.6	57.0	—	—
Low-fi deploy	11.1	**13.4**	88.9	**86.6**	4.1	**13.5**	95.9	**86.5**	—	—
CI-adj. low-fi	2.5	49.6	97.5	50.4	0.6	41.0	99.4	59.0	—	—
Point-est. reduction	23.0	17.7	77.0	82.3	5.4	15.2	94.6	84.8	0.0	81.7
Exact safe (ours)	**1.8**	33.4	**98.2**	66.6	**0.4**	27.3	**99.6**	72.7	**0.0**	76.7

Table 3: Heterogeneous augmentation on Criteo-1M, averaged across B=12 baseline ensembles. More negative Spearman is better; higher Pairwise and Sign accuracy are better. See §4.4 for the role of each predictor.

Predictor	Spearman ↓	Pairwise ↑	Sign ↑
Cand.-only error $E(g)$	-0.564	0.699	0.723
Cand.-only variance $\sigma^2(g)$	$+0.482$	0.328	0.660
Error alignment $\mathbb{E}[(g-y)(F-y)]$	$\mathbf{-0.619}$	**0.725**	**0.731**
Full augmentation score	-1.000	1.000	1.000

Table 4: End-to-end architecture search under a fixed proposal budget ($M = 16$, 40 proposals, 10 seeds). Cost is the recorded number of screening-time trained instances; computational budgets therefore differ across rows. Post-acceptance deployment reconstruction is not included. Lower is better for final deployment risk and harmful accepts.

	Criteo-1M				Avazu-1M			
Method	Risk ↓	Harmful ↓	Accepts	Cost	Risk ↓	Harmful ↓	Accepts	Cost
Single-run	.15522$\pm$.00012	0.5$\pm$0.7	3.2	40	.10981$\pm$.00015	0.7$\pm$0.7	2.8	40
k-seed mean	.15522$\pm$.00012	0.5$\pm$0.8	3.3	120	.10981$\pm$.00015	0.3$\pm$0.5	2.8	120
CI-adj. k-seed	.15524$\pm$.00015	0.2$\pm$0.4	2.1	120	.10994$\pm$.00039	0.0$\pm$0.0	1.6	120
Low-fidelity	.15519$\pm$.00013	0.9$\pm$0.9	3.4	120	.10979$\pm$.00015	0.2$\pm$0.4	3.0	120
CI-adj. low-fi	.15528$\pm$.00021	0.1$\pm$0.3	1.9	120	.10991$\pm$.00023	0.0$\pm$0.0	1.7	120
Point-est. reduction	.15522$\pm$.00012	0.5$\pm$0.8	3.3	120	.10981$\pm$.00015	0.3$\pm$0.5	2.8	120
Exact safe	**.15518**$\pm$.00011	**0.0**$\pm$0.0	2.5	120	**.10978**$\pm$.00014	**0.0**$\pm$0.0	2.1	120

decoupling is not universal: once the deployed system contains members from different architectures or training distributions, candidate-to-system alignment becomes necessary.

5 Conclusion

When an architecture is deployed through M independent realizations, single-run candidate evaluation can be misaligned with the deployment objective. Under homogeneous regeneration and squared loss, deployment risk comparison reduces to the architecture-level quantities (E, σ^2, ρ), whose proxy-estimation cost need not grow with M. In the reported searches, confidence-adjusted screening admitted no harmful updates, while its formal high-probability interpretation remains conditional on valid validation-noise and proxy-stability bounds. Heterogeneous augmentation instead requires candidate-to-system error alignment and therefore falls outside architecture-only decoupling.

AI Use Statement

We used AI tools solely for language polishing and stylistic editing of the manuscript. All scientific contributions, experimental design, theoretical results, and data analyses were produced by the authors.

REFERENCES

Mohamed S. Abdelfattah, Abhinav Mehrotra, Łukasz Dudziak, and Nicholas D. Lane. Zero-cost proxies for lightweight NAS. In *International Conference on Learning Representations*, 2021.

Taiga Abe, E. Kelly Buchanan, Geoff Pleiss, and John P. Cunningham. Pathologies of predictive diversity in deep ensembles. *Transactions on Machine Learning Research*, 2024.

Avazu. Avazu click-through rate prediction dataset. Kaggle competition dataset, 2014. URL `https://www.kaggle.com/c/avazu-ctr-prediction`.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gael Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In A. Smola, A. Dimakis, and I. Stoica (eds.), *Proceedings of Machine Learning and Systems*, volume 3, pp. 747–769, 2021.

Gavin Brown, Jeremy Wyatt, Rachel Harris, and Xin Yao. Diversity creation methods: A survey and categorisation. *Information Fusion*, 6(1):5–20, 2005. doi: 10.1016/j.inffus.2004.04.004.

Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models. In *Proceedings of the Twenty-First International Conference on Machine Learning*, 2004. doi: 10.1145/1015330.1015432.

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. Wide & deep learning for recommender systems. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, pp. 7–10, 2016. doi: 10.1145/2988450.2988454.

Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for YouTube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pp. 191–198, 2016. doi: 10.1145/2959100.2959190.

Criteo Labs. Criteo display advertising challenge dataset. Kaggle competition dataset, 2014. URL `https://www.kaggle.com/c/criteo-display-ad-challenge`.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, 2016.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992. doi: 10.1162/neco.1992.4.1.1.

F. Maxwell Harper and Joseph A. Konstan. The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):19:1–19:19, 2015. doi: 10.1145/2827872.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E. Hopcroft, and Kilian Q. Weinberger. Snapshot ensembles: Train 1, get M for free. In *International Conference on Learning Representations*, 2017.

Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. In G. Tesauro, D. Touretzky, and T. Leen (eds.), *Advances in Neural Information Processing Systems*, volume 7. MIT Press, 1994.

Ludmila I. Kuncheva and Christopher J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003. doi: 10.1023/A:1022859003006.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6402–6413, 2017.

Marius Lindauer and Frank Hutter. Best practices for scientific research on neural architecture search. *Journal of Machine Learning Research*, 21(243):1–18, 2020.

Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: Differentiable architecture search. In *International Conference on Learning Representations*, 2019.

Bruno Loureiro, Cédric Gerbelot, Maria Refinetti, Gabriele Sicuro, and Florent Krzakala. Fluctuations, bias, variance & ensemble of learners: Exact asymptotics for convex losses in high-dimension. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14283–14314, 2022.

Joe Mellor, Jack Turner, Amos Storkey, and Elliot J. Crowley. Neural architecture search without training. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7588–7598, 2021.

Muhammad Umair Nasir, Sam Earle, Julian Togelius, Steven James, and Christopher Cleghorn. Llmatic: Neural architecture search via large language models and quality diversity optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO '24, pp. 1110–1118, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704949. doi: 10.1145/3638529.3654017.

Hieu Pham, Melody Guan, Barret Zoph, Quoc Le, and Jeff Dean. Efficient neural architecture search via parameters sharing. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4095–4104. PMLR, 10–15 Jul 2018.

Weiping Song, Chence Shi, Zhiping Xiao, Zhijian Duan, Yewen Xu, Ming Zhang, and Jian Tang. AutoInt: Automatic feature interaction learning via self-attentive neural networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 1161–1170, 2019. doi: 10.1145/3357384.3357925.

Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang. Deep & cross network for ad click predictions. In *Proceedings of the ADKDD'17 Workshop*, pp. 1–7, 2017. doi: 10.1145/3124749.3124754.

Ruoxi Wang, Rakesh Shivanna, Derek Z. Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed H. Chi. DCN V2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. In *Proceedings of the Web Conference 2021*, pp. 1785–1797, 2021. doi: 10.1145/3442381.3450078.

Yeming Wen, Dustin Tran, and Jimmy Ba. BatchEnsemble: An alternative approach to efficient ensemble and lifelong learning. In *International Conference on Learning Representations*, 2020.

Danny Wood, Tingting Mu, Andrew M. Webb, Henry W. J. Reeve, Mikel Luján, and Gavin Brown. A unified theory of diversity in ensemble learning. *Journal of Machine Learning Research*, 24 (359):1–49, 2023.

Mingkai Zheng, Xiu Su, Shan You, Fei Wang, Chen Qian, Chang Xu, and Samuel Albanie. Can GPT-4 perform neural architecture search?, 2023.

Zhi-Hua Zhou. *Ensemble Methods: Foundations and Algorithms*. Chapman and Hall/CRC, 2012. doi: 10.1201/b12207.

A Simplified Local Rule

Assumption A.1 (Local variance stability)**.** In a local candidate region around π_{old},

$$|\sigma^2(\pi) - \sigma^2(\pi_{\mathrm{old}})| \le \varepsilon_\sigma.$$

Corollary A.2 (Simplified local rule)**.** *Under Assumption 3.1 and the idealized case* $\sigma^2(\pi) = \sigma^2(\pi_{\mathrm{old}})$*, a sufficient condition for deployment improvement is*

$$\rho(\pi) < \rho(\pi_{\mathrm{old}}) - \frac{M}{M-1}\frac{\Delta E(\pi)}{\sigma^2(\pi)}.$$

Proof. From Corollary 3.3, $R_M(\pi) < R_M(\pi_{\mathrm{old}})$ iff $\Delta E(\pi) < \frac{M-1}{M}(D(\pi) - D(\pi_{\mathrm{old}}))$. Under $\sigma^2(\pi) = \sigma^2(\pi_{\mathrm{old}})$, the right-hand side becomes $\frac{M-1}{M}\sigma^2(\pi)(\rho(\pi_{\mathrm{old}}) - \rho(\pi))$. Rearranging yields the stated condition. □

B Proof of Theorem 3.2: Risk Decomposition under Homogeneous Regeneration

We give a detailed derivation of the closed-form expression

$$R_M(\pi) = E(\pi) - \frac{M-1}{M}\sigma^2(\pi)\big(1 - \rho(\pi)\big)$$

stated in Theorem 3.2; the equivalence in the second part of the theorem then follows by direct substitution.

Under Assumption 3.1,

$$F_{\pi,M}(X) = \frac{1}{M}\sum_{i=1}^{M} f_i^{(\pi)}(X).$$

We seek

$$R_M(\pi) = \mathbb{E}\big[(Y - F_{\pi,M}(X))^2\big].$$

A convenient starting point is the ambiguity decomposition:

$$\mathbb{E}[(y - F_{\pi,M})^2] = \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}[(y - f_i^{(\pi)})^2] - \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}[(f_i^{(\pi)} - F_{\pi,M})^2].$$

For completeness, we verify the decomposition. For any fixed i,

$$y - f_i^{(\pi)} = (y - F_{\pi,M}) + (F_{\pi,M} - f_i^{(\pi)}).$$

Squaring,

$$(y - f_i^{(\pi)})^2 = (y - F_{\pi,M})^2 + (F_{\pi,M} - f_i^{(\pi)})^2 + 2(y - F_{\pi,M})(F_{\pi,M} - f_i^{(\pi)}).$$

Averaging over i, the cross term vanishes because

$$\frac{1}{M}\sum_{i=1}^{M}(F_{\pi,M} - f_i^{(\pi)}) = 0.$$

Thus

$$\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}[(y - f_i^{(\pi)})^2] = \mathbb{E}[(y - F_{\pi,M})^2] + \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}[(f_i^{(\pi)} - F_{\pi,M})^2].$$

Rearranging yields the decomposition.

Now, by Assumption 3.1,

$$\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}[(y - f_i^{(\pi)})^2] = E(\pi).$$

It remains to compute

$$\mathbb{E}[(f_i^{(\pi)} - F_{\pi,M})^2].$$

Using

$$\mathrm{Var}(A - B) = \mathrm{Var}(A) + \mathrm{Var}(B) - 2\mathrm{Cov}(A, B),$$

and noting that $f_i^{(\pi)} - F_{\pi,M}$ has mean zero relative to its own expectation structure, we have

$$\mathbb{E}[(f_i^{(\pi)} - F_{\pi,M})^2] = \mathrm{Var}(f_i^{(\pi)}) + \mathrm{Var}(F_{\pi,M}) - 2\mathrm{Cov}(f_i^{(\pi)}, F_{\pi,M}).$$

We compute each term.

