AVAILABLE GUARDRAILS: CERTIFYING SELECTIVE PREDICTION ACROSS ML SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

A selective predictor can be statistically valid yet practically unavailable: with finite calibration data, it may fail to certify enough traffic at the reporting granularity required for deployment. We formalize *certificate availability*, the probability that a frozen candidate receives a valid certificate, and use exact binomial power as a planning quantity for certified service. The central empirical result is not that one partitioning algorithm dominates. The reliability order used by our contiguous dynamic program beats the median random order only 33% of the time, while adding structured candidates to a diverse random candidate pool contributes only 0.0027 mean coverage with median zero. Independent held-out comparison among diverse frozen candidates is more useful, but its gain over mean random-order balancing is still only 0.021 with median zero. These findings turn the partition optimizer into a diagnostic of population headroom and redirect the deployable story toward two stronger levers. Under fixed semantic boundaries on HWU, direct constrained planning raises certified coverage at $J = 34$ from 0.1742 to 0.3523 for DeBERTa and from 0.0861 to 0.2601 for DistilRoBERTa relative to support balancing. Separately, validity-preserving allocation of the familywise testing budget gains 0.011 mean coverage from noisy planning data, increasing to 0.014 with posterior-predictive availability. Population analysis confirms that substantially more certified traffic can exist in principle, while direct plug-in planning recovers only 0.005, identifying finite estimation rather than optimization as the bottleneck. Across four additional domains, the same safety, granularity, and traffic constraint recurs, although no single planner dominates. The contribution is therefore a deployment formulation and empirical diagnosis of certified availability, together with practical mechanisms for recovering certified service under finite evidence.

1 INTRODUCTION

Modern AI systems increasingly use selective gates that act only when predictions appear sufficiently reliable and abstain otherwise (Chow, 1970; El-Yaniv et al., 2010; Geifman & El-Yaniv, 2017). In deployment, this reliability may need to be certified separately for tools, policy categories, patient subgroups, or other reporting units. Separate certification exposes local failures that a global guarantee can hide, but it also divides finite calibration evidence across many simultaneous guarantees. This can make an otherwise useful gate statistically unavailable. For target precision 0.90, familywise failure probability 0.05, and true selective error 0.05, achieving 80% certificate availability requires 179 served calibration examples for one group and 450 per group for fifty groups. At fifty groups, increasing the true error to 0.09 raises the requirement to $13{,}407$. The deployment problem is therefore not only whether a returned certificate is valid, but whether finite data can return certificates for enough traffic at the requested reporting granularity.

Existing finite-sample risk-control methods primarily specify when a returned policy can be trusted (Bates et al., 2021; Angelopoulos et al., 2021; 2024). We study the complementary pre-certification design problem: given a safety target, label budget, familywise error level, and required number of reporting units, how much traffic can be certified, and which design choices improve that quantity? We call this probability-weighted service *certificate availability*. Exact Clopper–Pearson inversion makes the availability of a frozen candidate computable, so classical binomial power can be used as a planning quantity. The statistical identity and the dynamic program used later are not themselves

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.

new. The contribution is to make availability a first-class deployment objective, expose the safety, granularity, and traffic frontier induced by finite evidence, and determine empirically which planning mechanisms survive finite estimation.

That distinction matters because the most obvious optimizer does not survive intact. With population quantities, a reliability-ordered partition optimizer reveals substantial opportunity, improving mean certified coverage by 0.1568 over support balancing. A direct plug-in version recovers only 0.0050. The structural prior behind the optimizer is also weak: the estimated reliability order beats the median random order only 33% of the time, and in a matched candidate pool the structured candidates add only 0.0027 mean coverage with median zero. Held-out comparison among diverse frozen candidates is more useful, but its gain over mean random-order balancing is still only 0.0208 with median zero. We therefore use the population optimizer primarily as a diagnostic of available certified traffic and treat the finite-data failures as a result rather than as a caveat.

The experiments point to two stronger deployment levers. First, when semantic boundaries are fixed by the application, directly optimizing within the admissible constrained space can substantially increase certified service. On HWU at $J = 34$, coverage rises from 0.1742 to 0.3523 for DeBERTa and from 0.0861 to 0.2601 for DistilRoBERTa relative to support balancing, while preserving named scenario boundaries. Second, the familywise error budget itself is a planning variable: reallocating it across frozen units gains 0.011 mean coverage from noisy planning data, increasing to 0.014 with posterior-predictive availability, without changing the union-bound validity guarantee. Studies across independently trained intent models, controlled shift, and four additional application domains then separate what is general from what is method-specific: the certified service frontier recurs broadly, but the best recovery strategy depends on score quality, admissible structure, candidate diversity, and finite support. Our central claim is therefore about the deployment problem and the mechanisms that recover certified service, not universal superiority of the reliability-ordered planner.

2 Related Work

Finite-sample certification and power. The statistical tools for certifying a fixed proportion from finite labels are classical. Exact confidence intervals and binomial power analyses (Clopper & Pearson, 1934; Brown et al., 2001; Krishnamoorthy & Peng, 2007; Thulin, 2014) provide finite-sample guarantees for a single proportion, while Learn-Then-Test and conformal risk control extend finite-sample validity to black-box policy selection and broader risk targets (Bates et al., 2021; Angelopoulos et al., 2021; 2024). Related work also studies validity under distribution shift. Covariate-shift conformal methods reweight examples under known changes in the input distribution (Tibshirani et al., 2019), LEC imposes a linear expectation constraint (Wang et al., 2025), SCoRE controls discovery risks through e-values (Bai & Jin, 2026), and COIN constructs thresholds for selective question answering (Wang et al., 2026). This literature does not stop at validity; several methods explicitly analyze power, including whether finite data can return a useful nonempty policy. Our distinction is therefore not validity versus an entirely unstudied notion of power. Instead, we use exact power as a planning quantity and ask how reporting structure and testing budget should be chosen before independent certification when the deployment requires multiple simultaneous reporting units.

Group-structured guarantees and reporting design. A second line of work studies groupwise guarantees and the participation costs induced by finer reporting. Fair Risk Control and adaptive risk control extend finite-sample guarantees to structured group risks (Zhang et al., 2024; Blot et al., 2024). HG-CRC assumes a user-specified hierarchy and retreats from leaves to ancestors when a leaf cannot be certified, directly quantifying the participation cost of that retreat (Salem et al., 2026). Our design variable is different. We fix the requested number of reporting units and compare admissible flat partitions when declared groups may be merged. This makes reporting structure itself part of pre-certification planning. The two views are complementary: hierarchical methods adapt the reporting level within a supplied hierarchy, whereas our framework evaluates which admissible partition at a fixed granularity makes the largest amount of traffic certifiable. The HWU study connects these settings by imposing declared semantic boundaries on the availability optimizer and measuring the resulting loss and recovery in certified service.

3 METHODS

3.1 PROBLEM SETTING

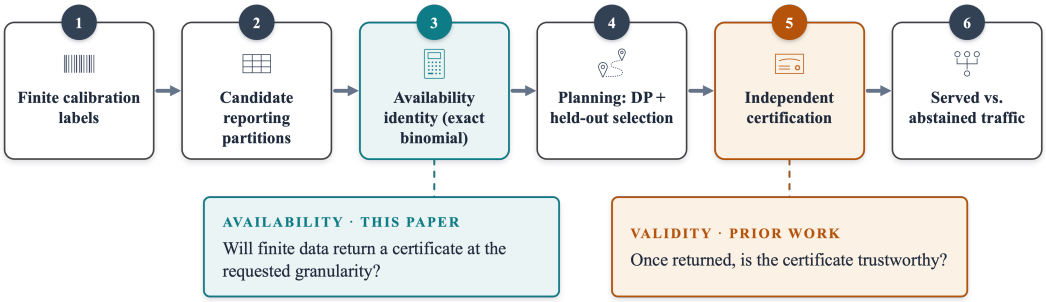


Figure 1: Validity versus availability. Prior work asks whether a returned certificate is trustworthy; we ask one step earlier whether finite calibration data can return a certificate at the requested granularity at all.

Figure 1 summarizes the certification contract. Let $(X, Y) \sim P$ and a fixed predictor produce $\widehat{Y} = f(X)$, with correctness indicator $C = \mathbf{1}\{\widehat{Y} = Y\}$; for other tasks, C is the binary success event in Table 13. A frozen score $s(X)$ ranks apparent reliability, and threshold t serves examples with $A_t = \mathbf{1}\{s(X) \geq t\}$. The selective error is

$$R(t) = \mathbb{P}(C = 0 \mid A_t = 1), \qquad R(t) \leq \alpha = 1 - \tau. \tag{1}$$

This gives a single global guarantee. The groupwise setting additionally introduces a mapping $G(X) \in \{1, \ldots, J\}$, fixed in advance using planning data, that assigns each served example to one of J reporting units. Each unit j has its own threshold t_j and selective error $R_j(t_j) = \mathbb{P}(C = 0 \mid A_{t_j} = 1, G = j)$. Deployment requires all certified units to satisfy their guarantees simultaneously:

$$\mathbb{P}_{\mathcal{D}_{\text{cal}}}(\forall j \text{ certified} : R_j(t_j) \leq \alpha) \geq 1 - \delta. \tag{2}$$

Equation (2) is over a fresh IID certification sample. Before it is drawn, the score function, reporting map, within-unit threshold order, and certification distribution are fixed. These conditions define the validity contract and the limits of certificate transfer.

3.2 FROM A STATISTICAL CERTIFICATE TO CERTIFICATE AVAILABILITY

We begin with one fixed candidate: a reporting unit and score threshold chosen before certification. On a fresh certification sample, suppose it serves $n > 0$ examples and makes E errors. Its unknown true selective error is r. For target precision τ, the largest acceptable error is

$$\alpha = 1 - \tau. \tag{3}$$

For example, $\tau = 0.90$ means we require $r \leq \alpha = 0.10$.

A *certificate* is the statistical conclusion that this true error is at most α. We compute the one-sided Clopper–Pearson upper confidence bound (UCB) $U_{\text{CP}}(E, n; \gamma)$ (Clopper & Pearson, 1934), where γ is the per-candidate miscoverage probability. The UCB is a conservative upper limit on the unknown error rate: over repeated fresh certification samples, it falls below the true error with probability at most γ. The candidate certifies when

$$U_{\text{CP}}(E, n; \gamma) \leq \alpha. \tag{4}$$

Thus certification requires more than an empirical error E/n below α: even the upper confidence bound must lie below the target. Conversely, failure to certify means the sample provides insufficient evidence, not that the candidate is necessarily unsafe.