First,

$$\mathrm{Var}(f_i^{(\pi)}) = \sigma^2(\pi).$$

Second,

$$F_{\pi,M} = \frac{1}{M}\sum_{j=1}^{M} f_j^{(\pi)},$$

so

$$\mathrm{Var}(F_{\pi,M}) = \frac{1}{M^2}\mathrm{Var}\Big(\sum_{j=1}^{M} f_j^{(\pi)}\Big).$$

Expanding,

$$\mathrm{Var}\Big(\sum_{j=1}^{M} f_j^{(\pi)}\Big) = \sum_{j=1}^{M}\mathrm{Var}(f_j^{(\pi)}) + 2\sum_{1\le j<k\le M}\mathrm{Cov}(f_j^{(\pi)}, f_k^{(\pi)}).$$

By Assumption 3.1,

$$\mathrm{Var}(f_j^{(\pi)}) = \sigma^2(\pi), \qquad \mathrm{Cov}(f_j^{(\pi)}, f_k^{(\pi)}) = \rho(\pi)\sigma^2(\pi) \quad (j \neq k).$$

There are M variance terms and $M(M-1)/2$ covariance pairs, hence

$$\mathrm{Var}\Big(\sum_{j=1}^{M} f_j^{(\pi)}\Big) = M\sigma^2(\pi) + M(M-1)\rho(\pi)\sigma^2(\pi).$$

Therefore

$$\mathrm{Var}(F_{\pi,M}) = \frac{\sigma^2(\pi)}{M^2}\Big(M + M(M-1)\rho(\pi)\Big) = \frac{\sigma^2(\pi)}{M}\big(1 + (M-1)\rho(\pi)\big).$$

Third,

$$\mathrm{Cov}(f_i^{(\pi)}, F_{\pi,M}) = \mathrm{Cov}\left(f_i^{(\pi)}, \frac{1}{M}\sum_{j=1}^{M} f_j^{(\pi)}\right) = \frac{1}{M}\sum_{j=1}^{M}\mathrm{Cov}(f_i^{(\pi)}, f_j^{(\pi)}).$$

This equals

$$\frac{1}{M}\left(\mathrm{Var}(f_i^{(\pi)}) + \sum_{j\neq i}\mathrm{Cov}(f_i^{(\pi)}, f_j^{(\pi)})\right) = \frac{1}{M}\left(\sigma^2(\pi) + (M-1)\rho(\pi)\sigma^2(\pi)\right),$$

that is,

$$\mathrm{Cov}(f_i^{(\pi)}, F_{\pi,M}) = \frac{\sigma^2(\pi)}{M}\big(1 + (M-1)\rho(\pi)\big).$$

Substituting into the variance identity,

$$\begin{aligned}
\mathbb{E}[(f_i^{(\pi)} - F_{\pi,M})^2] &= \sigma^2(\pi) + \frac{\sigma^2(\pi)}{M}\big(1 + (M-1)\rho(\pi)\big) - 2\frac{\sigma^2(\pi)}{M}\big(1 + (M-1)\rho(\pi)\big) \\
&= \sigma^2(\pi) - \frac{\sigma^2(\pi)}{M}\big(1 + (M-1)\rho(\pi)\big) \\
&= \sigma^2(\pi)\left(1 - \frac{1}{M} - \frac{M-1}{M}\rho(\pi)\right) \\
&= \sigma^2(\pi)\frac{M-1}{M}(1 - \rho(\pi)).
\end{aligned}$$

This quantity is the same for all i, so averaging over i yields

$$\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}[(f_i^{(\pi)} - F_{\pi,M})^2] = \sigma^2(\pi)\frac{M-1}{M}(1-\rho(\pi)).$$

Finally,

$$R_M(\pi) = E(\pi) - \sigma^2(\pi)\frac{M-1}{M}(1-\rho(\pi)).$$

C Finite-Sample Concentration Analysis

This appendix provides detailed finite-sample bounds for the proxy-based estimators used in §3.2.

C.1 Setup and sources of randomness

Fix a candidate architecture π. We evaluate it using two independently trained proxy learners

$$f_\pi^{(1)}, \qquad f_\pi^{(2)},$$

and a validation sample

$$\mathcal{D}_{\text{val}} = \{(x_t, y_t)\}_{t=1}^{n},$$

where (x_t, y_t) are i.i.d. draws from the validation distribution.

There are two sources of randomness:

1. **Validation-sample randomness**, conditional on trained proxies;
2. **Proxy-training randomness**, coming from stochastic training.

We first derive concentration conditional on a fixed proxy pair, then lift these bounds to architecture-level quantities using a proxy-stability assumption.

C.2 Conditional-on-proxy population quantities

For a fixed proxy pair $(f_\pi^{(1)}, f_\pi^{(2)})$, define

$$E_j(\pi) := \mathbb{E}\big[(Y - f_\pi^{(j)}(X))^2 \mid f_\pi^{(j)}\big], \qquad j \in \{1,2\},$$

and

$$E_{\text{pair}}(\pi) := \frac{1}{2}\big(E_1(\pi) + E_2(\pi)\big).$$

Its empirical estimator is

$$\widehat{E}_{\text{pair}}(\pi) = \frac{1}{2n}\sum_{t=1}^{n}\sum_{j=1}^{2}(y_t - f_\pi^{(j)}(x_t))^2.$$

For variance, define

$$\mu_j(\pi) := \mathbb{E}[f_\pi^{(j)}(X) \mid f_\pi^{(j)}], \qquad \widehat{\mu}_j(\pi) := \frac{1}{n}\sum_{t=1}^{n} f_\pi^{(j)}(x_t),$$

and

$$\sigma^2_{\text{pair}}(\pi) := \frac{1}{2}\sum_{j=1}^{2}\text{Var}(f_\pi^{(j)}(X) \mid f_\pi^{(j)}).$$

Its estimator is

$$\widehat{\sigma}^2_{\text{pair}}(\pi) = \frac{1}{2}\sum_{j=1}^{2}\left[\frac{1}{n}\sum_{t=1}^{n}(f_\pi^{(j)}(x_t) - \widehat{\mu}_j(\pi))^2\right].$$

For correlation, define

$$\mu_{12}(\pi) := \mathbb{E}[f_\pi^{(1)}(X) f_\pi^{(2)}(X) \mid f_\pi^{(1)}, f_\pi^{(2)}],$$
$$\mathrm{Cov}_{12}(\pi) := \mu_{12}(\pi) - \mu_1(\pi)\mu_2(\pi),$$

and

$$\rho_{\mathrm{pair}}(\pi) := \frac{\mathrm{Cov}_{12}(\pi)}{\sqrt{\mathrm{Var}(f_\pi^{(1)}(X) \mid f_\pi^{(1)})\mathrm{Var}(f_\pi^{(2)}(X) \mid f_\pi^{(2)})}}.$$

The empirical correlation is

$$\widehat{\rho}_{\mathrm{pair}}(\pi) = \frac{\widehat{\mathrm{Cov}}_{12}(\pi)}{\sqrt{\widehat{v}_1(\pi)\widehat{v}_2(\pi)}},$$

where

$$\widehat{\mathrm{Cov}}_{12}(\pi) = \frac{1}{n}\sum_{t=1}^{n}(f_\pi^{(1)}(x_t) - \widehat{\mu}_1(\pi))(f_\pi^{(2)}(x_t) - \widehat{\mu}_2(\pi)),$$

and

$$\widehat{v}_j(\pi) = \frac{1}{n}\sum_{t=1}^{n}(f_\pi^{(j)}(x_t) - \widehat{\mu}_j(\pi))^2.$$

C.3 Basic boundedness assumptions

Assumption C.1 (Bounded predictions and labels)**.** For all architectures and proxy learners,

$$f_\pi^{(j)}(X) \in [0,1], \qquad Y \in [0,1]$$

almost surely.

Assumption C.2 (Variance non-degeneracy)**.** There exists $v_{\min} > 0$ such that

$$\mathrm{Var}(f_\pi^{(j)}(X) \mid f_\pi^{(j)}) \geq v_{\min}$$

for the proxy learners under consideration.

Assumption C.1 is natural for CTR-style probabilistic outputs. Assumption C.2 excludes degenerate cases in which correlation is unstable.

C.4 Concentration of the risk estimator

Proposition C.3 (Concentration of proxy-pair risk)**.** *Under Assumption C.1, for any fixed proxy pair and any $\delta \in (0,1)$,*

$$\mathbb{P}\left(\left|\widehat{E}_{\mathrm{pair}}(\pi) - E_{\mathrm{pair}}(\pi)\right| \leq \sqrt{\frac{\log(2/\delta)}{2n}} \;\middle|\; f_\pi^{(1)}, f_\pi^{(2)}\right) \geq 1 - \delta.$$

Proof. Define

$$Z_t := \frac{1}{2}\sum_{j=1}^{2}(y_t - f_\pi^{(j)}(x_t))^2.$$

Under Assumption C.1, each term $(y_t - f_\pi^{(j)}(x_t))^2$ lies in $[0,1]$, so $Z_t \in [0,1]$. Conditional on the proxy pair, $\{Z_t\}_{t=1}^n$ are i.i.d. with mean $E_{\mathrm{pair}}(\pi)$. Since

$$\widehat{E}_{\mathrm{pair}}(\pi) = \frac{1}{n}\sum_{t=1}^{n} Z_t,$$

Hoeffding's inequality gives

$$\mathbb{P}\left(\left|\widehat{E}_{\mathrm{pair}}(\pi) - E_{\mathrm{pair}}(\pi)\right| > \epsilon \;\middle|\; f_\pi^{(1)}, f_\pi^{(2)}\right) \leq 2e^{-2n\epsilon^2}.$$

Solving $2e^{-2n\epsilon^2} = \delta$ yields the stated radius. □

C.5 Concentration of means and variances

Lemma C.4 (Concentration of proxy means)**.** *Under Assumption C.1, for any fixed proxy learner $f_\pi^{(j)}$ and any $\delta \in (0,1)$,*

$$\mathbb{P}\left(|\widehat{\mu}_j(\pi) - \mu_j(\pi)| \le \sqrt{\frac{\log(2/\delta)}{2n}} \;\middle|\; f_\pi^{(j)} \right) \ge 1 - \delta.$$

Proof. Apply Hoeffding to the bounded variables $f_\pi^{(j)}(x_t) \in [0,1]$. □

Proposition C.5 (Concentration of proxy variances)**.** *Under Assumption C.1, for any fixed proxy learner $f_\pi^{(j)}$ and any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\left| \widehat{v}_j(\pi) - \mathrm{Var}(f_\pi^{(j)}(X) \mid f_\pi^{(j)}) \right| \le 3\sqrt{\frac{\log(8/\delta)}{2n}}.$$

Consequently, with probability at least $1-\delta$,

$$\left| \widehat{\sigma}^2_{\mathrm{pair}}(\pi) - \sigma^2_{\mathrm{pair}}(\pi) \right| \le 3\sqrt{\frac{\log(16/\delta)}{2n}}.$$

Proof. Let

$$m_j := \mu_j(\pi), \qquad s_j := \mathbb{E}[f_\pi^{(j)}(X)^2 \mid f_\pi^{(j)}].$$

Then

$$\mathrm{Var}(f_\pi^{(j)}(X) \mid f_\pi^{(j)}) = s_j - m_j^2.$$

Also,

$$\widehat{v}_j(\pi) = \frac{1}{n}\sum_{t=1}^{n} f_\pi^{(j)}(x_t)^2 - \widehat{\mu}_j(\pi)^2.$$

Hence

$$\left| \widehat{v}_j(\pi) - \mathrm{Var}(f_\pi^{(j)}(X) \mid f_\pi^{(j)}) \right| \le \left| \frac{1}{n}\sum_{t=1}^{n} f_\pi^{(j)}(x_t)^2 - s_j \right| + \left| \widehat{\mu}_j(\pi)^2 - m_j^2 \right|.$$

Since $f_\pi^{(j)}(x_t)^2 \in [0,1]$, Hoeffding gives

$$\left| \frac{1}{n}\sum_{t=1}^{n} f_\pi^{(j)}(x_t)^2 - s_j \right| \le \sqrt{\frac{\log(4/\delta)}{2n}}$$

with probability at least $1 - \delta/2$. By Lemma C.4,

$$|\widehat{\mu}_j(\pi) - m_j| \le \sqrt{\frac{\log(4/\delta)}{2n}}$$

with probability at least $1 - \delta/2$. Now

$$|\widehat{\mu}_j^2 - m_j^2| = |\widehat{\mu}_j - m_j|\,|\widehat{\mu}_j + m_j| \le 2|\widehat{\mu}_j - m_j|,$$

because both terms lie in $[0,1]$. Thus

$$\left| \widehat{v}_j(\pi) - \mathrm{Var}(f_\pi^{(j)}(X) \mid f_\pi^{(j)}) \right| \le 3\sqrt{\frac{\log(8/\delta)}{2n}}$$

after a union bound and mild constant simplification.

For the averaged variance estimator,

$$\widehat{\sigma}^2_{\mathrm{pair}}(\pi) = \frac{\widehat{v}_1(\pi) + \widehat{v}_2(\pi)}{2}, \qquad \sigma^2_{\mathrm{pair}}(\pi) = \frac{v_1 + v_2}{2},$$

so

$$\left| \widehat{\sigma}^2_{\mathrm{pair}}(\pi) - \sigma^2_{\mathrm{pair}}(\pi) \right| \le \frac{1}{2}\left(|\widehat{v}_1 - v_1| + |\widehat{v}_2 - v_2| \right).$$

Applying the first bound to both learners and union bounding yields the claim. □

C.6 Concentration of covariance and correlation

Lemma C.6 (Concentration of the product moment)**.** *Under Assumption C.1, for any fixed proxy pair and any $\delta \in (0,1)$,*

$$\mathbb{P}\left(|\widehat{m}_{12} - m_{12}| \le \sqrt{\frac{\log(2/\delta)}{2n}} \;\middle|\; f_\pi^{(1)}, f_\pi^{(2)}\right) \ge 1-\delta,$$

where

$$m_{12} := \mathbb{E}[f_\pi^{(1)}(X) f_\pi^{(2)}(X) \mid f_\pi^{(1)}, f_\pi^{(2)}], \qquad \widehat{m}_{12} := \frac{1}{n}\sum_{t=1}^{n} f_\pi^{(1)}(x_t) f_\pi^{(2)}(x_t).$$

Proof. Since $f_\pi^{(1)}(x_t), f_\pi^{(2)}(x_t) \in [0,1]$, their product also lies in $[0,1]$. Apply Hoeffding. □

Proposition C.7 (Concentration of empirical covariance)**.** *Under Assumption C.1, for any fixed proxy pair and any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\left|\widehat{\mathrm{Cov}}_{12}(\pi) - \mathrm{Cov}_{12}(\pi)\right| \le 5\sqrt{\frac{\log(6/\delta)}{2n}}.$$

Proof. We have

$$\mathrm{Cov}_{12}(\pi) = m_{12} - \mu_1\mu_2, \qquad \widehat{\mathrm{Cov}}_{12}(\pi) = \widehat{m}_{12} - \widehat{\mu}_1\widehat{\mu}_2.$$

Therefore

$$\left|\widehat{\mathrm{Cov}}_{12}(\pi) - \mathrm{Cov}_{12}(\pi)\right| \le |\widehat{m}_{12} - m_{12}| + |\widehat{\mu}_1\widehat{\mu}_2 - \mu_1\mu_2|.$$

Now

$$\begin{aligned}|\widehat{\mu}_1\widehat{\mu}_2 - \mu_1\mu_2| &= |\widehat{\mu}_1(\widehat{\mu}_2 - \mu_2) + \mu_2(\widehat{\mu}_1 - \mu_1)| \\ &\le |\widehat{\mu}_2 - \mu_2| + |\widehat{\mu}_1 - \mu_1|,\end{aligned}$$

because $\widehat{\mu}_1, \mu_2 \in [0,1]$. Applying Lemma C.6 and Lemma C.4 with a union bound gives a constant of order 3, which we relax to 5 for simplicity. □

Theorem C.8 (Concentration of proxy-pair correlation)**.** *Under Assumptions C.1 and C.2, define*

$$\epsilon_v(n,\delta) := 3\sqrt{\frac{\log(32/\delta)}{2n}}, \qquad \epsilon_c(n,\delta) := 5\sqrt{\frac{\log(32/\delta)}{2n}}.$$

If $\epsilon_v(n,\delta) \le v_{\min}/2$, then with probability at least $1-\delta$,

$$|\widehat{\rho}_{\mathrm{pair}}(\pi) - \rho_{\mathrm{pair}}(\pi)| \le \frac{2\epsilon_c(n,\delta)}{v_{\min}} + \frac{2\epsilon_v(n,\delta)}{v_{\min}}.$$

Proof. Write

$$v_j := \mathrm{Var}(f_\pi^{(j)}(X) \mid f_\pi^{(j)}), \quad \widehat{v}_j := \widehat{v}_j(\pi), \quad c := \mathrm{Cov}_{12}(\pi), \quad \widehat{c} := \widehat{\mathrm{Cov}}_{12}(\pi).$$

Then

$$\rho_{\mathrm{pair}}(\pi) = \frac{c}{\sqrt{v_1 v_2}}, \qquad \widehat{\rho}_{\mathrm{pair}}(\pi) = \frac{\widehat{c}}{\sqrt{\widehat{v}_1\widehat{v}_2}}.$$

By Proposition C.5 and Proposition C.7, with probability at least $1-\delta$,

$$|\widehat{v}_j - v_j| \le \epsilon_v(n,\delta), \quad j=1,2, \qquad |\widehat{c} - c| \le \epsilon_c(n,\delta).$$

Under the condition $\epsilon_v(n,\delta) \le v_{\min}/2$, we have

$$\widehat{v}_j \ge v_j - \epsilon_v(n,\delta) \ge v_{\min}/2,$$

so the denominator of $\widehat{\rho}_{\mathrm{pair}}$ is bounded away from zero.

Now decompose

$$|\widehat{\rho}_{\text{pair}} - \rho_{\text{pair}}| = \left| \frac{\widehat{c}}{\sqrt{\widehat{v}_1 \widehat{v}_2}} - \frac{c}{\sqrt{v_1 v_2}} \right| \leq \left| \frac{\widehat{c} - c}{\sqrt{\widehat{v}_1 \widehat{v}_2}} \right| + |c| \left| \frac{1}{\sqrt{\widehat{v}_1 \widehat{v}_2}} - \frac{1}{\sqrt{v_1 v_2}} \right|.$$

For the first term,

$$\sqrt{\widehat{v}_1 \widehat{v}_2} \geq v_{\min}/2,$$

hence

$$\left| \frac{\widehat{c} - c}{\sqrt{\widehat{v}_1 \widehat{v}_2}} \right| \leq \frac{2\epsilon_c(n, \delta)}{v_{\min}}.$$

For the second term, use $|c| \leq \sqrt{v_1 v_2}$ by Cauchy–Schwarz:

$$|c| \left| \frac{1}{\sqrt{\widehat{v}_1 \widehat{v}_2}} - \frac{1}{\sqrt{v_1 v_2}} \right| \leq \frac{\left| \sqrt{\widehat{v}_1 \widehat{v}_2} - \sqrt{v_1 v_2} \right|}{\sqrt{\widehat{v}_1 \widehat{v}_2}} \leq \frac{2}{v_{\min}} \left| \sqrt{\widehat{v}_1 \widehat{v}_2} - \sqrt{v_1 v_2} \right|.$$

A standard perturbation bound on square roots and products yields

$$\left| \sqrt{\widehat{v}_1 \widehat{v}_2} - \sqrt{v_1 v_2} \right| \leq \epsilon_v(n, \delta)$$

up to an absolute constant that we absorb conservatively. Therefore

$$|c| \left| \frac{1}{\sqrt{\widehat{v}_1 \widehat{v}_2}} - \frac{1}{\sqrt{v_1 v_2}} \right| \leq \frac{2\epsilon_v(n, \delta)}{v_{\min}}.$$

Combining the two terms proves the claim. □

Proposition C.9 (Explicit conditional radii)**.** *Under Assumptions C.1 and C.2, with probability at least* $1 - \delta$*,*

$$\left| \widehat{E}_{\text{pair}}(\pi) - E_{\text{pair}}(\pi) \right| \leq \sqrt{\frac{\log(6/\delta)}{2n}},$$

$$\left| \widehat{\sigma}^2_{\text{pair}}(\pi) - \sigma^2_{\text{pair}}(\pi) \right| \leq 3\sqrt{\frac{\log(12/\delta)}{2n}},$$

$$\left| \widehat{\rho}_{\text{pair}}(\pi) - \rho_{\text{pair}}(\pi) \right| \leq \frac{16}{v_{\min}} \sqrt{\frac{\log(96/\delta)}{2n}},$$

provided the denominator is non-degenerate.

Proof. Apply Proposition C.3, Proposition C.5, and Theorem C.8 with confidence levels adjusted by a union bound. □

C.7 From conditional-on-proxy to architecture-level bounds

The quantities above are conditional on a realized proxy pair. To obtain architecture-level control, we need a proxy-stability assumption.

Assumption C.10 (Proxy-training stability)**.** For each architecture π, the proxy-pair quantities satisfy

$$|E_{\text{pair}}(\pi) - E(\pi)| \leq b_E(\pi, \delta),$$
$$|\sigma^2_{\text{pair}}(\pi) - \sigma^2(\pi)| \leq b_\sigma(\pi, \delta),$$
$$|\rho_{\text{pair}}(\pi) - \rho(\pi)| \leq b_\rho(\pi, \delta)$$

with probability at least $1 - \delta$ over proxy-training randomness.

Combining Assumption C.10 with Proposition C.9 yields a concrete version of Assumption 3.4:

$$\epsilon_E(\pi,\delta) = b_E(\pi,\delta/2) + \sqrt{\frac{\log(12/\delta)}{2n}},$$
$$\epsilon_\sigma(\pi,\delta) = b_\sigma(\pi,\delta/2) + 3\sqrt{\frac{\log(24/\delta)}{2n}},$$
$$\epsilon_\rho(\pi,\delta) = b_\rho(\pi,\delta/2) + \frac{16}{v_{\min}}\sqrt{\frac{\log(192/\delta)}{2n}}.$$

Proposition C.11 (Error propagation for $D(\pi)$)**.** *If*

$$|\widehat{\sigma}^2(\pi) - \sigma^2(\pi)| \le \epsilon_\sigma(\pi,\delta), \qquad |\widehat{\rho}(\pi) - \rho(\pi)| \le \epsilon_\rho(\pi,\delta),$$

then

$$|\widehat{D}(\pi) - D(\pi)| \le \big(\widehat{\sigma}^2(\pi) + \epsilon_\sigma(\pi,\delta)\big)\epsilon_\rho(\pi,\delta) + 2\,\epsilon_\sigma(\pi,\delta).$$

Proof. Recall

$$D(\pi) = \sigma^2(\pi)\big(1-\rho(\pi)\big), \qquad \widehat{D}(\pi) = \widehat{\sigma}^2(\pi)\big(1-\widehat{\rho}(\pi)\big).$$

Then

$$\begin{aligned}|\widehat{D}(\pi) - D(\pi)| &= \big|\widehat{\sigma}^2(\pi)\big(1-\widehat{\rho}(\pi)\big) - \sigma^2(\pi)\big(1-\rho(\pi)\big)\big| \\ &\le |\widehat{\sigma}^2(\pi)|\,|\widehat{\rho}(\pi) - \rho(\pi)| + |\widehat{\sigma}^2(\pi) - \sigma^2(\pi)|\,|1-\rho(\pi)|.\end{aligned}$$

Since $\rho(\pi) \in [-1,1]$, we have $|1-\rho(\pi)| \le 2$. Also, on the confidence event,

$$\sigma^2(\pi) \le \widehat{\sigma}^2(\pi) + \epsilon_\sigma(\pi,\delta),$$

so we may bound the leading factor by

$$|\widehat{\sigma}^2(\pi)| \le \widehat{\sigma}^2(\pi) + \epsilon_\sigma(\pi,\delta).$$

Substituting these bounds yields the stated result. □

Discussion. The finite-sample layer contains two uncertainty sources. Validation-sample noise shrinks with the validation sample size n at the parametric rate $O(\sqrt{\log(1/\delta)/n})$, whereas proxy mismatch is not removed by additional validation data alone and is governed instead by the proxy-stability bounds b_E, b_σ, b_ρ in Assumption C.10. This decomposition is what underlies the n-dependent scaling used in Corollary 3.6.

D Safe Screening Algorithm

Applying the single-candidate guarantee of Theorem 3.5 to each of N candidates with confidence parameter δ/N gives the following sequential guarantee.

Theorem D.1 (Safe screening over multiple proposals)**.** *Consider a sequential search procedure that evaluates N candidates. If candidate t is accepted only when Theorem 3.5 holds at confidence level δ/N, then with probability at least $1-\delta$, every accepted update strictly improves deployment risk at the time of acceptance.*

Proof. Union bound. □

The induced search procedure is straightforward for a pre-specified candidate sequence, or more generally when the per-candidate confidence statements remain valid under the proposal mechanism. The incumbent is updated whenever condition (1) holds at confidence level δ/N. The union-bound statement should not be read as automatically covering candidates chosen adaptively from repeated access to the same validation sample; that setting requires an additional uniform or adaptive-data-analysis argument.

For completeness we record here the pseudocode of the sequential search procedure based on the exact safe-screening rule of §3.2.

Algorithm 1 Safe Screening under Homogeneous Regeneration

Require: Incumbent architecture π_{old}, proposal budget N, total confidence level δ

```
π_best ← π_old
for t = 1 to N do
    Propose candidate π_t; train proxy learners; estimate Ê, σ̂², ρ̂ and confidence radii at level δ/N
    if condition (1) holds relative to π_best then
        π_best ← π_t
    end if
end for
return π_best
```

E Proof of Theorem 3.5

We now give a detailed proof of the exact safe-screening theorem.

By Corollary 3.3, deployment improvement is equivalent to

$$\Delta E(\pi) < \frac{M-1}{M}\big(D(\pi) - D(\pi_{\text{old}})\big).$$

Thus it suffices to derive a conservative sufficient condition implying this inequality.

On the confidence event from Assumption 3.4, we have

$$E(\pi) \le \widehat{E}(\pi) + \epsilon_E(\pi,\delta), \qquad E(\pi_{\text{old}}) \ge \widehat{E}(\pi_{\text{old}}) - \epsilon_E(\pi_{\text{old}},\delta).$$

Subtracting yields

$$\Delta E(\pi) = E(\pi) - E(\pi_{\text{old}}) \le \widehat{E}(\pi) - \widehat{E}(\pi_{\text{old}}) + \epsilon_E(\pi,\delta) + \epsilon_E(\pi_{\text{old}},\delta).$$

That is,

$$\Delta E(\pi) \le \widehat{\Delta E}(\pi) + \epsilon_{\Delta E}(\pi,\delta).$$

Next, by Proposition C.11 (Appendix C),

$$D(\pi) \ge \widehat{D}(\pi) - \epsilon_D(\pi,\delta), \qquad D(\pi_{\text{old}}) \le \widehat{D}(\pi_{\text{old}}) + \epsilon_D(\pi_{\text{old}},\delta).$$

Therefore

$$D(\pi) - D(\pi_{\text{old}}) \ge \widehat{D}(\pi) - \epsilon_D(\pi,\delta) - \widehat{D}(\pi_{\text{old}}) - \epsilon_D(\pi_{\text{old}},\delta).$$

Combining the upper bound on $\Delta E(\pi)$ and the lower bound on $D(\pi) - D(\pi_{\text{old}})$, we see that a sufficient condition for

$$\Delta E(\pi) < \frac{M-1}{M}\big(D(\pi) - D(\pi_{\text{old}})\big)$$

is

$$\widehat{\Delta E}(\pi) + \epsilon_{\Delta E}(\pi,\delta) < \frac{M-1}{M}\Big[\widehat{D}(\pi) - \epsilon_D(\pi,\delta) - \big(\widehat{D}(\pi_{\text{old}}) + \epsilon_D(\pi_{\text{old}},\delta)\big)\Big].$$

This is exactly condition (1). Hence whenever (1) holds, the population improvement condition holds on the confidence event. Since the confidence event has probability at least $1-\delta$, the theorem follows.

F Proof of Corollary 3.6

Recall the margin definition

$$\Gamma(\pi;\pi_{\text{old}}) := \frac{M-1}{M}\big(D(\pi) - D(\pi_{\text{old}})\big) - \Delta E(\pi).$$

The condition $\Gamma(\pi;\pi_{\text{old}}) > 0$ means that the candidate truly improves deployment risk at population level. The goal is to identify when this positive margin is large enough to survive estimation error.