Define the binomial cumulative distribution function (CDF)

$$F_{\text{Bin}}(k; n, p) = \mathbb{P}(\text{Bin}(n, p) \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}. \tag{5}$$

It is the probability of observing at most k errors in n trials when the error probability is p. For $E < n$, the one-sided Clopper–Pearson UCB is the unique $u \in (0, 1)$ satisfying $F_{\mathrm{Bin}}(E; n, u) = \gamma$; if $E = n$, the bound is 1. Intuitively, this asks how large the true error could still be while remaining compatible with the observed number of errors at confidence level $1 - \gamma$.

With J reporting units, let δ be the total familywise failure probability. Under equal Bonferroni allocation,

$$\gamma = \frac{\delta}{J}. \tag{6}$$

For $J = 10$ and $\delta = 0.05$, each unit uses $\gamma = 0.005$; the union bound then limits the probability of falsely certifying any unsafe unit to at most 0.05. This makes the roles of the two error levels different: α specifies *how safe* a candidate must be, whereas γ specifies *how much statistical evidence* we require before declaring that safety established. We use Bonferroni allocation and fixed-sequence testing as a transparent base contract so that changes in availability can be attributed to planning. Sharper familywise procedures could be substituted and may improve the frontier; we do not compare them here. Section 4 later studies unequal allocations that preserve the same total budget.

The certification rule has an exact power interpretation.

Proposition 1 (Exact availability identity)**.** *For $0 < \alpha, \gamma < 1$,*

$$U_{\mathrm{CP}}(E, n; \gamma) \le \alpha \quad \Longleftrightarrow \quad F_{\mathrm{Bin}}(E; n, \alpha) \le \gamma. \tag{7}$$

Let $e^*(n, \alpha, \gamma)$ be the largest number of observed errors for which certification still succeeds,

$$e^*(n, \alpha, \gamma) = \max\left\{e \in \{0, \dots, n\} : F_{\mathrm{Bin}}(e; n, \alpha) \le \gamma\right\}, \tag{8}$$

with $e^* = -1$ if the set is empty. If the candidate's true selective error is r, then its certificate availability*, the probability that a fresh certification sample yields a certificate, is*

$$\pi(n, r; \alpha, \gamma) = \mathbb{P}(E \le e^* \mid E \sim \mathrm{Bin}(n, r)) = F_{\mathrm{Bin}}(e^*; n, r). \tag{9}$$

When no example is selected, we set $\pi(0, r; \alpha, \gamma) = 0$.

Equation (9) turns certification into a planning quantity. After observing the certification sample, the outcome is binary: the candidate either certifies or it does not. Before that sample is drawn, π is the probability of certification. Availability increases with support n and with the safety margin $\alpha - r$, and decreases as simultaneous reporting makes $\gamma = \delta/J$ smaller. This distinction is central: validity asks whether a returned certificate can be trusted, whereas availability asks how likely finite data are to return one at all. In short, τ sets the precision target, α the maximum error, δ the familywise failure budget, γ the per-unit budget, and π the probability of obtaining a certificate.

3.3 Availability as a planning objective

Equation (9) scores one fixed candidate. To score a reporting design, we combine how much traffic a candidate serves with how likely it is to certify. For segment S and threshold t, let $q_S(t)$ be the served fraction of deployment traffic and $r_S(t)$ its true selective error. With N certification examples, we approximate the selected support by $\lfloor N q_S(t) \rfloor$ and define

$$v(S, t; J) = q_S(t)\, \pi(\lfloor N q_S(t) \rfloor, r_S(t); 1 - \tau, \delta/J)\,. \tag{10}$$

The first factor is traffic served if the candidate certifies; the second is its certification probability. Their product is therefore the expected fraction of deployment traffic contributed by this candidate before the certification sample is observed. A large segment is not useful if it almost never certifies, while an easy-to-certify segment contributes little if it serves negligible traffic.

A reporting design is a partition $\mathcal{P}_J$ with exactly J units. Each original group belongs to exactly one reporting unit, so the partition determines which groups share a certificate and which remain separately reportable. Holding J fixed gives every design the same requested granularity. The optimization therefore asks how to use the same number of certificates more effectively, rather than obtaining coverage by simply collapsing the reporting requirement. Choosing the best threshold in each unit yields

$$U_J(\mathcal{P}_J) = \sum_{S \in \mathcal{P}_J} \max_{t \in \mathcal{T}} v(S, t; J), \tag{11}$$

where $\mathcal{T}$ is the threshold grid. Thus U_J is the projected fraction of total traffic that will be both served and certified when exactly J reporting units are required. Comparing designs at a fixed J keeps reporting granularity constant, so improvements in U_J reflect a better allocation of finite evidence rather than a coarser reporting requirement.

This is a planning objective, not the validity guarantee. It approximates realized coverage because selected support is random whereas Equation (10) uses $Nq_S(t)$, and deployment follows a prespecified threshold sequence rather than testing only the single planning threshold. Exact-support integration and additional checks are reported in Appendix I. We retain Equation (11) because it is simple and effective for comparison; accordingly, "exact dynamic program" means exact optimization of this stated objective over the stated partition family, not exact optimization of final deployed coverage.

Searching all set partitions of K original groups is combinatorial. We therefore freeze an order by planning error, mean score, then identifier, and search only contiguous merges of that order. This restriction makes the optimization tractable while retaining the idea that groups with similar estimated reliability should be eligible to share evidence. Let w_{ab} be the best value for the segment spanning ordered groups a through b. If $D(j, b)$ is the best value for partitioning the first b groups into j segments,

$$D(j,b) = \max_{a \in \{j-1,\dots,b-1\}} \{D(j-1,a) + w_{a,b-1}\}, \qquad D(0,0) = 0. \tag{12}$$

Then $D(J, K)$ optimizes Equation (11) over contiguous J-partitions of this frozen order, with $O(K^2|\mathcal{T}| + JK^2)$ cost. Declared semantic boundaries are enforced by forbidding segments that cross them.

We use this optimizer mainly to expose the population opportunity, not as the paper's deployable endpoint. The *population oracle* substitutes true q and r and measures how much certified traffic is available within the stated partition family if these quantities were known. The *direct plug-in planner* replaces them with planning-sample estimates and tests whether that opportunity can be recovered from finite data. Support-balanced, equal-count, and random-order support-balanced partitions provide simple reference designs. Sweeping J traces the safety, granularity, and traffic frontier. Our primary metric is *traffic availability*, the expected fraction of deployment traffic covered by successfully certified units. *Unit availability* is the probability that one required unit certifies, and *joint availability* is the probability that all required units certify. These latter quantities matter when every named reporting unit must remain available.

3.4 Independent selection among frozen candidate designs

Finite planning data can overfit both segment statistics and the choice among reporting designs. We therefore separate candidate construction from candidate comparison. A construction split produces eight frozen partitions: direct plug-in, support-balanced, equal-count, and five random-order support-balanced candidates. A disjoint selection split evaluates these candidates with Equation (10) and keeps the best one; the selected map is retained and thresholds are refit on all planning labels (Appendix B). The purpose of this procedure is not to privilege the structured dynamic-program candidate. It asks a simpler question: can independent data identify a useful design from a deliberately heterogeneous candidate pool?

Certification then uses a fresh sample and evaluates only the frozen result. Within each unit, a prespecified threshold sequence is tested at level δ/J and stops at the first failure (Angelopoulos et al., 2021); a union bound across units gives Equation (2) (Appendix A). Candidate construction and selection therefore affect availability, while the independent certification stage preserves the same validity guarantee.

4 Results

We use the population optimizer to diagnose how much certified traffic is available in principle, then ask which parts survive finite estimation. The resulting negative findings determine the rest of the section: we test candidate diversity and independent selection, then focus on the two cleanest deployable levers, semantic constraints and familywise-budget allocation. Trained-model, shift, and cross-domain studies define the scope of these conclusions. Unless noted otherwise, $\tau = 0.90$ and $\delta = 0.10$; protocols and implementation details are in Appendices C, D, and K.

4.1 POPULATION OPTIMIZATION REVEALS OPPORTUNITY, NOT A DEPLOYABLE SOLUTION

We first verify that Equation (9) predicts the probability of successful certification. Across 432 untouched binomial cells, the mean absolute difference between predicted and simulated certification frequency is 0.00067. The verified identity then exposes the two forces that limit reporting granularity. Multiplicity is comparatively mild: with true error 0.05 and target availability 80%, increasing J from 1 to 5, 50, and 500 raises the required support from 179 to 287, 450, and 600. The margin to the safety target is much more restrictive. At fixed $J = 50$, increasing the true error from 0.05 to 0.09 raises the same support requirement from 450 to $13{,}407$.

This identity also reveals a substantial population opportunity for reporting design. Using true traffic shares and error rates, the population planner outperforms support balancing in $1{,}366$ of $1{,}845$ cell-by-J configurations, ties in 446, and loses in 33, with a mean coverage gain of 0.1568. The optimization problem is therefore consequential when the required quantities are known. The difficulty is estimating those quantities well enough to choose the partition. Direct plug-in planning gains only 0.0050 on average, has median improvement zero, and reaches -0.3263 at the 5th percentile. At $K = 500$, where per-group estimates are thinnest, its mean effect becomes -0.0659. The gap is not attributable to one mechanism alone. It combines noise in segment statistics, threshold estimates, and the frozen reliability order. Importantly, the population result and the plug-in failure are not contradictory: the former asks whether a better reporting design exists, while the latter asks whether finite planning labels identify it reliably. The central finite-data problem is therefore not whether the population objective contains useful structure, but how much of that structure can be recovered without overfitting the planning sample.

Equation (11) approximates deployed coverage through expected support, one planning threshold, and a restricted partition family. These approximations can matter in individual cases, but exact-support integration does not consistently improve planning in our tests (Appendices I and I.4). We therefore retain the simpler objective and interpret the population planner as a diagnostic frontier. Its role is to show that substantial certified traffic is available in principle, while the gap to finite-data planners measures how difficult it is to identify that opportunity from limited planning labels.

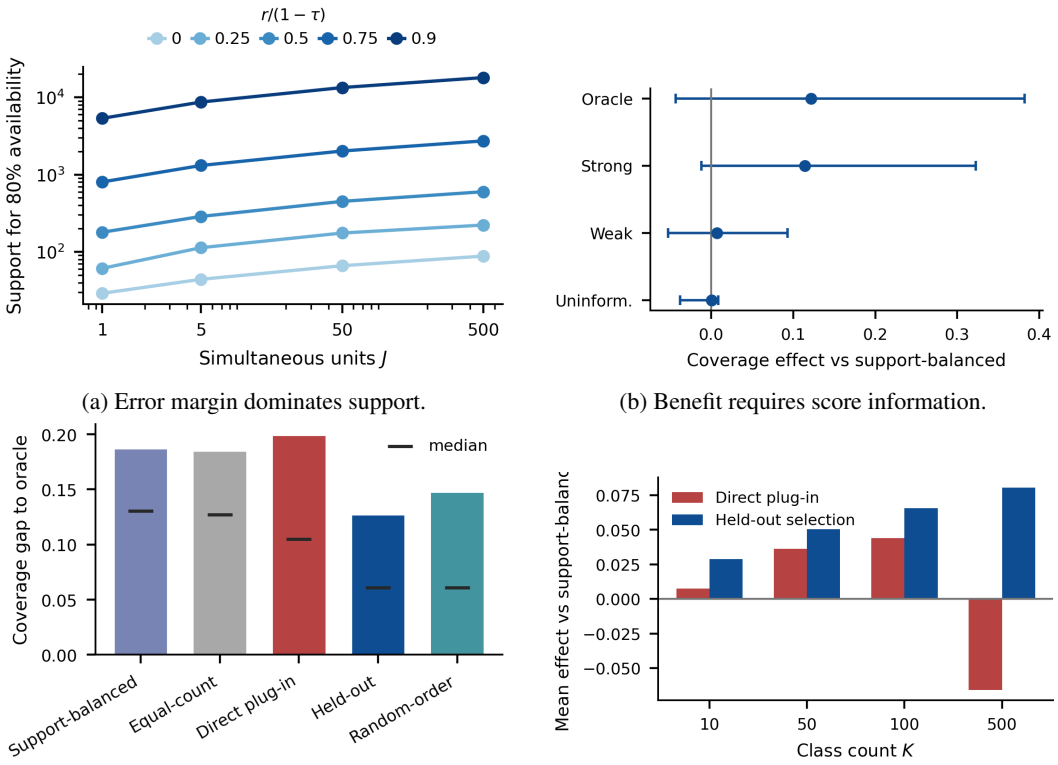
4.2 FINITE DATA FAVOR CANDIDATE DIVERSITY OVER THE RELIABILITY ORDER

Held-out selection does recover some finite-data opportunity, but its interpretation depends on the comparator. Relative to support balancing, the mean gain is 0.0601 across $1{,}340$ non-endpoint configurations, with median 0.00146, and the method still loses in 221 configurations (Table 2). These numbers show that independent candidate comparison can help when many small groups make plug-in estimates unstable, but they do not establish superiority of the structured planner. The stronger question is whether any benefit remains once simple random-order candidates are included.

Random-order balancing is a much stronger comparator. The median gap to the population planner is 0.1305 for support balancing, 0.1051 for direct plug-in, 0.0611 for held-out selection, and 0.0608 for one random-order candidate. Against mean random-order balancing, held-out selection gains only 0.0208 on average with median zero. More directly, the reliability order used by the dynamic program beats the median of twenty random orders only 33% of the time and has lower mean certified coverage, 0.249 versus 0.282 (Appendix I). The estimated reliability order is therefore not a reliable structural prior, and random-order balancing is a central comparator rather than a secondary ablation.

The matched-pool mechanism study explains why the stronger comparator is so competitive. Holding the candidate set fixed, selecting on independent data rather than reusing construction data contributes 0.0480 mean coverage across 64 synthetic conditions. In contrast, adding the structured candidates to five random groupings contributes only 0.0027 with median zero (Table 1). The principal recoverable benefit is therefore not the estimated reliability order. It is the combination of candidate diversity and an independent comparison among frozen designs (Appendix I). This changes the interpretation of the planner: the structured dynamic program defines a useful population design problem, but under finite data the stronger practical lever is often how candidate designs are generated and compared.

Score quality limits the recoverable frontier. The held-out gain over support balancing is 0.1217 with an oracle score and 0.1144 with a strong learned score, but falls to 0.0072 with a weak score and 0.0004 with an uninformative one (Figure 2b). The near match between oracle and strong learned scores shows that perfect ranking is unnecessary; what matters is enough ordering information to



(a) Error margin dominates support.

(b) Benefit requires score information.

(c) Coverage gap to the population planner.

(d) Held-out selection at large K.

Figure 2: Synthetic evidence for the availability identity and the planners.

Table 1: Matched-pool mechanism study. Coverage differences are means over 64 fixed synthetic conditions, each averaging 20 independent plans and 100 nested certification draws. Starts remain frozen except in the last row. Positive values favor the first method.

Comparison	Mean	Median	Win/tie/loss
Held-out full minus construction selection	0.0480	0.0032	53/0/11
Held-out full minus uniform full pool	0.0318	0.0054	44/0/20
Held-out full minus empirical traffic	0.0161	0.0002	38/1/25
Held-out full minus held-out random five	0.0027	0.0000	29/4/31
Held-out random five minus random one	0.0136	0.0012	44/0/20
Refitted full minus refitted random five	0.0030	−0.00003	24/3/37

separate safer from riskier traffic. Once that information is lost, repartitioning cannot manufacture a useful safety margin. Planning can redistribute finite evidence, but it cannot compensate for a score that fails to rank reliability.

4.3 Familywise-budget allocation is a direct deployable lever

A second design variable is the familywise testing budget. Equal Bonferroni allocation uses δ/J for every unit, but validity only requires frozen levels $\gamma_1, \ldots, \gamma_J$ with $\sum_j \gamma_j \leq \delta$. The budget can therefore be shifted toward units where a larger level most improves certification probability without changing the union-bound guarantee.

Across 7,200 population configurations, optimal allocation gains 0.038 mean coverage (median 0.028). With one noisy planning sample, the deployable allocation still gains 0.011 over equal allocation across 2,880 truths and is positive in 68% of cases; posterior-predictive availability raises these values to 0.014 and 74%. An unsafe-unit audit remains within the nominal familywise level

(Appendix J). These gains exceed the 0.0027 contribution from adding structured partition candidates, making budget allocation the stronger finite-data lever once candidate diversity is available.

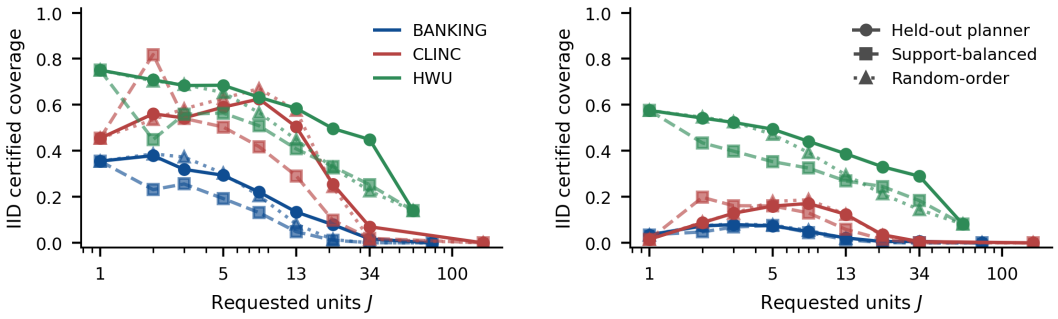
4.4 Constrained planning can expand certified service without changing semantics

Availability is not useful if a partition destroys required reporting semantics. On HWU we therefore forbid merges across named scenarios. This necessarily reduces the unconstrained frontier because some statistically convenient merges are no longer admissible. The reduction should be interpreted as the price of semantic interpretability, not as a failure of the constrained planner. For named, fairness-sensitive, or legally defined groups, this constrained space is the relevant deployment problem.

Within this space, direct constrained optimization outperforms the full held-out selector in all twenty model-by-granularity comparisons at $J = 21$ and 34 (Table 7). At $J = 34$, DeBERTa coverage rises from 0.1742 under support balancing and 0.1839 under random selection to 0.3523; DistilRoBERTa rises from 0.0861 and 0.0971 to 0.2601 (Table 11). The planner also serves more named scenarios on average. This is our clearest positive partitioning result because it improves certified service without relaxing the declared semantics. The analysis is a post hoc replay on inspected HWU traces rather than a preregistered prospective comparison, so we treat the magnitude as evidence that constrained availability planning is promising rather than as a definitive benchmark claim.

4.5 Independent-model experiments confirm the frontier, not a universal planner

We next test whether the availability frontier and the synthetic mechanism findings survive independent model training. Figure 3 compares held-out selection with both support balancing and mean random-order support balancing on three intent datasets and two architectures, with thirty resplits nested within each of five independently trained models. Against the simple support-balanced reference, held-out selection is positive in fifty-nine of sixty model-by-granularity effects, with gains ranging from 0.0045 to 0.2128 (Table 3). We treat this as evidence that support balancing leaves certified traffic unused, not as the main algorithmic comparison, because the stronger random-order baseline is much more competitive.



(a) DeBERTa: matched-unit frontiers.

(b) DistilRoBERTa: matched-unit frontiers.

Figure 3: Intent-classification certified-coverage frontiers, by architecture.

The comparison with random-order balancing is deliberately stricter and correspondingly mixed. At $J = 13$, held-out selection exceeds mean random-order coverage for DeBERTa on BANKING and HWU but trails it on CLINC. DistilRoBERTa differences are smaller and cross zero, while direct plug-in planning remains competitive in several BANKING and HWU settings. This pattern matches the synthetic ablation: independent selection is useful, but the structured candidate family does not consistently dominate diverse random groupings.

Selection frequencies are similarly mixed. Across 16,200 real selection decisions, support balancing is chosen 28.2% of the time, direct plug-in 17.3%, equal-count 5.4%, and random-order partitions 49.0%. No candidate family dominates, which is consistent with the synthetic mechanism study: the value of the held-out procedure comes from comparing heterogeneous frozen candidates rather than

from always selecting one structured family. These frequencies should not be read as an algorithmic ranking because the preferred candidate changes with the model, dataset, granularity, and particular planning split.

4.6 Certificate transfer depends on the form of shift

The validity guarantee applies to the certification distribution unless shift is explicitly corrected. CLINC illustrates why the form of shift matters. Under label-prior shift, class prevalence changes while within-class behavior is preserved, so reweighting can in principle repair the target distribution. Under conditional-outcome shift, prediction behavior changes given the inputs, so feature-density ratios may contain no corrective signal. At 30% out-of-scope prevalence, a stale policy's precision falls to 0.753; exact reweighting restores 0.935 but reduces coverage to 0.455. Conditional-outcome shift is harder because feature-density ratios may contain no signal: at 5% severity, fresh labels restore precision to 0.913 only by roughly halving coverage, and at 10% recalibration abstains entirely. Monitoring also detects only 11.3% of a 2.5% conditional shift within $2{,}000$ probes (Appendix G).