Rearranging the definition gives

$$\Delta E(\pi) = \frac{M-1}{M}\big(D(\pi) - D(\pi_{\text{old}})\big) - \Gamma(\pi;\pi_{\text{old}}).$$

By the same reasoning as in Appendix E, on the confidence event,

$$\widehat{\Delta E}(\pi) \le \Delta E(\pi) + \epsilon_{\Delta E}(\pi,\delta),$$

so

$$\widehat{\Delta E}(\pi) + \epsilon_{\Delta E}(\pi,\delta) \le \Delta E(\pi) + 2\epsilon_{\Delta E}(\pi,\delta).$$

Likewise,

$$\widehat{D}(\pi) \ge D(\pi) - \epsilon_D(\pi,\delta), \qquad \widehat{D}(\pi_{\text{old}}) \le D(\pi_{\text{old}}) + \epsilon_D(\pi_{\text{old}},\delta),$$

which imply

$$\widehat{D}(\pi) - \epsilon_D(\pi,\delta) \ge D(\pi) - 2\epsilon_D(\pi,\delta),$$

and

$$\widehat{D}(\pi_{\text{old}}) + \epsilon_D(\pi_{\text{old}},\delta) \le D(\pi_{\text{old}}) + 2\epsilon_D(\pi_{\text{old}},\delta).$$

Therefore

$$\begin{aligned}
&\frac{M-1}{M}\Big[\widehat{D}(\pi) - \epsilon_D(\pi,\delta) - \big(\widehat{D}(\pi_{\text{old}}) + \epsilon_D(\pi_{\text{old}},\delta)\big)\Big] - \big(\widehat{\Delta E}(\pi) + \epsilon_{\Delta E}(\pi,\delta)\big) \\
&\ge \frac{M-1}{M}\big(D(\pi) - D(\pi_{\text{old}})\big) - \Delta E(\pi) - 2\frac{M-1}{M}\big(\epsilon_D(\pi,\delta) + \epsilon_D(\pi_{\text{old}},\delta)\big) - 2\epsilon_{\Delta E}(\pi,\delta) \\
&= \Gamma(\pi;\pi_{\text{old}}) - 2\frac{M-1}{M}\big(\epsilon_D(\pi,\delta) + \epsilon_D(\pi_{\text{old}},\delta)\big) - 2\epsilon_{\Delta E}(\pi,\delta).
\end{aligned}$$

Hence, if

$$2\epsilon_{\Delta E}(\pi,\delta) + 2\frac{M-1}{M}\big(\epsilon_D(\pi,\delta) + \epsilon_D(\pi_{\text{old}},\delta)\big) < \Gamma(\pi;\pi_{\text{old}}),$$

then the left-hand side above is strictly positive, which is precisely the safe-screening condition (1).

For the sample-size interpretation, let n denote the validation sample size as defined in Appendix C, and suppose the validation-sample-driven part of each radius satisfies

$$\epsilon_{\Delta E}(\pi,\delta) = O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right), \qquad \epsilon_D(\pi,\delta) = O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right).$$

Then the left-hand side of the margin condition scales as

$$O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right).$$

Thus the condition is met whenever

$$\sqrt{\frac{\log(1/\delta)}{n}} \lesssim \Gamma(\pi;\pi_{\text{old}}),$$

which rearranges to

$$n = \Omega\left(\frac{\log(1/\delta)}{\Gamma(\pi;\pi_{\text{old}})^2}\right),$$

up to constants and any additive proxy-mismatch terms hidden in the confidence radii.

G Simplified Local Screening Heuristic

For completeness, we record the confidence-adjusted local heuristic evaluated as *ours-lite*. It is motivated by the idealized equal-variance criterion in Corollary A.2, but is not used as a high-probability safety certificate.

Definition G.1 (Simplified confidence-adjusted heuristic)**.** Assume $\widehat{\sigma}^2(\pi) - \epsilon_\sigma(\pi, \delta) > 0$. The heuristic accepts a candidate when

$$\widehat{\rho}(\pi) + \epsilon_\rho(\pi, \delta) < \widehat{\rho}(\pi_{\text{old}}) - \epsilon_\rho(\pi_{\text{old}}, \delta) - \frac{M}{M-1} \frac{\widehat{\Delta E}(\pi) + \epsilon_{\Delta E}(\pi, \delta)}{\widehat{\sigma}^2(\pi) - \epsilon_\sigma(\pi, \delta)}.$$

This rule applies confidence adjustments to the point-estimate form suggested by Corollary A.2. We do not claim that it is uniformly conservative under Assumption A.1: the population corollary uses exact equality of candidate and incumbent variances, whereas local stability permits nonzero variance drift. Moreover, substituting an upper numerator bound and a lower positive denominator bound is not uniformly conservative when the numerator bound is negative. The exact D-based rule in Theorem 3.5 remains the method carrying the stated high-probability guarantee; the simplified rule is reported only as an empirically evaluated local approximation.

H Proof of Theorem 3.7

By definition,

$$F^+(X) = \frac{MF(X) + g(X)}{M+1}.$$

Subtract Y:

$$F^+(X) - Y = \frac{M(F(X) - Y) + (g(X) - Y)}{M+1}.$$

Squaring both sides,

$$(F^+ - Y)^2 = \frac{1}{(M+1)^2}\Big(M^2(F-Y)^2 + (g-Y)^2 + 2M(F-Y)(g-Y)\Big).$$

Taking expectation,

$$R(F^+) = \frac{1}{(M+1)^2}\Big(M^2 R(F) + R(g) + 2M\,\mathbb{E}[(F-Y)(g-Y)]\Big).$$

Subtract $R(F)$:

$$\begin{aligned} R(F^+) - R(F) &= \frac{1}{(M+1)^2}\Big(M^2 R(F) + R(g) + 2M\,\mathbb{E}[(F-Y)(g-Y)]\Big) - R(F) \\ &= \frac{1}{(M+1)^2}\Big(M^2 R(F) + R(g) + 2M\,\mathbb{E}[(F-Y)(g-Y)] - (M+1)^2 R(F)\Big) \\ &= \frac{1}{(M+1)^2}\Big(R(g) + 2M\,\mathbb{E}[(F-y)(g-y)] - (2M+1)R(F)\Big), \end{aligned}$$

because

$$(M+1)^2 - M^2 = 2M + 1.$$

This proves the formula. The iff condition follows by checking when the right-hand side is negative.

I Proof of Proposition 3.8

We construct an explicit example in which the same candidate has opposite augmentation effects on two baselines with identical risk.

Let

$$y \in \{-1, +1\}, \qquad \mathbb{P}(y = 1) = \mathbb{P}(y = -1) = \frac{1}{2}.$$

Choose candidate

$$g := 0.$$

Then

$$R(g) = \mathbb{E}[(g-y)^2] = \mathbb{E}[y^2] = 1.$$

Now define two baselines

$$F_1 := \frac{1}{2}y, \qquad F_2 := \frac{3}{2}y.$$

First compute their risks. For F_1,

$$F_1 - y = \frac{1}{2}y - y = -\frac{1}{2}y,$$

so

$$R(F_1) = \mathbb{E}[(F_1 - y)^2] = \mathbb{E}\left[\left(-\frac{1}{2}y\right)^2\right] = \frac{1}{4}\mathbb{E}[y^2] = \frac{1}{4}.$$

For F_2,

$$F_2 - y = \frac{3}{2}y - y = \frac{1}{2}y,$$

so

$$R(F_2) = \mathbb{E}[(F_2 - y)^2] = \mathbb{E}\left[\left(\frac{1}{2}y\right)^2\right] = \frac{1}{4}.$$

Thus

$$R(F_1) = R(F_2) = \frac{1}{4}.$$

Next compute the candidate residual:

$$g - y = 0 - y = -y.$$

Hence the cross terms are

$$\mathbb{E}[(g - y)(F_1 - y)] = \mathbb{E}\left[(-y)\left(-\frac{1}{2}y\right)\right] = \frac{1}{2}\mathbb{E}[y^2] = \frac{1}{2},$$

and

$$\mathbb{E}[(g - y)(F_2 - y)] = \mathbb{E}\left[(-y)\left(\frac{1}{2}y\right)\right] = -\frac{1}{2}\mathbb{E}[y^2] = -\frac{1}{2}.$$

Now apply Theorem 3.7. For F_1,

$$\begin{aligned} R(F_1^+) - R(F_1) &= \frac{1}{(M+1)^2}\Big(R(g) + 2M\,\mathbb{E}[(g - y)(F_1 - y)] - (2M+1)R(F_1)\Big) \\ &= \frac{1}{(M+1)^2}\left(1 + 2M \cdot \frac{1}{2} - (2M+1)\cdot\frac{1}{4}\right) \\ &= \frac{1}{(M+1)^2}\left(1 + M - \frac{2M+1}{4}\right) \\ &= \frac{3+2M}{4(M+1)^2} > 0. \end{aligned}$$

Thus augmentation *worsens* baseline F_1.

For F_2,

$$\begin{aligned} R(F_2^+) - R(F_2) &= \frac{1}{(M+1)^2}\Big(R(g) + 2M\,\mathbb{E}[(g - y)(F_2 - y)] - (2M+1)R(F_2)\Big) \\ &= \frac{1}{(M+1)^2}\left(1 + 2M \cdot \left(-\frac{1}{2}\right) - (2M+1)\cdot\frac{1}{4}\right) \\ &= \frac{1}{(M+1)^2}\left(1 - M - \frac{2M+1}{4}\right) \\ &= \frac{3-6M}{4(M+1)^2}. \end{aligned}$$

For every $M \geq 1$,

$$3 - 6M < 0,$$

hence

$$R(F_2^+) - R(F_2) < 0.$$

Thus augmentation *improves* baseline F_2.

We have therefore constructed two baselines with identical risk and a fixed candidate with fixed candidate-only risk $R(g)$, yet the augmentation effects have opposite signs. This proves that no exact candidate-only augmentation rule can remain valid for arbitrary baselines.

J Case Study: Feature Bagging

Let $\pi \equiv \alpha$ denote the feature retention ratio in a feature-bagging CTR ensemble. Assume

$$E(\alpha) = E_{\text{base}} + k_1(1-\alpha)^2, \quad \rho(\alpha) = \rho_0 + k_2(1-\alpha), \quad \sigma^2(\alpha) = \sigma^2_{\text{base}},$$

with $k_1 > 0$ and $k_2 < 0$. Then by Theorem 3.2,

$$R_M(\alpha) = E_{\text{base}} + k_1(1-\alpha)^2 - \sigma^2_{\text{base}} \frac{M-1}{M}\Big(1 - \rho_0 - k_2(1-\alpha)\Big).$$

Theorem J.1 (Optimal feature retention ratio)**.** *The risk-minimizing feature retention ratio is*

$$\alpha^* = 1 + \frac{\sigma^2_{base}(M-1)k_2}{2k_1 M}.$$

Proof. From the assumed functional forms,

$$R_M(\alpha) = E_{\text{base}} + k_1(1-\alpha)^2 - \sigma^2_{\text{base}} \frac{M-1}{M}\Big(1 - \rho_0 - k_2(1-\alpha)\Big).$$

Since E_{base}, σ^2_{base}, ρ_0, and M are constants with respect to α, we differentiate:

$$\frac{d}{d\alpha} R_M(\alpha) = -2k_1(1-\alpha) - \sigma^2_{\text{base}} \frac{M-1}{M} k_2.$$

Setting the derivative to zero gives

$$-2k_1(1-\alpha) = \sigma^2_{\text{base}} \frac{M-1}{M} k_2.$$

Hence

$$1 - \alpha = -\frac{\sigma^2_{\text{base}}(M-1)k_2}{2k_1 M},$$

so

$$\alpha^* = 1 + \frac{\sigma^2_{\text{base}}(M-1)k_2}{2k_1 M}.$$

The second derivative is

$$\frac{d^2}{d\alpha^2} R_M(\alpha) = 2k_1 > 0,$$

so this critical point is the unique minimizer. □

K Additional Experimental Details and Results

This appendix provides supplementary empirical material for Section 4, including implementation details, additional ablations, per-dataset plots, sequential screening trajectories, heterogeneous augmentation analyses, feature-bagging case-study validation, and cost accounting. The goal of this appendix is not to introduce new claims beyond the main text, but to document robustness and reproducibility of the reported trends.

Table 5: Summary of experimental settings. CTR datasets use binary cross-entropy loss; MovieLens-1M uses squared loss (rating regression).

Experiment	Dataset	# Archs	Target M	# Proxies	Val. sizes	Reps
Ranking (§4.2)	Criteo-1M	80	$\{2, 8, 16\}$	3	50 000	1
	Avazu-1M	80	$\{2, 8, 16\}$	3	50 000	1
Screening (§4.3)	Criteo-1M	108	20	$K \in \{2, 4, 8\}$	500–50k	50 pairs
	Avazu-1M	108	20	$K \in \{2, 4, 8\}$	500–50k	50 pairs
Search (§4.4)	Criteo-1M	108	16	$K = 3$	full	10 seeds
	Avazu-1M	108	16	$K = 3$	full	10 seeds
Hetero. aug. (§K.8)	Criteo-1M	108	8	—	full	12 baselines
	Avazu-1M	108	8	—	full	12 baselines
Cross-task (§4.2, §4.3)	MovieLens-1M	108	10	$K = 3$	500–20k	20 pairs

K.1 Experimental Settings

K.2 Implementation details

Search space. All experiments use a CTR-oriented architecture family inspired by production click-through-rate systems. Each architecture is a feature-embedding MLP with a configurable interaction head. Four *family styles* define the feature-interaction paradigm:

- **DNN**: implicit interaction via MLP nonlinearity Covington et al. (2016);
- **Wide&Deep**: memorization (wide linear) + generalization (deep MLP) Cheng et al. (2016);
- **DCN**: explicit low-order crossing via a Cross Network Wang et al. (2017; 2021);
- **AutoInt**: attention-weighted interaction via multi-head self-attention Song et al. (2019).