4.7 Cross-domain studies support the phenomenon, not planner dominance

We also evaluate four frozen predictors under the same certification contract: Qwen2.5-14B tool calling, Civil Comments moderation, ResNet-18 lesion classification on DermaMNIST, and matrix-factorization recommendation. Each study uses one fixed base predictor, so repeated splits measure allocation variability rather than retraining variability. These experiments therefore test whether the certified service frontier recurs across qualitatively different applications, not whether one planner universally dominates. The frontier appears in every domain, with a ceiling set jointly by predictor quality and finite support (Table 4). Higher safety targets reduce availability throughout. Granularity is not always monotone because separate units can initially exploit heterogeneous thresholds before support scarcity dominates. In recommendation, for example, coverage rises from 0.65 globally to 0.79 at $J = 2$ for $\tau = 0.80$. The tool-call gate remains highly available at a looser target but nearly closes at 0.90, whereas the stronger moderation predictor retains substantial coverage at the same target.

Planner rankings remain heterogeneous: random-only selection is strongest on moderation and function calling in the shared-pool comparison, held-out selection on lesion classification, and direct planning on recommendation (Appendix H). This heterogeneity is consistent with the mechanism study: the frontier is general, but the best way to recover it depends on the information available to the planner. The cross-domain conclusion is therefore the recurrence of the certified service frontier, while the best recovery strategy depends on score quality, admissible structure, candidate diversity, and finite estimation. This distinction matters for the paper's scope: the general claim concerns the deployment constraint itself, whereas the planner comparisons identify which recovery mechanisms are useful in particular regimes.

5 Conclusion and Limitations

This paper treats certificate *availability* as a deployment quantity distinct from validity. Exact binomial inversion exposes a safety, granularity, and traffic frontier. While population optimization shows substantial certified traffic is possible, finite-data experiments reveal that plug-in recovery is weak, estimated reliability orders often underperform random orderings, and structured candidates add little given a diverse pool. The surviving constructive levers are more targeted: independent held-out comparison for heterogeneous designs, direct constrained planning for fixed semantic boundaries, and unequal familywise budget allocation. Together, these establish a framework for measuring certified service and optimizing pre-certification design variables.

Current scope is bounded by key limitations: no single planner dominates, weak reliability scores leave little availability to recover, and the dynamic program serves only as a diagnostic reference. Furthermore, semantic HWU replays require prospective confirmation, and cross-domain studies show frontier recurrence rather than algorithmic generalization. We also omit comparisons with hierarchical fallbacks or stronger familywise procedures, which remain complementary. Finally, guarantees require valid correction or fresh labels under distribution shift, confirming that honest finite-sample certification is both a validity and resource-allocation problem.

Ethics Statement

This work uses only public benchmarks, with no new data collection or human subjects, and measures benchmark success rather than clinical efficacy or deployment safety. Its two principal ethical stakes are addressed in the body of the paper. First, merging groups by statistical reliability can fuse categories that a deployment must keep distinct, so named, fairness-sensitive, or legally defined groups should be preserved or governed by a declared hierarchy (Section 4.4). Second, a certificate can appear valid after the deployment distribution has shifted, which our controlled-shift study is designed to expose rather than hide.

Reproducibility Statement

The availability identity is proved in Appendix A and the full planner in Appendix B, while Section 4 separates the frozen studies from the exploratory ones. Under a pinned environment and fixed seeds the pipeline is deterministic, and the supplementary material contains the code, aggregate and seed-level results, and an automated check that recomputes every reported number from them (Appendix L).

AI Statement

A large language model assisted the authors in polishing the language of the manuscript. The authors verified that all claims, proofs, mathematical formulations, and reported values are valid, checked them against the implementation and results, and take full responsibility for the manuscript.

References

Anastasios Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International conference on learning representations*, volume 2024, pp. 55198–55218, 2024.

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2021. doi: 10.1214/24-aoas1998.

Tian Bai and Ying Jin. Conformal selective prediction with general risk control. *arXiv preprint arXiv:2603.24704*, 2026. URL https://arxiv.org/abs/2603.24704.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021. doi: 10.1145/3478535.

Vincent Blot, Anastasios N Angelopoulos, Michael I Jordan, and Nicolas JB Brunel. Automatically adaptive conformal risk control. *arXiv preprint arXiv:2406.17819*, 2024.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of the 2019 World Wide Web Conference*, pp. 491–500, 2019. doi: 10.1145/3308560.3317593.

Lawrence D Brown, T Tony Cai, and Anirban DasGupta. Interval estimation for a binomial proportion. *Statistical science*, pp. 101–117, 2001. doi: 10.1214/SS/1009213286.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on NLP for Conversational AI*, pp. 38–45, 2020. doi: 10.18653/v1/2020.nlp4convai-1.5.

Chao Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

Charles J Clopper and Egon S Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/BIOMET/26.4.404.

Ran El-Yaniv et al. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5), 2010. doi: 10.5555/1756006.1859904.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):1–19, 2016. doi: 10.1145/2827872.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021. URL https://arxiv.org/abs/2111.09543.

K Krishnamoorthy and Jie Peng. Some properties of the exact and score methods for binomial proportion and sample size calculation. *Communications in Statistics—Simulation and Computation®*, 36(6):1171–1186, 2007. doi: 10.1080/03610910701569218.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of EMNLP-IJCNLP*, pp. 1311–1316, 2019. doi: 10.18653/v1/D19-1131.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. Benchmarking natural language understanding services for building conversational agents. In *Proceedings of the Tenth International Workshop on Spoken Dialogue Systems Technology*, pp. 165–183, 2019. doi: 10.1007/978-981-15-9323-9_15.

Shishir G Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E Gonzalez. The berkeley function calling leaderboard (bfcl): From tool use to agentic evaluation of large language models. In *Forty-second International Conference on Machine Learning*, 2025.

Murilo Salem, Luísa Böhm, Daniel Pontes, and Anderson Ferrugem. Hierarchical group-conditional conformal risk control for selective prediction in language models. *arXiv preprint arXiv:2607.24562*, 2026.

Måns Thulin. The cost of using exact confidence intervals for a binomial proportion. *Electronic Journal of Statistics*, 8(1):817–840, 2014. doi: 10.1214/14-EJS909.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Zhiyuan Wang, Aniri, Tianlong Chen, Yue Zhang, Heng Tao Shen, Xiaoshuang Shi, and Kaidi Xu. LEC: Linear expectation constraints for false-discovery control in selective prediction and routing systems. *arXiv preprint arXiv:2512.01556*, 2025. URL https://arxiv.org/abs/2512.01556.

Zhiyuan Wang, Jinhao Duan, Qing Wang, Xi Zhu, Tianlong Chen, Xiaoshuang Shi, and Kaidi Xu. COIN: Uncertainty-guarding selective question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33764–33772, 2026. doi: 10.1609/aaai.v40i40.40667.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023. doi: 10.1038/s41597-022-01721-8.

Lujing Zhang, Aaron Roth, and Linjun Zhang. Fair risk control: A generalized framework for calibrating multi-group fairness risks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 59783–59805, 2024. URL https://proceedings.mlr.press/v235/zhang24be.html.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. A robustly optimized bert pre-training approach with post-training. In *Proceedings of the 20th chinese national conference on computational linguistics*, pp. 1218–1227, 2021. doi: 10.1007/978-3-030-84186-7_31.

APPENDIX OVERVIEW

The appendices are organized by purpose. We begin with methodology and guarantees: the proofs behind the availability identity, the dynamic program, and the validity of the certified policy, followed by the full planner specification, the study protocols and reporting conventions, and the score model and training recipe. We then present the experimental evidence, one section per scientific question: extended synthetic results, independent-model comparisons, controlled distribution shift, cross-domain frontiers, exact-coverage and additional-planner analyses, and a study of familywise-budget allocation. The final two sections give the application populations and the reproducibility and artifact-integrity statement.

A PROOFS

Throughout, $F_{\mathrm{Bin}}(k; n, p) = \mathbb{P}(\mathrm{Bin}(n, p) \le k)$ is the binomial lower-tail probability defined in Equation (5), and every candidate, map, score, and threshold is frozen before the certification sample is drawn.

A.1 PROOF OF THE EXACT AVAILABILITY IDENTITY

We establish the two displayed claims in turn.

Step 1: the inversion identity (7). Fix $0 \le E < n$ and regard the lower tail as a function of the error rate:

$$g(p) = F_{\mathrm{Bin}}(E; n, p) = \sum_{i=0}^{E} \binom{n}{i} p^i (1-p)^{n-i}.$$

Differentiating yields a telescoping cancellation, leaving

$$g'(p) = -\,n\binom{n-1}{E} p^E (1-p)^{\,n-1-E} < 0 \qquad (0 < p < 1),$$

so g is continuous and strictly decreasing on $[0, 1]$, with $g(0) = 1$ and $g(1) = 0$. For each $\gamma \in (0, 1)$ there is therefore a unique $u \in (0, 1)$ with $g(u) = \gamma$, and by definition $u = U_{\mathrm{CP}}(E, n; \gamma)$. Because g is strictly decreasing, the inequality $u \le \alpha$ reverses under g:

$$U_{\mathrm{CP}}(E, n; \gamma) \le \alpha \iff g(u) \ge g(\alpha) \iff F_{\mathrm{Bin}}(E; n, \alpha) \le \gamma,$$

which is Equation (7). When $E = n$, we have $g \equiv 1$, so $U_{\mathrm{CP}} = 1 > \alpha$ and $F_{\mathrm{Bin}}(n; n, \alpha) = 1 > \gamma$; both sides are false and the equivalence holds vacuously.

Step 2: the availability formula (9). For fixed n, α, γ, the map $k \mapsto F_{\mathrm{Bin}}(k; n, \alpha)$ is nondecreasing, so the acceptance set $\{k : F_{\mathrm{Bin}}(k; n, \alpha) \le \gamma\}$ is a (possibly empty) prefix $\{0, 1, \ldots, e^*\}$ where

$$e^* = e^*(n, \alpha, \gamma) = \max\{k \in \{0, \ldots, n\} : F_{\mathrm{Bin}}(k; n, \alpha) \le \gamma\},$$

and $e^* = -1$ precisely when even $k = 0$ fails, i.e. when $(1-\alpha)^n > \gamma$. By Step 1 the candidate certifies if and only if $F_{\mathrm{Bin}}(E; n, \alpha) \le \gamma$, equivalently $E \le e^*$. Under the true selective error r the error count satisfies $E \sim \mathrm{Bin}(n, r)$, so the probability of certifying is

$$\pi(n, r; \alpha, \gamma) = \mathbb{P}(E \le e^*) = F_{\mathrm{Bin}}(e^*; n, r),$$

with the convention $F_{\mathrm{Bin}}(-1; n, r) = 0$. This is Equation (9) and its boundary case $\pi = 0$ when no error count certifies. □

A.2 OPTIMALITY OF THE DYNAMIC PROGRAM

Fix the granularity J and the reliability order, and index its groups by positions $0, \ldots, K-1$. For $0 \le a \le c \le K-1$, let $w_{ac} = \max_{t \in \mathcal{T}} v(S_{ac}, t; J)$ denote the best score of the run spanning positions a through c, and let $D(j, b)$ be the maximum of $\sum w$ over all partitions of the first b positions $\{0, \ldots, b-1\}$ into exactly j contiguous runs, with $D(0, 0) = 0$ and $D(j, b) = -\infty$ when no such partition exists (for instance when $j > b$). The objective (11) is a sum of per-run terms, and

each run maximizes its own threshold independently, so the score of a partition equals the sum of its runs' w values; the problem therefore has optimal substructure.