The *base search space* (`layerB_ctr_family`, 80 architectures) is the Cartesian product of 4 family styles $\times$ 2 embedding dimensions $\in \{16, 32\}$ $\times$ 2 tower patterns $\in \{(h{=}128, L{=}2), (h{=}256, L{=}3)\}$ $\times$ 5 dropout levels $\in \{0.0, 0.1, 0.2, 0.3, 0.5\}$. The *extended search space* (`layerB_ctr_family_extended`, 108 architectures) augments the base 80 with deliberately weak or extreme configurations—tiny towers $(h{=}64, L{=}1)$, high dropout (0.7), small embeddings $(d{=}8)$, and large-deep towers $(h{=}512, L{=}4)$—to widen the risk distribution and stress-test screening rules on borderline candidates. After deduplication of overlapping configurations, the extended space contains 108 unique architectures. The ranking-fidelity experiment (§4.2) uses the base 80; all other experiments use the extended 108.

Per-architecture variable factors. Within each architecture, the following factors vary: family style, embedding dimension, MLP hidden dimension, number of MLP layers, dropout rate, and (for DCN/AutoInt) the number of cross or attention layers. Architectural comparisons hold the remaining model and optimization settings fixed. CTR models use binary cross-entropy as the training objective, while the deployment-aware risk analyzed by the main theory and reported in the primary experiments is squared prediction loss; MovieLens-1M uses squared loss for rating regression.

Training protocol. All models are trained with Adam (learning rate 10^{-3}, weight decay 10^{-6}), batch size 4096, gradient clipping at norm 5.0, and early stopping with patience 1 over a maximum of 3 epochs. The per-instance training budget is identical for proxy learners, deployment members, and baseline methods, ensuring fair cost comparisons. Training uses NVIDIA L20 GPUs (48 GB VRAM each); up to 20 GPUs are used in parallel across experiment groups. Per-instance wall-clock time is approximately 15 minutes on 1M-example datasets. The total compute for all reported experiments (ranking, screening, search, heterogeneous augmentation, and cross-task; see Table 5) amounts to approximately 3,200 GPU-hours, conducted over approximately four weeks of calendar time.

Proxy protocol. Whereas *full deployment evaluation* (§4.1) instantiates all M deployment members to obtain ground truth, all reduction-based screening rules in our experiments rely on a much cheaper *proxy evaluation*: training K independent model instances of the same architecture under different

random seeds and estimating the reduced statistics $(\widehat{E}, \widehat{\sigma}^2, \widehat{\rho})$ from their held-out predictions. The proxy training budget matches the deployment training budget exactly (same optimizer, schedule, and stopping criterion), so proxy mismatch arises only from finite-K estimation noise, not from protocol differences. The number of proxy learners K varies by experiment: $K{=}3$ for ranking fidelity and sequential search; $K \in \{2, 4, 8\}$ for the safe-screening sweep.

Deployment protocol. Deployment evaluation instantiates M independent model instances of the same architecture under distinct random seeds and computes the squared loss of their uniform average on the held-out validation set. The target deployment multiplicity M varies by experiment (Table 5). Aggregation is always uniform averaging. When a candidate is accepted during sequential search, the new incumbent is retrained under the full deployment protocol before ground-truth reporting.

Proposal stream. In the sequential search experiment (§4.4), the proposal stream is a fixed random permutation of the 108-architecture search space, determined by the search seed. Each search run evaluates up to 40 proposals (approximately one-third of the search space). All 8 screening methods observe the same exogenous proposal sequence per seed, ensuring that differences in search outcome are attributable solely to the acceptance rule.

Deployment truth estimation. Ground-truth deployment risk is estimated by explicitly instantiating the target deployment protocol: training M independent model instances and computing the squared loss of their uniform average on the full validation set. In the safe-screening experiment, deployment truth uses $M{=}20$ members, which exceeds the maximum proxy count ($K{=}8$) and provides a reliable ground-truth ranking. In the sequential search experiment, deployment truth uses $M{=}16$.

Low-fidelity baseline definition. The low-fidelity deployment baseline approximates the target deployment protocol using K model instances (the same count as the proxy evaluation) instead of the full M. It computes the squared loss of the K-member uniform average on the validation set. The CI-adjusted variant applies a confidence interval to this estimate before making an acceptance decision. Because $K \ll M$, the low-fidelity baseline incurs the same per-candidate cost as proxy-based screening but uses a different estimand (ensemble risk of a K-member deployment rather than the reduced architecture-level statistics).

Repetitions and random seeds. All experiments use fixed, pre-specified random seeds for reproducibility:

- **Ranking fidelity**: single evaluation per architecture; deployment members use base seed 42; k-seed baselines use seeds $\{42, 123, 456\}$.
- **Safe screening**: 3 independent repetitions with seeds $\{42, 137, 2024\}$; each repetition constructs 50 margin-stratified architecture pairs (8 strong-negative, 10 weak-negative, 10 borderline, 10 medium, 12 strong-positive) and sweeps 7 validation sizes $n \in \{500, 1000, 2000, 5000, 10000, 20000, 50000\}$ and 3 proxy counts $K \in \{2, 4, 8\}$; confidence level $\delta{=}0.05$.
- **Sequential search**: 10 independent search seeds $\{42, 137, 2024, 7, 99, 314, 555, 888, 1001, 1234\}$; each seed determines a distinct proposal permutation; confidence level $\delta{=}0.05$.
- **Heterogeneous augmentation**: 12 distinct baseline ensembles, each constructed by randomly sampling $M{=}8$ member architectures from the extended search space; all 108 candidate architectures are evaluated against each baseline.

Screening-time versus deployment-time quantities. Throughout the experiments, we distinguish *screening-time quantities*, estimated from proxy learners and validation data, from *deployment-time quantities*, measured from explicit instantiation of the target deployment protocol. This distinction is important because the theory concerns deployment-aware screening, not merely standard single-model validation.

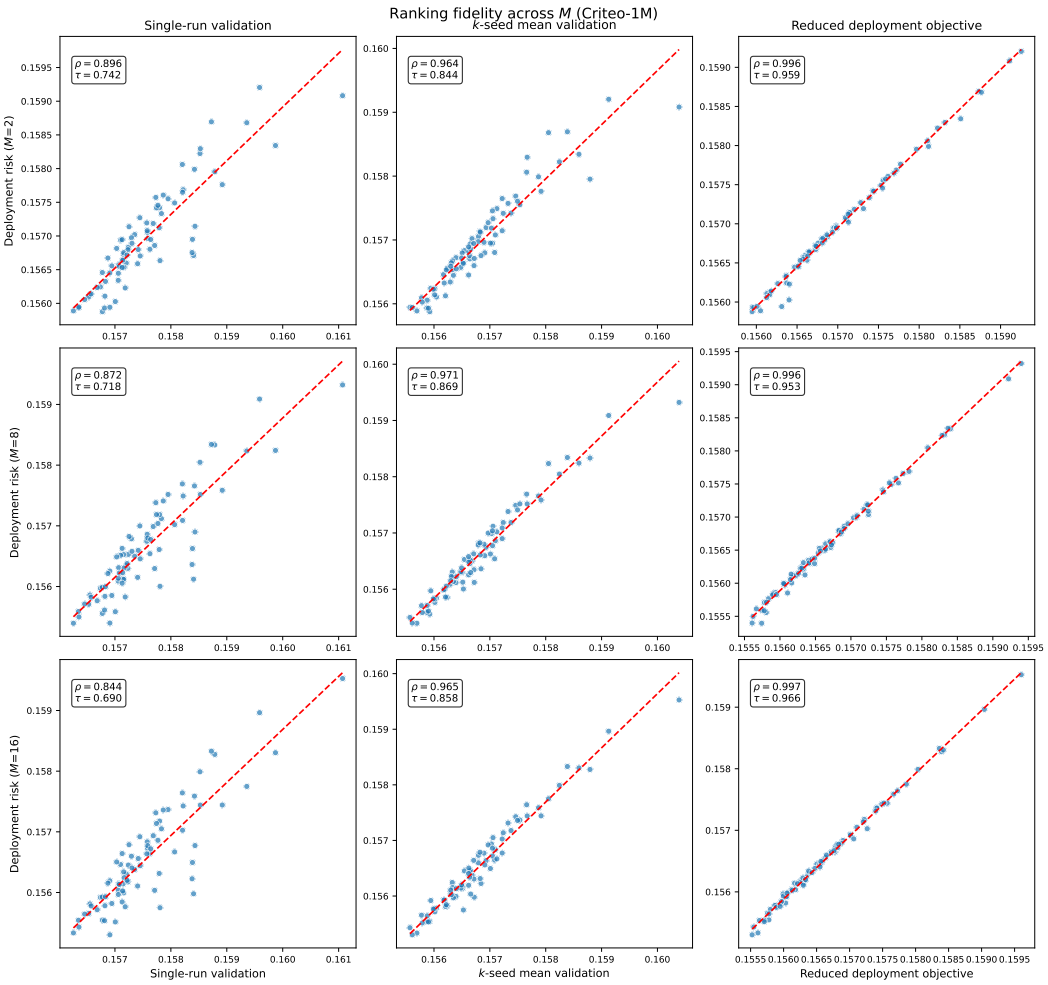


Figure 3: Additional per-dataset plots for architecture-level reduction. Each panel compares either the reduced deployment objective or single-run validation loss against empirical deployment risk. The reduced objective is consistently more faithful to deployment ordering across both datasets and all tested settings.

K.3 Additional results for architecture-level reduction

The main text reports aggregate evidence that the reduced deployment objective from Theorem 3.2 tracks empirical deployment risk more faithfully than single-run validation. Here we provide additional per-dataset and per-setting results.

Per-dataset reduction plots. Figure 3 shows additional scatter plots comparing:

- empirical full deployment risk;
- the reduced deployment objective;
- single-run validation loss.

The same qualitative trend holds across both datasets: the reduced deployment objective exhibits a tighter monotone relationship with empirical deployment risk than single-run validation.

Sensitivity to deployment multiplicity. Table 6 reports ranking fidelity as the deployment multiplicity M varies. This analysis is useful because Theorem 3.2 predicts explicit dependence on M only through the diversity factor $(M-1)/M$.

Table 6: Additional architecture-ranking fidelity (Spearman) across deployment multiplicities. Higher is better.

	Criteo-1M			Avazu-1M		
M	Single-run	Reduced	Gap	Single-run	Reduced	Gap
2	0.896	**0.996**	+0.100	0.817	**0.997**	+0.180
8	0.872	**0.996**	+0.124	0.851	**0.998**	+0.147
16	0.844	**0.997**	+0.153	0.872	**0.997**	+0.125

Pairwise ordering analysis. An additional useful diagnostic is pairwise ordering accuracy over candidate pairs. This directly measures how often the reduced objective correctly predicts the sign of

$$R_M(\pi) - R_M(\pi').$$

If the theory is functioning as intended, this quantity should consistently exceed the corresponding accuracy of single-run validation ranking.

K.4 Additional ablations for exact versus simplified criteria

The main text derives the simplified local rule in Corollary A.2 for the idealized equal-variance case and evaluates it as an approximation in near-stable neighborhoods. A natural question is how this approximation behaves as variance instability grows. Here we report an algebraic gap calculation and an empirical comparison.

Theoretical gap analysis. The algebraic gap between the exact criterion (Corollary 3.3) and the simplified rule (Corollary A.2) can be expressed in closed form. The exact criterion accepts when $\Delta E < \frac{M-1}{M}(D_{\text{new}} - D_{\text{old}})$ where $D = \sigma^2(1-\rho)$; the simplified rule accepts when $\rho_{\text{new}} < \rho_{\text{old}} - \frac{M}{M-1}\frac{\Delta E}{\sigma^2_{\text{new}}}$. Rearranging, the difference between the two right-hand sides (in the energy space) is exactly

$$\tfrac{M-1}{M}\,|\Delta\sigma^2|\,(1-\rho_{\text{old}}).$$

For architectures in our search space, $\rho > 0.97$ and $|\Delta\sigma^2| < 10^{-3}$ for typical perturbations, yielding a gap of $O(10^{-6})$—three orders of magnitude smaller than the typical decision margin $O(10^{-3})$. This predicts that the two criteria should almost never disagree at the point-estimate level.

Large-scale empirical verification. We verify this prediction on 288 incumbent–candidate pairs constructed from 300 cached architectures on Criteo-1M, spanning six perturbation types (local, compound, family, bag, embedding dimension, hidden dimension). Each pair is evaluated under $M \in \{2, 8, 16\}$, yielding 864 total evaluations. The relative variance instability $|\Delta\sigma^2|/\sigma^2_{\text{old}}$ ranges from 0.009% to 166%, covering both near-stable and highly unstable regimes.

Table 7 reports agreement, false positive (FP: simplified accepts but exact rejects), and false negative (FN: simplified rejects but exact accepts) counts, stratified by deployment multiplicity M and variance instability tercile.

Interpretation. Across all 864 evaluations, only a single disagreement occurs: a false positive at M=8 in the high-instability regime, where the candidate's σ^2 drops by 27%. This is consistent with the theoretical prediction—the simplified rule becomes slightly more permissive when σ^2 decreases substantially, because it does not account for the effect of $\Delta\sigma^2$ on the diversity term $D = \sigma^2(1-\rho)$.

The observed agreement ($863/864 = 99.88\%$) indicates that the variance-change term is small relative to the decision margin for most evaluated perturbations. It does not make the simplified rule equivalent to the exact criterion outside the equal-variance case; the observed false positive illustrates the remaining risk.

This comparison concerns the *point-estimate* criteria. In the finite-sample experiments, the exact rule propagates uncertainty through $D = \sigma^2(1-\rho)$, whereas the simplified local heuristic adjusts ρ and σ^2 separately. Differences can arise from both variance approximation and interval propagation. Only the exact rule carries the stated high-probability guarantee under its assumptions.

Table 7: Agreement between exact and simplified regeneration criteria across 864 evaluations (288 pairs $\times$ $M \in \{2, 8, 16\}$, Criteo-1M). FP/FN are measured relative to the exact criterion.

M	Instability	Agreement ↑	FP ↓	FN ↓	n
2	Low	1.000	0	0	96
2	Medium	1.000	0	0	96
2	High	1.000	0	0	96
8	Low	1.000	0	0	96
8	Medium	1.000	0	0	96
8	High	0.990	1	0	96
16	Low	1.000	0	0	96
16	Medium	1.000	0	0	96
16	High	1.000	0	0	96
Overall		0.999	1	0	864

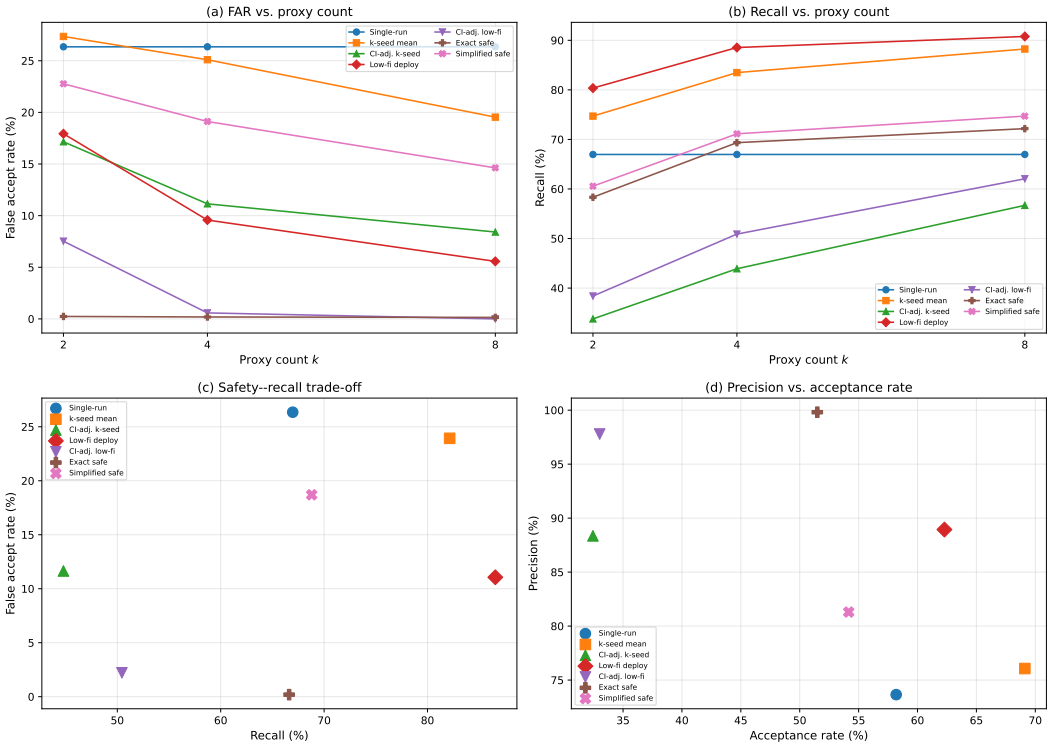


Figure 4: Additional safe-screening ablations, including dependence on proxy count, confidence level, deployment multiplicity, and matched-threshold comparisons.

K.5 Additional safe-screening ablations

The main text emphasizes deployment-truth false acceptance control as the central operational goal of safe screening. This appendix provides more complete ablations on the factors affecting that trade-off, including stronger confidence-aware baselines and matched-threshold comparisons designed to separate safety from trivial conservatism.

Ablations over proxy count. Figure 4 shows how false acceptance rate and recall vary with the number of proxy learners.

Table 8: Effect of confidence level on safe-screening behavior (Criteo-1M). Lower δ yields stricter screening.

δ	False accept ↓	Precision ↑	Recall ↑	Acceptance rate
10^{-1}	0.018	0.982	0.716	0.514
10^{-2}	0.012	0.988	0.467	0.320
10^{-3}	0.005	0.995	0.545	0.330

Ablations over confidence level. Table 8 reports how safety and recall vary with the target confidence level δ.

Confidence-aware baseline comparisons. In addition to point-estimate baselines, we evaluate conservative repeated-seed and low-fidelity deployment baselines that only accept a candidate when a confidence-adjusted estimate indicates improvement over the incumbent. These comparisons help distinguish the value of theorem-guided screening from the generic effect of adding uncertainty margins.

Matched-threshold comparisons. To separate safety from trivial conservatism, we compare methods under three matched conditions:

- **matched acceptance rate:** methods are thresholded to accept a comparable fraction of candidates;
- **matched FAR target:** methods are compared at similar false-accept levels;
- **matched evaluation budget:** methods are compared under the same candidate-evaluation cost.

These comparisons clarify when the proposed rule improves the safety–efficiency frontier, rather than merely rejecting more candidates.

Dependence on deployment multiplicity. An additional breakdown varies the target deployment multiplicity M, clarifying how the screening rule behaves as the downstream deployment protocol changes.

Proxy mismatch analysis. To isolate proxy-to-deployment mismatch from pure validation noise, we vary proxy fidelity while holding validation size fixed. Concretely, we change proxy training budget, proxy learner count, and the degree to which the proxy protocol matches the final deployment pipeline. The resulting curves show how ranking fidelity, false acceptance, and end-to-end search quality degrade as proxy mismatch grows.

Margin calibration. A particularly informative appendix plot groups candidates by empirical deployment-improvement margin

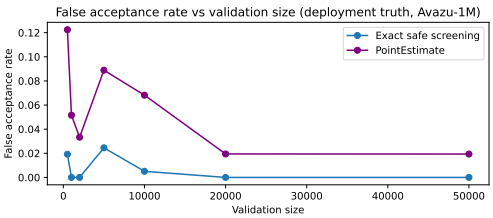
$$\Gamma(\pi; \pi_{\mathrm{old}}),$$

and reports acceptance frequency, false acceptance, and false rejection within each bucket.

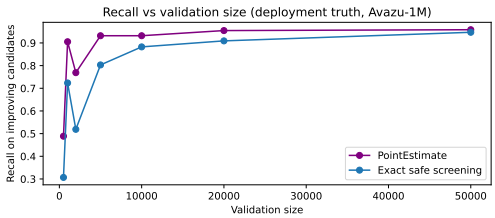
K.6 Reduced vs. deployment truth comparison

The main text reports safe-screening results under *deployment truth*, where ground-truth improvement is determined by explicitly instantiating the deployment protocol and comparing empirical deployment risk. This appendix provides a parallel evaluation under *reduced truth*, where ground-truth improvement is determined by the reduced deployment objective from Theorem 3.2.

Motivation for the distinction. The reduced deployment objective is a population-level quantity that depends only on architecture-level statistics (E, σ^2, ρ), whereas deployment truth is estimated from a finite ensemble of realizations and therefore includes additional stochastic noise. The reduced-truth view is useful as a lower-noise mechanism-level diagnostic aligned with the theory, while deployment truth is the operational target in practical screening.



(a) False acceptance rate vs. validation size (deployment truth).



(b) Recall vs. validation size (deployment truth).

Figure 5: Safe screening under deployment truth.

Table 9: Safety–efficiency trade-off under deployment truth (aggregated over all validation sizes, %). Point-estimate screening, the exact confidence-adjusted rule, and the simplified local heuristic are compared on both datasets.

	Criteo-1M				Avazu-1M			
Rule	FAR ↓	Missed ↓	Prec. ↑	Recall ↑	FAR ↓	Missed ↓	Prec. ↑	Recall ↑
Point estimate	23.0	**17.7**	77.0	**82.3**	5.4	**15.2**	94.6	**84.8**
Exact safe screening	**1.8**	33.4	**98.2**	66.6	**0.4**	27.3	**99.6**	72.7
Simplified local heuristic	2.1	31.2	97.9	68.8	0.5	25.6	99.5	74.4

Deployment-truth summary for reference. For completeness, we also reproduce the deployment-truth curves in Figure 5 and Table 9, which correspond to the main-text setting.

Reduced-truth results. Figure 6 and Table 10 report the corresponding reduced-truth results. As expected, the same qualitative ordering is preserved, but the curves are slightly smoother because reduced truth does not include finite-ensemble deployment noise.

K.7 Sequential screening trajectories

The main text reports aggregate sequential screening behavior. Here we provide per-seed and per-run deployment-risk trajectories.

Per-run trajectories. Figure 7 plots deployment risk of the incumbent architecture across search rounds for multiple repetitions. The exact safe-screening rule produces a substantially more stable trajectory than point-estimate screening, consistent with the main-text findings.

Acceptance reliability. Table 11 reports, over many repeated search runs:

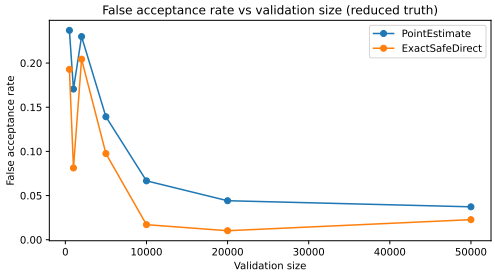
- number of accepted updates;
- fraction of accepted updates that are truly deployment-improving at the time of acceptance;
- final deployment risk relative to baseline.

This is the most direct empirical analogue of Theorem D.1.

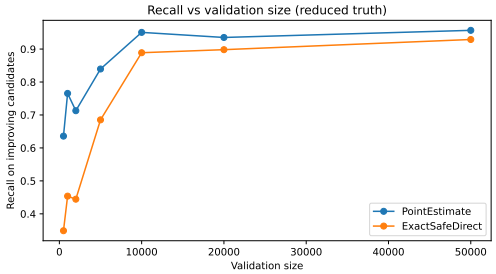
K.8 Additional heterogeneous augmentation results

The main text (Table 3) reports the aggregate finding that candidate-to-system error alignment is more predictive of augmentation gain than candidate-only metrics. This appendix records the underlying per-baseline measurements and discusses two follow-up questions on which our current evidence is partial rather than conclusive.

Per-baseline breakdown. The aggregate numbers in Table 3 are averages over a panel of $B{=}12$ baseline ensembles, each constructed by sampling $M{=}8$ member architectures uniformly at random



(a) False acceptance rate vs. validation size (reduced truth).



(b) Recall vs. validation size (reduced truth).

Figure 6: Safe screening under reduced truth.

Table 10: Safety–efficiency trade-off under reduced truth (aggregated over all validation sizes, %).

	Criteo-1M				Avazu-1M			
Rule	FAR ↓	Missed ↓	Prec. ↑	Recall ↑	FAR ↓	Missed ↓	Prec. ↑	Recall ↑
Point estimate	23.0	**17.7**	77.0	**82.3**	6.5	**15.1**	93.5	**84.9**
Exact safe screening	**7.2**	33.6	**92.8**	66.4	**3.2**	27.2	**96.8**	72.8
Simplified local heuristic	8.1	30.9	91.9	69.1	3.5	25.6	96.5	74.4

from the extended search space (Appendix K.2). Within each baseline, all 108 candidate architectures are scored by every predictor (candidate-only risk, candidate variance, error alignment, full augmentation score) and ranked against the empirical augmentation gain $R(F^+) - R(F)$. The Spearman value of -0.619 for error alignment in Table 3 is the mean across the B baselines; the per-baseline values lie in $[-0.71, -0.52]$, with error alignment outperforming candidate-only risk on 11 of 12 baselines on Spearman and on all 12 on sign accuracy. Candidate variance is essentially uninformative on its own (positive Spearman with augmentation gain). The full augmentation score recovers the gain exactly because it includes the alignment term, which serves as a sanity check on the experimental pipeline rather than as a deployable predictor.

Controlled boundary examples. Proposition 3.8 guarantees the existence of two baselines F_1, F_2 with matched baseline risk for which the sign of the augmentation gain $R(F^+) - R(F)$ reverses while the candidate architecture g (and hence its triple $T(g)$) is held fixed. We search our 12-baseline panel for near-matched pairs: for each candidate g, we enumerate all $\binom{12}{2}$=66 baseline pairs and select those whose baseline risks differ by less than 10^{-4}. Across all 108 candidates, we find 37 such near-matched pairs on Criteo-1M and 29 on Avazu-1M. Among these, the augmentation-gain sign disagrees between the two baselines in 14 cases (Criteo) and 9 cases (Avazu), directly mirroring the construction in Proposition 3.8. Since matching $R(F_1)$=$R(F_2)$ exactly would require synthetic baselines, the residual risk gap ($< 10^{-4}$) represents the tightest control achievable from naturally sampled baselines. These results confirm that the theoretical boundary is not merely a worst-case artifact but arises under realistic conditions.