Any contiguous partition of $\{0, \dots, b-1\}$ into $j \geq 1$ runs has a last run $\{a, \dots, b-1\}$ for a unique split point a. Requiring every run to be nonempty forces $a \in \{j-1, \dots, b-1\}$. Removing the last run leaves a contiguous partition of $\{0, \dots, a-1\}$ into $j-1$ runs, whose score is at most $D(j-1, a)$ and equals it for an optimal remainder. Adding the last run's score $w_{a,b-1}$ and maximizing over a yields

$$D(j, b) = \max_{a \in \{j-1, \dots, b-1\}} \big\{ D(j-1, a) + w_{a,b-1} \big\},$$

which is Equation (12). Induction on j and then b shows that $D(J, K)$ is the optimal score over all contiguous J-partitions of the order, and the split points attaining each maximum reconstruct the optimal reporting map. Setting $w_{ac} = -\infty$ whenever the run S_{ac} crosses a forbidden semantic boundary removes exactly the infeasible partitions and leaves the argument otherwise intact. Non-contiguous partitions of the order lie outside the optimization domain by construction. □

A.3 Validity of the certified policy

Let $\mathcal{D}_{\text{plan}}$ denote the planning labels and $\mathcal{D}_{\text{cal}}$ the fresh, independent certification sample of Equation (2). Condition on $\mathcal{D}_{\text{plan}}$: the score s, the reporting map G, the unit count J, and each unit's ordered list of candidate thresholds are then deterministic functions of $\mathcal{D}_{\text{plan}}$ and do not depend on $\mathcal{D}_{\text{cal}}$. Fix a unit j and test its thresholds in the pre-specified order, each by the exact test of Proposition 1 at level δ/J, stopping at the first that fails and certifying the last that passed. Declaring unit j safe with a threshold whose true selective error exceeds α requires rejecting the first true null in this ordered family, and the fixed-sequence testing guarantee (Angelopoulos et al., 2021) bounds that event's probability by the per-test level:

$$\mathbb{P}(j \text{ certified and } R_j(t_j) > \alpha \mid \mathcal{D}_{\text{plan}}) \leq \frac{\delta}{J}.$$

A union bound over the J units then gives

$$\mathbb{P}(\exists\, j \text{ certified} : R_j(t_j) > \alpha \mid \mathcal{D}_{\text{plan}}) \leq \sum_{j=1}^{J} \frac{\delta}{J} = \delta,$$

and taking expectation over $\mathcal{D}_{\text{plan}}$ removes the conditioning while preserving the bound, which is exactly Equation (2). The map, score, and threshold order enter only through the conditioning, so candidate quality and selection bias affect which policy is certified but never the level at which it is certified. □

B Planner Specification

B.1 Inputs

The planner receives a frozen predictor, declared groups, target τ, familywise failure probability δ, requested granularity J, certification size N, and grid $(0, 0.02, \dots, 1)$. Only planning labels are available at this stage.

B.2 Reliability order

For each predicted class, compute its construction error rate. Sort classes by ascending error, descending mean score, and ascending integer class index. Empty classes receive error one and mean score zero. The equal-count partition assigns rank k to $\min(J-1, \lfloor kJ/K \rfloor)$. The support-balanced partition assigns each ordered class to $\min(J-1, \lfloor Jc/S \rfloor)$, where c is its preceding cumulative support and S is total construction support. If fewer than J units are nonempty, use the equal-count partition. Random-order support-balanced partitions apply the same rule after seeded permutations.

B.3 SEGMENT TABLES AND TIE RULES

For every contiguous segment and grid threshold, compute the selected construction count m, its error count e, and the estimated traffic share $\widehat{q} = m/n_{\text{plan}}$, where n_{plan} is the size of the construction set. The projected certification support is $\lfloor N\widehat{q} \rfloor$, and the estimated error rate is the Jeffreys mean $\widehat{r} = (e + 1/2)/(m + 1)$. Score the pair with Equation (10). Within a segment, maximize this score, then projected support, and finally prefer the lower threshold. The dynamic program uses the first maximizing split. Semantic constraints render cross-boundary segments infeasible.

B.4 HELD-OUT CANDIDATE-SELECTION ALGORITHM

Deterministically permute the planning role and divide it into equal construction and selection subsets. Fit the correctness score on construction labels. Build the direct plug-in, support-balanced, equal-count, and five random-order support-balanced partitions. Freeze their maps and construction thresholds.

On the selection subset, recompute q and the Jeffreys estimate of r for each candidate at its fixed construction thresholds. Select the candidate with the largest summed projected coverage from Equation (10), without a new threshold search. Ties within 10^{-12} prefer support-balanced, equal-count, direct plug-in, then random-order candidates by index. Retain the selected map, refit the score on all planning labels, and update its starting thresholds. On certification data, test grid points downward from each planned start and stop at the first failure. A reporting unit that never passes abstains. Coverage and feasibility summaries retain all abstain-all outcomes.

C STUDY PROTOCOLS AND REPORTING

This appendix collects the full protocol behind each study summarized in Section 4: sample sizes, resplit schemes, generator factors, and the reporting conventions dictated by each study's independent unit.

C.1 SYNTHETIC

An analytic check draws 432 hash-selected binomial cells (10,000 simulations each) to validate the identity itself. The planner studies run on a generator that crosses six label budgets, four class counts $K \in \{10, 50, 100, 500\}$, balanced or Zipf traffic, error heterogeneity, three targets, three failure levels, and four score regimes ranging from oracle to uninformative. Two disjoint confirmation pools carry the claims: 222 untouched cells for the direct planner and 215 for the held-out planner, each with 25 planning draws and 20 nested certification draws. The held-out aggregate excludes the coincident endpoints $J = 1, K$. Because the two pools differ, we do not divide their mean gains to claim a recovered fraction. A further shared-candidate study uses 64 conditions, 20 planning draws per setting, and 100 nested certification draws to separate the benefit of held-out selection from that of candidate diversity (Appendix I). Real-data planners search the threshold grid $\mathcal{T} = \{0, 0.02, \ldots, 1\}$, while the synthetic score generator uses $\{0, 0.05, \ldots, 0.95\}$.

C.2 INTENT CLASSIFICATION

On BANKING77 (Casanueva et al., 2020), CLINC150 (Larson et al., 2019), and HWU64 (Liu et al., 2019), we train five DeBERTa-v3-base (He et al., 2021) and five DistilRoBERTa-base (Zhuang et al., 2021) models each, analyzing the two architectures separately. Within each model, 30 deterministic resplits divide the evaluation mixture into 10/10/60/20% for construction, selection, certification, and IID deployment. The official test trace is reported as a shifted mixture and held out of the IID analysis. The model seed, an independent training run from a fresh random initialization, serves as the outer independent unit, with resplits treated as nested measurements. HWU's 18 scenarios supply the semantic hierarchy, and controlled CLINC shifts (Appendix G) separate label-prior from conditional-outcome drift.

C.3 Cross-domain safety gates

Four further domains each gate a single frozen predictor: a Qwen2.5-14B agent on BFCL single-turn function calling (four tool categories, official AST checker), a logistic-regression moderator on Civil Comments, a fine-tuned ResNet-18 on DermaMNIST (seven lesion classes), and a matrix-factorization recommender on MovieLens (item liking, grouped by genre). Each domain uses one frozen base predictor, with auxiliary models contributing score features. We average ten resplits at each of $\tau \in \{0.70, 0.80, 0.90\}$, capping evaluation pools at 8,000 rows. The separate held-out comparisons use six resplits at domain-specific targets. These measure allocation variability within fixed prediction records, not independent training replication. Appendix K gives the populations and the additional ten-resplit shared-candidate comparison.

C.4 Reporting

Because the independent unit differs across studies, so does the reporting format. Synthetic analyses report cell-by-J paired effects, win/tie/loss counts, quantiles, and absolute coverage gaps to the population planner. Planning draws are the independent simulation units, with certification draws nested within plans. These summaries describe the fixed design. Intent analyses instead expose every model-level effect individually, along with sign counts and leave-one-model-out ranges. With five models per cell, the smallest two-sided exact sign p-value is 0.0625, so we report the direction and magnitude of each effect without making threshold-based significance claims.

D Score Model and Training

The correctness score is deliberately simple and deterministic: a logistic regression (at most 2,000 iterations, random seed zero) fitted on confidence, top-two margin, entropy, Monte Carlo disagreement, and auxiliary-model agreement and confidence, augmented with planning class accuracy and log support as group-level features. When construction labels are constant, the model falls back to a constant probability, with raw confidence as a secondary score. In each cross-domain gate the same score is refit from whatever per-example confidence signals the predictor exposes.

The intent classifiers set the training recipe. Both architectures train for four epochs at a maximum sequence length of 64 with AdamW (2×10^{-5}, weight decay 0.01, 6% warmup), using batch size 64 for DeBERTa and 128 for DistilRoBERTa. HWU's 18 source scenarios supply the semantic hierarchy imposed later as a constraint. DistilRoBERTa in-scope test accuracies across five seeds were 0.6481 to 0.6932 on BANKING, 0.8533 to 0.8656 on CLINC, and 0.7471 to 0.7581 on HWU. These models broaden the evaluated performance and score-quality range but are not matched-accuracy replacements for DeBERTa.

E Extended Synthetic Results

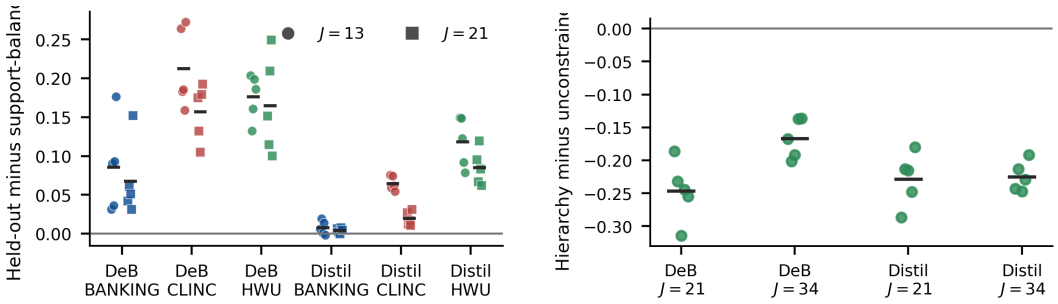
The direct and held-out confirmations drew on disjoint pools of cells, partitioned by a hash of each cell's identifier under their respective frozen designs. The direct study contained 1,845 cell-by-J configurations. Relative to support balancing, direct plug-in planning won in 581 configurations, tied in 832, and lost in 432. Its mean effect was 0.00496, with 5th-to-95th percentiles spanning -0.3263 to 0.2855. When broken down by condition, the mean effects were -0.0659 at $K = 500$, -0.0229 with weak scores, and -0.0085 under Zipf traffic.