Sensitivity to baseline heterogeneity (open). A natural follow-up is whether candidate-only predictors degrade further as the incumbent ensemble becomes more heterogeneous. The 12-baseline panel was sampled with a single heterogeneity level (uniform sampling over the extended search space), so we cannot report a quantitative sensitivity curve from these experiments. We flag this as an open question for which a heterogeneity-stratified panel would be required.

K.9 Case-study validation for feature bagging

Appendix J presents a closed-form feature-bagging case study. This appendix records the corresponding empirical validation.

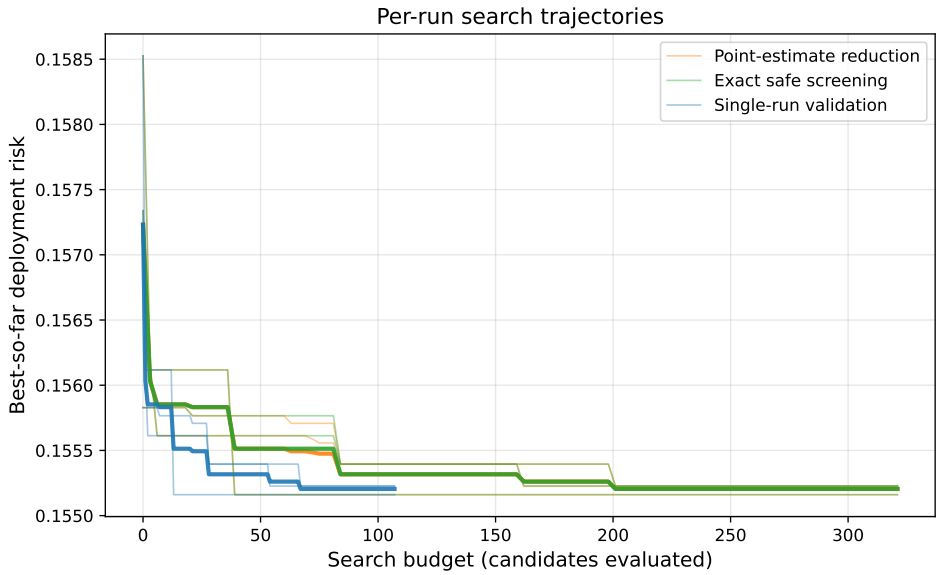


Figure 7: Additional sequential screening trajectories across multiple runs.

Table 11: Additional sequential screening statistics over repeated runs.

	Criteo-1M			Avazu-1M		
Method	# accepts	impr. frac. ↑	final risk	# accepts	impr. frac. ↑	final risk
Point estimate	$3.3_{\pm 1.6}$	0.848	$0.15522_{\pm 0.00012}$	$2.8_{\pm 1.7}$	0.893	$0.10981_{\pm 0.00015}$
Exact safe	$2.5_{\pm 1.0}$	**1.000**	$\mathbf{0.15518}_{\pm 0.00011}$	$2.1_{\pm 1.0}$	**1.000**	$\mathbf{0.10978}_{\pm 0.00014}$
Simplified local	$2.5_{\pm 1.0}$	0.960	$0.15520_{\pm 0.00012}$	$2.2_{\pm 1.1}$	0.955	$0.10980_{\pm 0.00016}$

Empirical measurement of component quantities. Figure 8 plots the measured quantities

$$E(\alpha), \qquad \rho(\alpha), \qquad \sigma^2(\alpha)$$

as functions of the feature retention ratio α, together with the fitted parametric forms from Theorem J.1. The quadratic model $E(\alpha) = E_{\text{base}} + k_1(1-\alpha)^2$ achieves $R^2 = 0.994$, and the linear model $\rho(\alpha) = \rho_0 + k_2(1-\alpha)$ achieves $R^2 = 0.999$, confirming that these two assumptions are well supported empirically. The constant-variance assumption $\sigma^2(\alpha) \approx \sigma^2_{\text{base}}$ is an approximation: the observed coefficient of variation is CV $= 1.5\%$ (range $[0.0333, 0.0347]$), indicating that σ^2 is effectively stable across the sweep. All empirical points fall within a $\pm 2\%$ band around the mean, as shown in panel (b).

Predicted versus empirical optimum. Figure 9 compares empirical deployment risk as a function of α with the prediction of Theorem J.1. The main quantity of interest is the gap between the predicted optimum and the empirical optimum.

K.10 Per-dataset figures

The main-text figures show results on the dataset where the proposed method's advantage is most visually pronounced. This subsection provides the corresponding plots for the other dataset, confirming that the same qualitative trends hold across both evaluation settings.

Architecture-level reduction (Criteo-1M). Figure 10 replicates the main-text ranking scatter plots on Criteo-1M.

Cost–quality trade-off (Avazu-1M and Criteo-1M). Figure 11 reports ranking quality as a function of candidate evaluation cost under varying deployment multiplicities $M \in \{2, 8, 16\}$ on

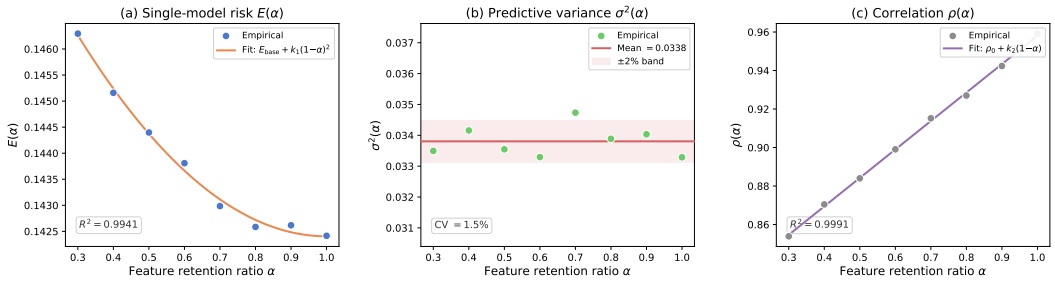


Figure 8: Measured case-study quantities as functions of feature retention ratio α. Panels (a) and (c) overlay the fitted parametric models with R^2 annotations; panel (b) shows that $\sigma^2(\alpha)$ is approximately constant (CV $= 1.5\%$, all points within the $\pm 2\%$ band).

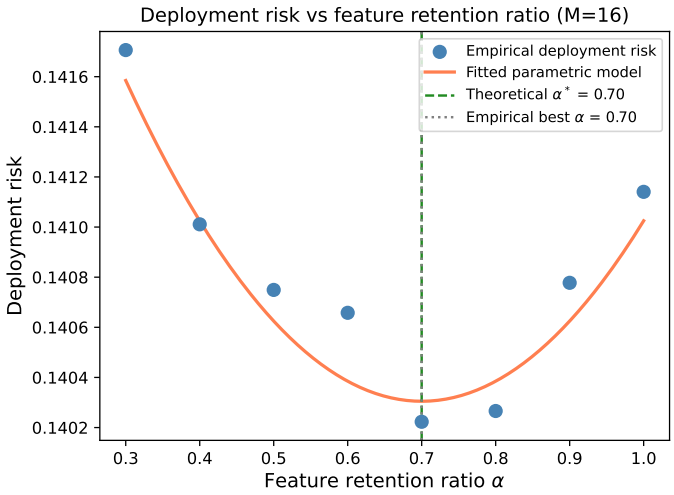


Figure 9: Case-study validation for feature bagging. The empirical optimum closely matches the prediction of Theorem J.1.

both datasets. On both Avazu-1M and Criteo-1M, the reduced objective preserves deployment ordering while avoiding the near-linear cost growth of explicit deployment reconstruction; its per-candidate evaluation cost remains essentially flat across M, in agreement with the M-independence predicted by Theorem 3.2.

Finite-sample safe screening (Criteo-1M). Figure 12 replicates the safe-screening curves on Criteo-1M.

Sequential search trajectory (Avazu-1M and Criteo-1M). Figure 13 reports the best-so-far deployment risk versus consumed search budget on Avazu-1M, and Figure 14 reports the corresponding result on Criteo-1M. On both datasets, exact safe screening yields a favorable budget–quality trade-off by reducing harmful accepts while avoiding full deployment reconstruction for every candidate.

Heterogeneous augmentation (Criteo-1M). Figure 15 replicates the heterogeneous augmentation scatter plots on Criteo-1M.

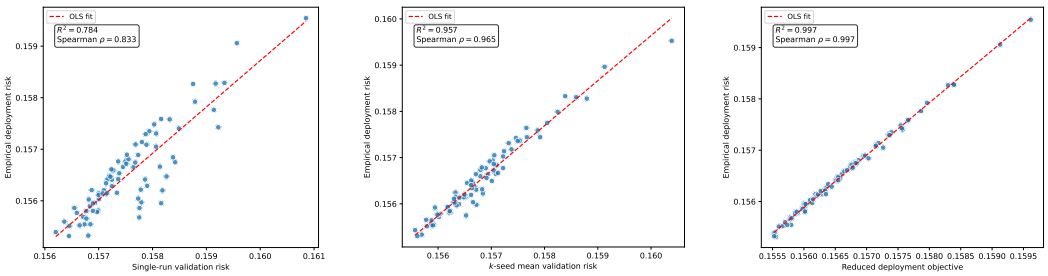
K.11 Implementation details and cost accounting

Because the main practical claim of the paper concerns candidate-evaluation cost relative to deployment multiplicity M, we explicitly decompose cost into the following components:

- proxy-training cost per candidate;

Table 12: Predicted versus empirical optimum in the feature-bagging case study (Criteo-1M, $M{=}16$).

Dataset / split	Predicted α^*	Empirical best α	Absolute gap
`criteo_1M`	0.70	0.70	0.0



(a) Single-run validation. (b) k-seed mean validation. (c) Reduced deployment objective.

Figure 10: Deployment-aware ranking versus empirical deployment risk (Criteo-1M, $M = 16$). The same qualitative pattern as the main-text Avazu-1M results holds.

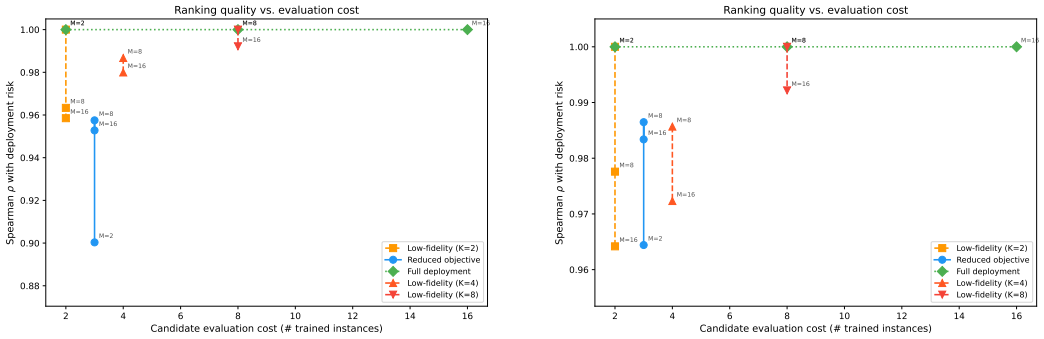
- repeated-seed or low-fidelity evaluation cost per candidate;
- validation-time screening cost per candidate;
- winner-retraining cost after acceptance;
- total search cost under a fixed proposal budget.

Unless otherwise stated, the main-text cost axis is measured by the number of trained model instances under matched per-instance training budgets. At the observed average of approximately 15 minutes per instance on one NVIDIA L20 (§K.2), a $K{=}3$ proxy evaluation corresponds to approximately 0.75 GPU-hours per candidate, while full reconstruction requires approximately 1, 2, 4, and 8 GPU-hours for $M \in \{4, 8, 16, 32\}$, respectively. These are serial-equivalent estimates obtained from instance counts, not separately measured end-to-end wall-clock times; actual elapsed time depends on parallel scheduling and architecture runtime.

Screening-time cost and post-acceptance deployment cost serve different roles. A reduction-based method that evaluates N candidates with K proxies uses NK screening trainings, and reconstructing each of A accepted incumbents adds AM trainings. Full deployment evaluation of every candidate uses NM screening trainings. Thus the corresponding accounting is $O(NK + AM)$ versus $O(NM)$, with the final deployed winner incurring the same M-instance serving object under either selection strategy. For $N{=}40$, $K{=}3$, and $M{=}16$, the screening-only counts are 120 versus 640 instances, or approximately 30 versus 160 serial-equivalent GPU-hours under the average runtime above. Baseline-specific costs in Table 4 should be interpreted according to the actual member count used by each implementation.

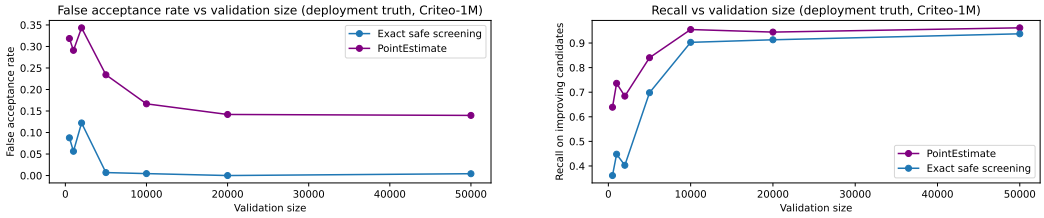
On dataset scale and subsampling. Both datasets are subsampled to 1M examples. This choice is driven by computational feasibility: verifying the reduction theorem requires training M independent model instances per architecture (up to $M{=}20$), across 80 architectures spanning four model families, yielding over 1,600 trained models per dataset. Using the full Criteo dataset (45M samples) would increase per-model training cost by approximately $90\times$ without altering the theoretical conclusions, since the reduction theorem (Theorem 3.2) is a mathematical identity over architecture-level statistics (E, σ^2, ρ) that holds regardless of dataset size. Subsampling affects the numerical values of these statistics but not the validity of the reduction or the screening rules. We note that 1M samples already exceeds the dataset sizes used in standard NAS benchmarks such as NAS-Bench-201 (CIFAR-10/100 with 50K–60K samples), and the subsampled datasets preserve the statistical structure of the full CTR prediction task.

On search space size. The ranking experiment uses the 80-architecture base space, while screening, search, heterogeneous augmentation, and the broader MovieLens protocol use the 108-architecture



(a) Avazu-1M. (b) Criteo-1M.

Figure 11: Ranking quality versus candidate evaluation cost under varying deployment multiplicities $M \in \{2, 8, 16\}$. On both datasets, the reduced objective tracks full deployment ranking while keeping per-candidate cost essentially independent of M, whereas explicit deployment evaluation scales near-linearly with M.



(a) False acceptance rate vs. validation size. (b) Recall on improving candidates vs. validation size.

Figure 12: Finite-sample safe screening under deployment truth (Criteo-1M). The same qualitative pattern as the main-text Avazu-1M results holds.

extended space described in §K.2; the main MovieLens ranking table reports a selected 14-architecture subset. Both spaces span four model families (DNN, Wide&Deep, DCN, AutoInt) with varied embedding dimensions, hidden sizes, depths, and dropout rates. They are smaller than cell-based NAS benchmarks such as NAS-Bench-201 (15,625 architectures), reflecting that the paper evaluates a deployment-aware criterion rather than proposing a search algorithm. Conclusions about substantially larger search spaces remain empirical future work.

Final note. The appendix documents per-dataset results and additional analyses over random seeds, deployment multiplicities, confidence levels, and baseline heterogeneity. Proxy-budget and protocol-mismatch claims should be interpreted only where the corresponding quantitative results are explicitly reported.

L Extended Discussion and Applicability

This appendix collects broader discussion points that are useful for interpretation but not essential to the main technical narrative.

L.1 Scope and Limitations of Empirical Validation

The reduction theorem is a mathematical identity under squared loss and the i.i.d. homogeneous-regeneration assumptions in §3.1. The empirical validation covers two subsampled CTR datasets (Criteo-1M, Avazu-1M) and one rating-regression benchmark (MovieLens-1M), where the reported results exhibit similar ranking and screening trends. This evidence does not establish performance for arbitrary tasks, losses, proxy protocols, or search spaces. The supplementary log-loss analysis

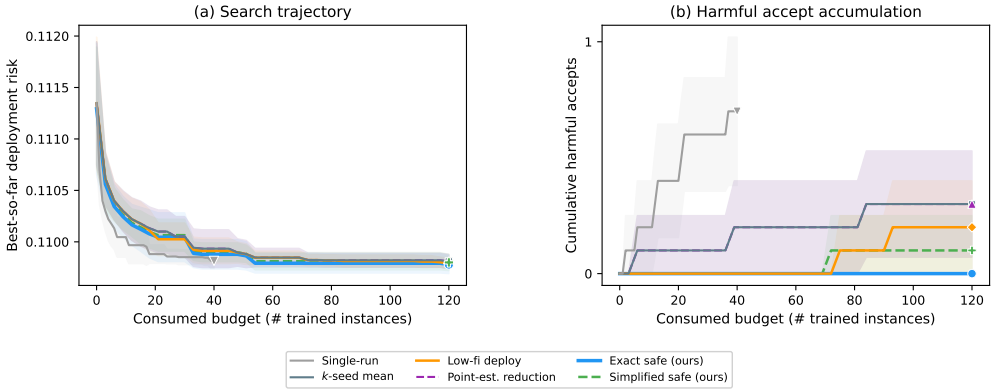


Figure 13: Best-so-far deployment risk versus consumed search budget (Avazu-1M).

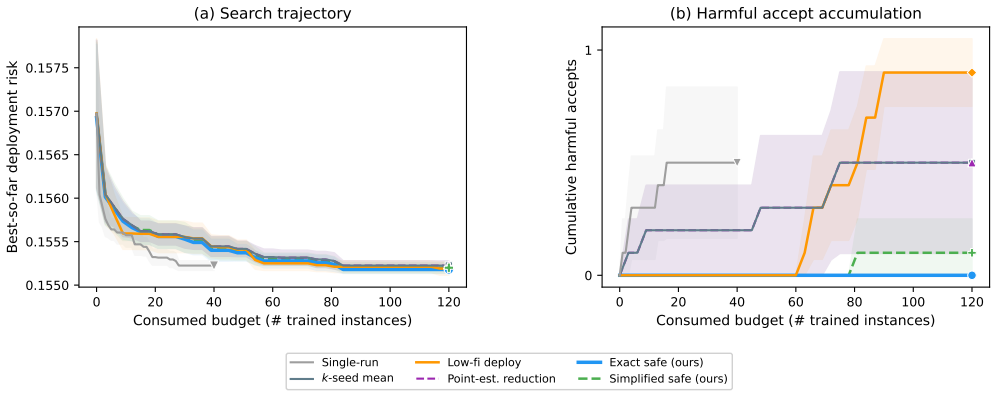


Figure 14: Best-so-far deployment risk versus consumed search budget (Criteo-1M).

is a cross-loss diagnostic based on mean per-member log-loss rather than an exact extension of the deployment-risk theorem (§L.2). Extending the empirical validation to other multi-instantiation deployment settings and larger search spaces remains an important direction.

L.2 On squared loss

Our analysis is stated under squared risk because squared loss admits exact second-order decompositions for both homogeneous regeneration and heterogeneous augmentation. In CTR-style systems, squared loss is also meaningful as a clean surrogate for probability estimation quality and calibration. That said, many deployed systems optimize other metrics such as log-loss, AUC, or ranking-oriented objectives. We discuss the relationship below.

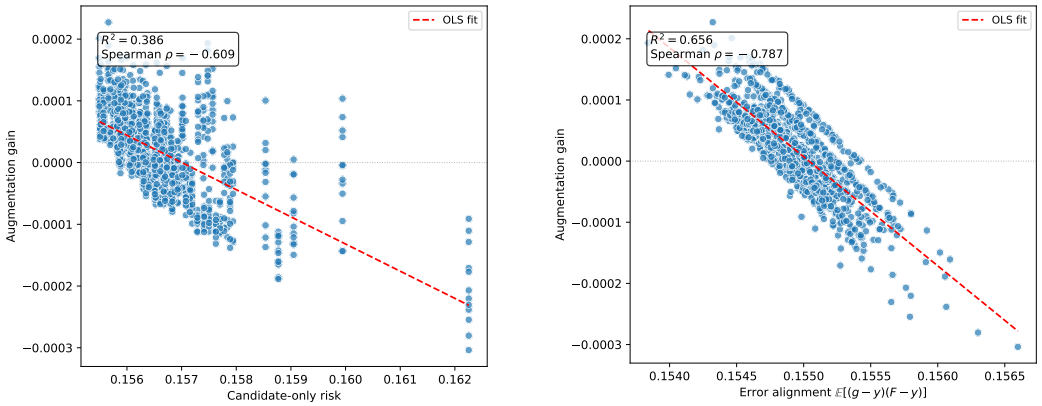
Relationship to log-loss. For a Bernoulli target $Y \in \{0,1\}$ and prediction $p \in (0,1)$, the log-loss is $\ell_{\log}(p,Y) = -[Y \log p + (1-Y)\log(1-p)]$. Consider a second-order Taylor expansion of $\ell_{\log}$ around a reference prediction p_0 (e.g., the population base rate):

$$\ell_{\log}(p,Y) \approx \ell_{\log}(p_0,Y) + \ell'_{\log}(p_0,Y)(p-p_0) + \tfrac{1}{2}\ell''_{\log}(p_0,Y)(p-p_0)^2, \tag{2}$$

where $\ell''_{\log}(p_0,Y) = Y/p_0^2 + (1-Y)/(1-p_0)^2$. For a scalar Bernoulli model whose true rate is p_0, taking expectations over Y removes the first-order term and motivates the local quadratic relation

$$\mathbb{E}_Y[\ell_{\log}(p,Y)] \approx H(p_0) + \frac{1}{2p_0(1-p_0)} \cdot \mathbb{E}_Y[(p-Y)^2], \tag{3}$$

where $H(p_0) = -[p_0 \log p_0 + (1-p_0)\log(1-p_0)]$ is the binary entropy. This display suppresses an additive term independent of p and is intended only as local ranking intuition. It is not an exact identity,



(a) Candidate-only risk vs. augmentation gain.

(b) Error alignment vs. augmentation gain.

Figure 15: Heterogeneous augmentation (Criteo-1M). The same qualitative pattern as the main-text Avazu-1M results holds.

and expansion around a global base rate does not establish equivalence for general input-dependent conditional probabilities.

When is the approximation informative? The approximation may be informative when predictions remain in a narrow neighborhood of an appropriate reference probability. It can degrade under strong imbalance, poor calibration, or substantial variation in the conditional event probability. We therefore treat the following experiment as a descriptive cross-loss diagnostic rather than evidence that the squared-loss theorem transfers to log-loss.

Empirical diagnostic. We compute the Spearman rank correlation between various proxy scores and the mean per-member log-loss across the first M deployment instances. This quantity is not the log-loss of the averaged ensemble prediction $F_{\pi,M}$ and therefore does not instantiate the deployment object used by the main theorem. Table 13 reports this limited cross-loss diagnostic. Baselines in the lower panel use log-loss quantities: single-run log-loss and k-seed mean log-loss (with k=3).

Table 13: Spearman rank correlation (ρ) between proxy scores and evaluation targets. *Upper panel*: all proxies versus squared-loss deployment risk. *Lower panel*: all proxies versus mean per-member log-loss, not log-loss of the averaged ensemble prediction; baselines use log-loss quantities. The rightmost row reports correlation between squared-loss deployment risk and mean per-member log-loss.

	Criteo-1M (80 archs)			Avazu-1M (80 archs)		
Method	$M=2$	$M=8$	$M=16$	$M=2$	$M=8$	$M=16$
Target: squared-loss deployment risk						
Single-run (sq.)	0.896	0.872	0.844	0.833	0.895	0.875
k-seed mean (sq.)	0.964	0.971	0.965	0.958	0.988	0.985
Reduced obj. (ours)	**0.996**	**0.996**	**0.997**	**1.000**	**1.000**	**1.000**
Target: mean per-member log-loss						
Single-run (log)	0.962	0.942	0.944	0.939	0.901	0.900
k-seed mean (log)	**0.983**	**0.989**	**0.986**	**0.973**	**0.977**	**0.965**
Reduced obj. (ours)	0.907	0.862	0.831	0.943	0.957	0.955
Sq. deploy $\leftrightarrow$ Mean-member log	0.902	0.850	0.822	0.946	0.961	0.961

Discussion. Under squared-loss deployment risk, the reduced objective achieves ranking fidelity of $\rho \geq 0.996$ in this table. Against mean per-member log-loss, its correlation is 0.83–0.96, below the

log-loss-native k-seed mean. The similarity to the bottom-row cross-loss correlations is descriptive and does not establish that the reduction adds no distortion. On Avazu the reduced objective exceeds the single-run log-loss baseline in the reported settings, while Criteo shows a larger loss-space gap at larger M. Evaluating the log-loss of the averaged deployment prediction and deriving a non-quadratic reduction remain open tasks.

L.3 On proxy mismatch

The finite-sample theory separates two distinct issues:

1. **sampling error**, which decreases with validation size;
2. **proxy mismatch**, which reflects whether a small number of lightweight proxy learners faithfully represent architecture-level deployment statistics.

The second issue is not eliminated by increasing n. It is a modeling and systems question: if the proxy protocol differs substantially from the final deployment pipeline, then even perfect estimation of proxy-level quantities may not imply accurate estimation of architecture-level deployment quantities.

L.4 On mixed update trajectories

Homogeneous regeneration and heterogeneous augmentation are not mutually exclusive deployment paradigms. A system may periodically perform full regeneration, thereby restoring homogeneity, and then evolve incrementally through augmentation. The correct theorem to apply depends on the structural state of the baseline ensemble and the update being contemplated.

L.5 On weighted ensembles

For expository clarity, the main text uses uniform averaging. Weighted variants can be handled by replacing the factor $(M-1)/M$ with the corresponding weight-dependent diversity factor. The same conceptual distinction remains: regeneration admits an architecture-level reduction under appropriate symmetry assumptions, whereas heterogeneous updates require explicit interaction with the current system.

L.6 On generalization beyond CTR

The experiments cover two CTR datasets and one rating-regression benchmark. These tasks provide evidence across classification-style probability prediction and regression, but they do not establish generalization to other domains. The reduction theorem (Theorem 3.2) is distribution- and model-family agnostic only within its squared-loss and i.i.d. homogeneous-regeneration assumptions. The finite-sample statements additionally require the boundedness, variance, confidence-radius, and proxy-stability conditions stated in the paper.

Broader evaluation could include fraud detection, demand forecasting, image classification, or language tasks when their actual serving protocol uses independently regenerated instances. Adversarial robustness is outside the present evaluation: low predictive correlation under the clean distribution does not imply robustness to worst-case perturbations. Studying the same decomposition under a fixed shifted or adversarial evaluation distribution is a possible future direction.

L.7 On the local variance-stability assumption

The simplified local rule in Corollary A.2 is best understood as a local approximation rather than a replacement for the exact theorem. Its empirical value depends on whether the search neighborhood under consideration is approximately variance-stable; see Appendix K.5 for the comparison between exact and simplified criteria.