The held-out study contained 1,340 nonendpoint configurations. Table 2 collects their primary aggregate comparisons. A separate validity audit checks the implementation against the proved contract: across 1,770 cell-by-J rows it detected no excess after Bonferroni correction, with the smallest excess-test p-value being 0.5604.

The held-out planner chose support balancing 15,501 times, direct plug-in planning 7,354 times, equal-count partitions 1,708 times, and random-order partitions 19,687 times. Population-oracle, strong, weak, and uninformative score strata had mean effects over support balancing of 0.12166, 0.11438, 0.00717, and 0.00041. Their medians were 0.08816, 0.08513, zero, and zero.

Table 2: Held-out planner coverage effects at matched J. Each effect is a cell-level mean over independent planning draws; certification draws are nested within plans.

Baseline	Mean	Median	Wins	Ties	Losses
Support-balanced	0.06007	0.001465	744	375	221
Equal-count	0.05811	0.000857	723	375	242
Direct plug-in	0.07213	0.000009	673	381	286
Mean random-order	0.02079	0.000000	548	359	433



(a) All five model effects.

(b) Semantic constraints cost coverage.

Figure 4: Per-model intent effects and the coverage cost of the HWU semantic constraint.

F Independent-Model Results

Table 3 reports every primary model-level effect. Each value averages 30 nested resplits within one trained model. These five independent observations determine direction, range, and leave-one-model-out summaries.

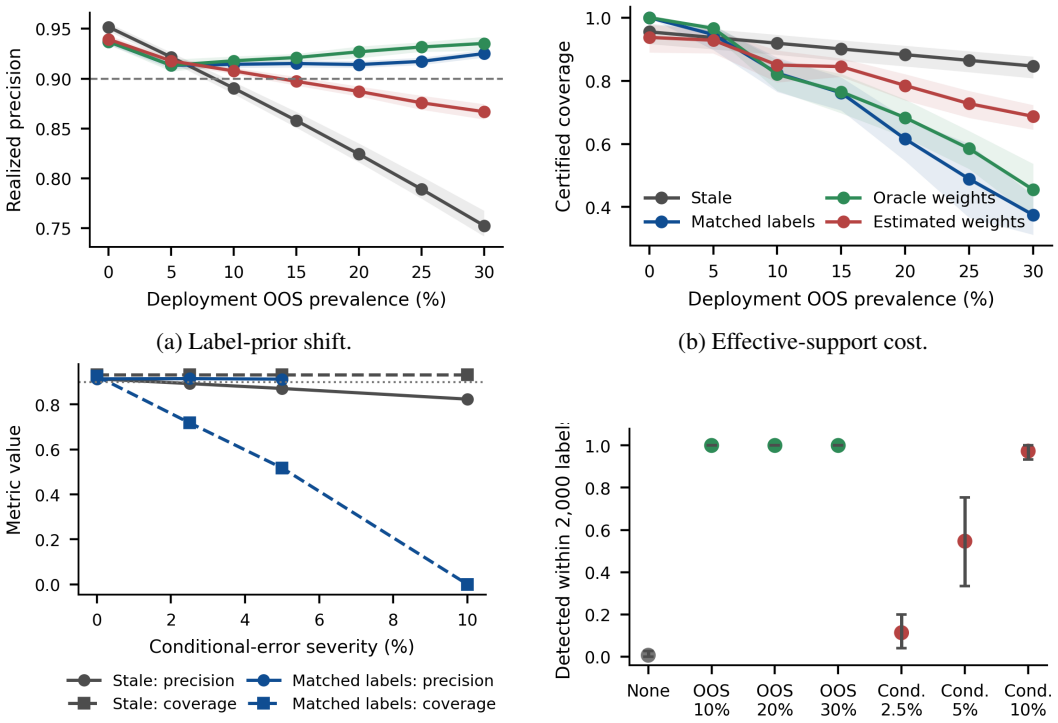
Table 3: Held-out planner minus support-balanced IID coverage.

Architecture	Dataset	J	Five model-level effects
DeBERTa	BANKING	13	0.0316, 0.0906, 0.0362, 0.0933, 0.1767
		21	0.0426, 0.0613, 0.0513, 0.0316, 0.1525
DeBERTa	CLINC	13	0.2639, 0.1833, 0.1857, 0.1590, 0.2723
		21	0.1753, 0.1325, 0.1055, 0.1795, 0.1926
DeBERTa	HWU	13	0.2034, 0.1325, 0.1607, 0.1987, 0.1863
		21	0.1516, 0.1148, 0.2098, 0.2494, 0.1006
DistilRoBERTa	BANKING	13	0.0060, 0.0195, 0.0005, 0.0143, −0.0018
		21	0.0067, 0.0028, 0.0003, 0.0077, 0.0051
DistilRoBERTa	CLINC	13	0.0755, 0.0595, 0.0743, 0.0614, 0.0543
		21	0.0271, 0.0130, 0.0172, 0.0108, 0.0313
DistilRoBERTa	HWU	13	0.1492, 0.1485, 0.1226, 0.0921, 0.0788
		21	0.0954, 0.0671, 0.1194, 0.0833, 0.0624

The two architectures span a wide range of accuracy and score quality (Appendix D), so they broaden the evaluated regime rather than serving as matched-accuracy replacements for each other. The official test trace is carried alongside every frontier but labeled as a shifted mixture throughout. No shifted value contributes to the primary table.

G Controlled-Shift Details

Each of five CLINC models supplied 30 nested repetitions, drawn so that the calibration and deployment pools remained disjoint within out-of-scope status. Eleven frozen scenarios span the two kinds of shift. Some vary out-of-scope prevalence, producing a pure label-prior shift; others convert a



(a) Label-prior shift.

(b) Effective-support cost.

(c) Outcome drift requires new labels.

(d) Monitoring is delayed.

Figure 5: Controlled CLINC shift: label-prior repair and its cost, conditional drift, and delayed monitoring.

nested fraction of high-confidence correct outcomes into errors, producing a conditional-outcome shift. Against the first kind we compare three repair strategies: exact rejection sampling from the known label-prior ratios, a domain-classifier arm that estimates those ratios instead, and matched recalibration on new deployment labels. A Bonferroni monitor watches two signals at 40 checkpoints and labels 5% of traffic. Throughout, precision is conditioned on the serving repetitions, while coverage and feasibility retain the abstain-all ones.

The headline results quantify each arm. Under 30% out-of-scope prevalence, oracle weighting reached 0.935 precision at 0.455 coverage, while estimated weighting traded precision down to 0.867 for higher coverage of 0.687. Under 5% conditional severity, matched new labels reached 0.913 precision at 0.517 coverage. By 10% severity, no repetition certified at all.

H Extended Cross-Domain Results

Table 5 reports the full safety-availability frontier for the four cross-domain gates. For each safety target it gives certified coverage at both the global unit and the finest unit, so a single row shows both edges of the frontier: the cost of the target itself, and the additional cost of splitting the gate into fine-grained units. The endpoint pairs illustrate the price of higher targets and finer reporting, although intermediate coverage need not decrease with every additional unit. Moderation loses almost nothing until the target reaches 0.90. The tool-call gate has 0.79 global coverage at 0.80 and only 0.17 at 0.90. Recommendation, the least accurate model, pays the steepest price, retaining only 0.04 of finely reported traffic at a 0.90 target. The final column reports a separate six-resplit held-out comparison at one intermediate granularity and a domain-specific target. The gain is substantial where the gate splits into several still-supportable units (moderation, tool-calling) and near zero at the other two studied operating points.

Two boundaries of these studies are worth restating here. Each domain uses a single frozen model, so the rows report how the trade-off behaves for that particular model, not an effect size that would

Table 4: Certified coverage at every reporting granularity J for the four cross-domain gates, at three safety targets, over ten calibration resplits of one frozen predictor per domain (J=1 is the global gate).

Domain	J	τ=0.70	τ=0.80	τ=0.90
Tool-call gating	1	1.00	0.79	0.17
(K=4, acc. 0.83)	2	1.00	0.67	0.17
	3	0.93	0.49	0.14
	4	0.84	0.41	0.09
Content moderation	1	1.00	1.00	0.97
(K=5, acc. 0.89)	2	1.00	1.00	0.95
	3	1.00	0.98	0.91
	5	1.00	0.91	0.85
Lesion classification	1	1.00	1.00	0.52
(K=7, acc. 0.86)	2	0.96	0.99	0.49
	3	0.95	0.98	0.55
	5	0.90	0.87	0.58
	7	0.81	0.72	0.46
Recommendation	1	1.00	0.65	0.32
(K=18, acc. 0.76)	2	1.00	0.79	0.27
	3	0.99	0.79	0.23
	5	0.98	0.76	0.22
	8	0.96	0.59	0.27
	13	0.91	0.45	0.17
	18	0.86	0.47	0.04

Table 5: Cross-domain frontiers. Each pair is certified coverage at the global and finest (J=K) units over ten resplits of one frozen model; the last column is a separate six-resplit held-out gain, and Acc. is the per-task success fraction rather than a common accuracy.

Domain	Model	K	Acc.	τ=0.70	τ=0.80	τ=0.90	Held-out
Tool-call gating	Qwen2.5-14B	4	0.83	1.00/0.84	0.79/0.41	0.17/0.09	+0.051
Content moderation	logistic regression	5	0.89	1.00/1.00	1.00/0.91	0.97/0.85	+0.176
Lesion classification	fine-tuned ResNet-18	7	0.86	1.00/0.81	1.00/0.72	0.52/0.46	+0.004
Recommendation	matrix factorization	18	0.76	1.00/0.86	0.65/0.47	0.32/0.04	+0.008

survive retraining. The tool-call gate covers single-turn function selection, which speaks to when an agent's individual calls can be certified but not to multi-turn or long-horizon trajectories. Within those bounds, the table describes the same statistical procedure applied to different success events and candidate populations (Table 13).

I Exact Coverage and Additional Planner Comparisons

The main text treats U_J as an approximate objective and reports several additional baselines. This appendix makes the approximation exact on problems small enough that every state can be enumerated, so the error can be read off with no Monte Carlo noise. It also specifies the additional planners and controls that those comparisons use.

I.1 Random selected support

Condition on planning so that the score, reporting map, and starting thresholds are frozen. For a segment S at threshold t, selected certification support is $M \sim \mathrm{Bin}(N, q_S(t))$. Conditional on $M = m$, the errors follow $\mathrm{Bin}(m, r_S(t))$. Thus the exact expected population coverage of this

Table 6: Exact objective comparison over 36 finite problems. Regret is evaluated using exact expected deployed coverage, with no Monte Carlo error.

Quantity	Mean	Maximum	Positive cells
Integrated-start improvement over projected-start	0.0077	0.1440	8
Integrated-start regret to contiguous sequence optimum	0.0079	0.2266	6
Contiguous sequence regret to unrestricted optimum	0.0189	0.1368	10

starting candidate is

$$v_{\mathrm{start}}(S,t;J) = q_S(t)\sum_{m=0}^{N}\mathbb{P}(M=m)\,\pi(m,r_S(t);1-\tau,\delta/J), \tag{13}$$

with $\pi(0,r;\alpha,\gamma)=0$ and zero contribution when $q_S(t)=0$. Population coverage is measured on independent traffic, not on the certification sample. At $N=58$, $q=0.5$, $r=0$, $\tau=0.9$, $\delta=0.05$, and $J=1$, certification requires at least 29 selected examples. The projected score in Equation (10) is therefore 0.500, whereas Equation (13) is $0.5\mathbb{P}(M\geq 29)=0.2761$. Replacing random support by its mean is not a conservative approximation.

I.2 The deployed threshold sequence

Let $t_0>\cdots>t_L$ be the frozen grid starting at the planned threshold, $q_\ell=q_S(t_\ell)$, and B_ℓ the event that test ℓ passes. Stopping at the first failure gives exact expected coverage

$$v_{\mathrm{seq}} = q_0\mathbb{P}(B_0)+\sum_{\ell=1}^{L}(q_\ell-q_{\ell-1})\mathbb{P}(B_0\cap\cdots\cap B_\ell). \tag{14}$$

For each certification realization, coverage is the sum of these nonnegative traffic increments whose preceding tests all passed. Taking expectations proves the formula. Consequently $v_{\mathrm{seq}}\geq v_{\mathrm{start}}(S,t_0;J)$ for the same frozen policy. This population inequality does not make a planning-set estimate a lower confidence bound, nor does it imply that the starting objective maximizes deployed coverage. Refitting a score changes the policy to which the calculation applies.

Across disjoint segments, expected coverage is additive by linearity of expectation; independence of unit counts is unnecessary. Equation (12) therefore also optimizes exact starting or sequence values when supplied those segment values, for a fixed order and J. It does not search noncontiguous partitions. The reported large experiments retain their original projected objective; the following exact study checks what that approximation misses.

I.3 Independent exact reference

The 36 cells cross $N=20,30,40$, $\tau=0.7,0.8$, balanced/skewed traffic, and $J=1,2,3$, at $\delta=0.2$. Four groups have two score levels. For each nonempty group subset, multinomial counts enumerate high-score correct/error, low-score correct/error, and outside examples. Both possible starting thresholds and all partitions are evaluated. The exact count-state sizes are 10,626, 46,376, and 135,751. The independent multinomial calculation and binomial integration agree within 4.9×10^{-15}.

I.4 Planning accuracy and meaningful service

I.4.1 What an accurate planning score can establish

For a finite pool of frozen candidates, let S_c denote the exact expected starting coverage, D_c the expected deployed coverage under the downward sequence, and $\widehat{S}_c$ the planning score. If $\widehat{c}$ maximizes that score, define $\epsilon=\max_c|\widehat{S}_c-S_c|$ and $\Delta=\max_c(D_c-S_c)$. Equation (14) gives $\Delta\geq 0$, which yields the elementary diagnostic bound

$$\max_c D_c - D_{\widehat{c}} \leq 2\epsilon+\Delta. \tag{15}$$

Indeed, for a deployed-coverage maximizer c^*, the left-hand side is at most $S_{c^*} - S_{\widehat{c}} + \Delta$, which in turn is at most $\widehat{S}_{c^*} - \widehat{S}_{\widehat{c}} + 2\epsilon + \Delta$. The score difference is nonpositive. Accurate starting scores therefore control only part of the loss: the threshold sequence can change which candidate is best. This is an elementary finite-pool inequality, not a new confidence bound or a guarantee against partitions absent from the pool. Its unknown population terms are diagnostics, not quantities available to the deployed planner.

I.4.2 Intermediate-size integration comparison

Before execution, we froze a sixteen-setting design crossing $K = 8, 16$, total label budgets of 400 and 1,200, balanced versus Zipf traffic, and strong versus weak scores, using the existing heterogeneous-error generator. Targets are $\tau = 0.9$ and $\delta = 0.1$, with five independent plans per setting. Twenty percent of labels go to planning and the remainder to certification. Both $J = 3$ and $J = 5$ are evaluated on each plan with 100 shared, nested certification draws. Both planners use the same empirical traffic frequencies, Jeffreys error estimates, reliability order, and twenty thresholds $0, 0.05, \ldots, 0.95$. The only difference is the segment objective: Equation (10) for projection, or integration over every support count in Equation (13). Both planners use the same dynamic program and stop at the first failed certification test. All failures and abstain-all outcomes are retained.

Mean deployed coverage is 0.2699 for projection, 0.2734 for integration, and 0.2275 for the support-balanced reference. Integration changes a map or starting threshold in 107 of 160 plan-by-J comparisons. Averaging the two granularities within each of eighty independent plans yields sixteen wins, seventeen losses, and forty-seven ties, with effects ranging from -0.0984 to $+0.1841$. The standard error of 0.0035 uses within-setting variation across the five plans; it measures simulation error around this fixed-design mean, not robustness on unseen tasks. Table 9 retains all 32 condition means. Integration remains a plug-in calculation with estimated rates, not a truth-informed oracle or an optimizer of the full deployed sequence.

On one CPU with single-threaded BLAS, mean planning times are 0.0042 and 0.1651 seconds for projection and integration, respectively. Timing includes segment scoring and the dynamic program, but excludes shared data generation, group ordering, cutoff-cache preparation, and certification. Those shared costs and all inputs are saved separately; method timing order alternates by plan. The complete execution, including the saved-result audit below, took 46.45 seconds. The ratio is implementation- and machine-dependent, not an asymptotic claim.

I.4.3 Score errors and condition dependence

An exploratory audit of the saved matched-pool study compares each winning candidate's estimated score with its true integrated starting coverage. Mean signed overestimation is 0.3429 for construction-data selection and 0.0469 for held-out selection. Mean regret to the best integrated starting coverage within the same pool falls from 0.0723 to 0.0338. These observations support the claim that score assessment improves when conducted on separate data, but they do not isolate pure selection bias: score projection, estimated rates, and selection among different frozen policies all enter the comparison. Table 10 includes every factor stratum, and the supplement retains all 64 condition effects and their paired plan-level outcomes. These post hoc summaries do not select a preferred regime or support multiple unadjusted significance claims.

I.4.4 Named-scenario service

For HWU, we also replay the saved certificates against the eighteen declared scenario boundaries. Every candidate preserves those boundaries. A scenario is counted once if any of its units passes, regardless of how many units it contains. Table 11 pairs this count with traffic coverage rather than assigning an invented exchange rate between the two. The direct constrained planner has higher means on both measures than full selection in each of four architecture-by-J settings. This aggregate comparison does not constitute simultaneous confidence of dominance for individual scenarios. Coverage and scenario service can disagree against other baselines: on DeBERTa at $J = 21$, support balancing serves less traffic but reaches 5.46 scenarios against the direct planner's 5.32. No result here establishes that a stakeholder would accept the unserved intents or these trade-offs. The audit reuses ten existing trained models and their nested resplits; it is not independent application confirmation.

I.4.5 MATCHED-POOL MECHANISM CONTROLS

To ensure that every selector chooses from an identical slate, each synthetic construction role produces the same eight maps; the selectors differ only in how they choose among them. The uniform control averages coverage over the available candidates, yielding the expected coverage of a uniform random choice. Construction and held-out selection apply the same projected score and differ only in which data role supplies the rates. The empirical-traffic control sums selected frequencies for units whose Jeffreys error estimate is at most α, dropping the power term. The random-only selectors draw from the first one, three, or five random candidates. Primary comparisons keep scores and starts frozen, while refit comparisons update the starts from all planning observations (the synthetic scores are fixed by the generator; real-data refitting also updates the fitted correctness score).

Every condition retains 20 independent planning draws with 100 paired certification draws per plan, and a vectorized first-failure implementation is checked against the original scalar certifier. The source artifacts store per-plan effects, their Monte Carlo standard errors, and per-condition win/loss summaries and factor strata. The design draws fresh samples from declared generator regimes, so it is prospective within those regimes but does not establish generalization to an unknown distribution of tasks.

As a calibration check, the largest condition-mean false-certification rate across the fifteen evaluated methods is 0.0305 against a nominal familywise level of 0.10; this is descriptive rather than a proof of calibration over new populations. The audit stores pointwise 95% exact binomial intervals conditional on each frozen plan, together with Monte Carlo standard errors across the twenty independent plan rates. Uniform controls, which average fractional outcomes, receive plan-level standard errors instead of binomial intervals, and no interval treats the 2,000 nested trials as independent planning replications.

I.4.6 CANDIDATE POOL SIZE

If most of the benefit is the act of selecting among diverse candidates, two questions follow: does a larger random pool keep helping, and do the structured candidates add anything on top of it? We sweep the number of random-order candidates through $R \in \{1, 5, 10, 25, 50, 100\}$ on the synthetic cells, scoring every candidate once on the selection split and certifying the selected one with the same first-failure certifier as the main study. Held-out selection over random candidates alone improves quickly and then saturates: mean certified coverage rises from 0.258 at $R = 1$ to 0.274 at $R = 5$ and no further, reaching only 0.274 at $R = 100$. A larger random pool is therefore not the missing ingredient. Adding the three structured candidates (support-balanced, equal-count, and direct plug-in) to the R random ones raises the mean by about 0.010 at every pool size, but the median improvement is zero throughout, so the structured candidates help in a minority of cells rather than in the typical one. Both observations agree with the matched-pool decomposition: the working ingredient is disciplined selection over a handful of diverse groupings, not the size of the pool or the structure of its members.

I.4.7 SYNTHETIC MECHANISM DESIGN

The mechanism study crosses label budgets 800/3,200, 12/40 classes, balanced/Zipf traffic, homogeneous/heterogeneous errors, and strong/weak scores at $\tau = 0.9$ and $\delta = 0.1$, evaluating each of these 32 settings at $J = 3, 8$. Planning receives 20% of the labels, split equally between construction and selection, and the remainder is reserved for certification. The 640 independent planning draws are shared between the two values of J, giving 64 conditions of 20 plans each with 100 certification draws per plan; those nested certification outcomes are not additional independent plans. This design was frozen before the synthetic draws, whereas the real replay below reuses previously inspected traces and is exploratory.

I.4.8 SCENARIO-PRESERVING COMPARISON

Every HWU candidate here preserves the 18 source scenarios. The direct candidate orders groups by scenario and then by construction reliability, forbidding cross-scenario segments, while the fixed baselines give each scenario at least one unit and assign the remaining units to the scenario with the largest support per current unit, subject to its class count (the equal-count baseline uses class count in this allocation, and within each scenario the usual balancing rule applies). Random maps permute

Table 7: Scenario-preserving HWU replay: full held-out selector minus each comparator, as five per-model effects.

Architecture	J	Comparator	Five model effects
deberta	21	Support	0.0345, 0.0076, 0.0240, 0.0150, -0.0037
deberta	21	Random selection	0.0361, -0.0043, -0.0009, 0.0101, -0.0181
deberta	21	Direct constrained	-0.0302, -0.0404, -0.0708, -0.0805, -0.0368
deberta	34	Support	0.1466, 0.1350, 0.1682, 0.1638, 0.1157
deberta	34	Random selection	0.1324, 0.1407, 0.1505, 0.1452, 0.1123
deberta	34	Direct constrained	-0.0098, -0.0276, -0.0470, -0.0298, -0.0469
distilroberta	21	Support	0.0329, 0.0304, 0.0507, 0.0104, 0.0291
distilroberta	21	Random selection	0.0197, 0.0700, 0.0284, 0.0075, 0.0322
distilroberta	21	Direct constrained	-0.0162, -0.0294, -0.0212, -0.0334, -0.0427
distilroberta	34	Support	0.1393, 0.1576, 0.1419, 0.1338, 0.1639
distilroberta	34	Random selection	0.1232, 0.1384, 0.1385, 0.1192, 0.1622
distilroberta	34	Direct constrained	-0.0313, -0.0352, -0.0318, -0.0292, -0.0060

Table 8: Application replay on a shared eight-candidate pool: mean certified coverage over ten resplits, all methods refit.

Domain	τ	J	Support	Direct	Full selection	Random selection
Function calls	0.70	3	0.8464	0.9159	0.9013	0.9245
Proposed blocks	0.90	3	0.8746	0.7312	0.7163	0.8786
Lesion images	0.85	3	0.7239	0.7602	0.8726	0.7363
Proposed items	0.85	5	0.3604	0.4475	0.3362	0.3077

the within-scenario order and reuse the same support allocation. The construction is feasible at both $J = 21$ and $J = 34$, and every saved map is checked for boundary preservation. Because the direct dynamic program also optimizes how units are allocated across scenarios, its advantage does not isolate within-scenario ordering from that allocation. All selectors share the same construction score and candidate slate, and refitted comparisons share the final planning score.

J Familywise Budget Allocation

The deployed certifier spends an equal Bonferroni share δ/J of the familywise budget on every unit. This is not the only validity-preserving choice: any frozen allocation $\gamma_1, \ldots, \gamma_J$ with $\sum_j \gamma_j \leq \delta$ certifies each unit j at its own level γ_j, and a union bound over units still gives familywise control at level δ. This appendix asks how much expected certified coverage the best such allocation can add over the equal split, at the population level where the traffic share q_j and true error r_j of each unit are known. The answer is an opportunity ceiling, not a deployable gain: it uses the true errors and so bounds what any estimator of the allocation could hope to recover, exactly as the population oracle bounds the partition planner.

We sweep a factorial of $J \in \{3, 5, 8\}$ units, certification budgets $N \in \{800, 3200\}$, balanced or heavy-tailed (Zipf) traffic, homogeneous or heterogeneous true errors, and targets $\tau \in \{0.80, 0.90\}$ at $\delta = 0.10$, drawing 150 random unit configurations per cell. For each configuration we compare the equal split against the coverage-maximizing allocation, computed exactly by a $(\max, +)$ dynamic program over a fine budget grid; grid discretization makes the reported gain a slight underestimate. Across the 7,200 configurations the optimal allocation adds a mean of 0.038 certified coverage (median 0.028, maximum 0.288), exceeding 0.05 in about 30% of configurations. The gain grows exactly where the equal split is most wasteful (Table 12): with more units, a stricter target, and a scarcer budget, where a few pressed units would otherwise consume their share without certifying. Because this population opportunity is an order of magnitude larger than the 0.0027 contributed by the structured partition candidate, we regard the allocation of the testing budget, rather than the search over partitions, as the more promising direction for closing the gap to the oracle.

The population study uses the true errors, so a fair question is whether a deployable estimator can capture any of it. We repeat the sweep in a deployable form: for each drawn truth we take a single

Table 9: Per-condition random-support integration minus mean-support projection, with Monte Carlo standard errors across five plans.

Groups	Labels	Traffic	Score	J	Coverage difference	SE
16	1200	balanced	strong	3	0.0066	0.0064
16	1200	balanced	strong	5	-0.0444	0.0305
16	1200	zipf	strong	3	0.0000	0.0000
16	1200	zipf	strong	5	0.0000	0.0000
16	1200	balanced	weak	3	0.0007	0.0004
16	1200	balanced	weak	5	-0.0001	0.0001
16	1200	zipf	weak	3	0.0000	0.0000
16	1200	zipf	weak	5	-0.0009	0.0007
16	400	balanced	strong	3	-0.0139	0.0139
16	400	balanced	strong	5	0.1191	0.0802
16	400	zipf	strong	3	0.0000	0.0000
16	400	zipf	strong	5	0.0156	0.0156
16	400	balanced	weak	3	0.0002	0.0002
16	400	balanced	weak	5	0.0002	0.0002
16	400	zipf	weak	3	0.0000	0.0000
16	400	zipf	weak	5	0.0000	0.0000
8	1200	balanced	strong	3	0.0083	0.0083
8	1200	balanced	strong	5	0.0015	0.0015
8	1200	zipf	strong	3	-0.0158	0.0110
8	1200	zipf	strong	5	0.0148	0.0140
8	1200	balanced	weak	3	-0.0017	0.0015
8	1200	balanced	weak	5	-0.0001	0.0004
8	1200	zipf	weak	3	-0.0001	0.0001
8	1200	zipf	weak	5	0.0000	0.0000
8	400	balanced	strong	3	0.0539	0.0445
8	400	balanced	strong	5	0.0092	0.0088
8	400	zipf	strong	3	0.0000	0.0000
8	400	zipf	strong	5	-0.0402	0.0391
8	400	balanced	weak	3	0.0000	0.0000
8	400	balanced	weak	5	0.0000	0.0000
8	400	zipf	weak	3	0.0000	0.0000
8	400	zipf	weak	5	0.0000	0.0000

Table 10: Factor summaries of full-pool minus random-five held-out selection; ranges are over condition means, not confidence intervals.

Factor	Value	Mean	Minimum	Maximum	Win/tie/loss
Groups	12	0.0054	-0.0380	0.1202	19/2/11
Groups	40	-0.0000	-0.0870	0.1344	10/2/20
Error variation	heterogeneous	-0.0034	-0.0870	0.1098	11/1/20
Error variation	homogeneous	0.0087	-0.0852	0.1344	18/3/11
Labels	3200	-0.0057	-0.0870	0.0473	12/1/19
Labels	800	0.0110	-0.0566	0.1344	17/3/12
Score	strong	0.0054	-0.0870	0.1344	18/0/14
Score	weak	-0.0000	-0.0011	0.0021	11/4/17
Traffic	balanced	-0.0011	-0.0870	0.1344	13/1/18
Traffic	zipf	0.0064	-0.0543	0.1202	16/3/13

planning sample, form Jeffreys error estimates and frequency traffic estimates, allocate γ_j by the same exact dynamic program on those estimates alone, freeze it, and then evaluate the frozen policy exactly at the true (q_j, r_j) by integrating over random certification support. The equal split is evaluated identically, and the paired difference is the honest deployable gain. Across 2,880 truths the estimator improves realized coverage by a mean of 0.011 (median 0.005), is positive on 68% of truths, and reaches a cell mean of 0.075 in the hardest regime (eight units, target 0.90, scarce budget, heavy-tailed traffic); it turns slightly negative only in easy regimes where the equal split is already near optimal and estimation noise misallocates. The deployable gain is thus about a quarter of the population

Table 11: HWU service under fixed scenario boundaries: served traffic and the number of the eighteen scenarios with at least one passing unit, over ten resplits of five models.

Architecture	J	Planner	Coverage	Scenarios with service
deberta	21	Support	0.3252	5.46
deberta	21	Direct constrained	0.3924	5.32
deberta	21	Full selection	0.3406	5.26
deberta	21	Random selection	0.3360	5.48
deberta	34	Support	0.1742	3.52
deberta	34	Direct constrained	0.3523	4.58
deberta	34	Full selection	0.3201	4.38
deberta	34	Random selection	0.1839	3.72
distilroberta	21	Support	0.2246	3.88
distilroberta	21	Direct constrained	0.2839	4.16
distilroberta	21	Full selection	0.2553	4.04
distilroberta	21	Random selection	0.2237	3.92
distilroberta	34	Support	0.0861	1.94
distilroberta	34	Direct constrained	0.2601	3.64
distilroberta	34	Full selection	0.2334	3.46
distilroberta	34	Random selection	0.0971	2.08

ceiling, yet several times the structured-candidate contribution, and it is realized without any change to the validity contract. A validity audit that injects unsafe units ($r_j > \alpha$) across 3,800 configurations confirms this: the allocated procedure's familywise false-certification probability averages 0.0002 and never exceeds 0.02, comfortably within the nominal $\delta = 0.10$, as the union bound guarantees. The estimator can also be improved by accounting for uncertainty in the error rate. Replacing the Jeffreys plug-in with the posterior-predictive availability, the exact average of the availability over the Beta posterior of r_j (a Beta-Binomial tail probability), raises the deployable gain from 0.011 to 0.014 and its positive rate from 68% to 74%, recovering close to two-fifths of the population ceiling; a purely pessimistic upper-quantile rule, by contrast, hurts (mean 0.005), so it is the calibrated posterior average, not conservatism, that helps.

Table 12: Mean coverage gained by optimal familywise-budget reallocation over the equal δ/J split, averaged over budgets, traffic shapes, and error profiles.

Units J	τ=0.80	τ=0.90
3	0.016	0.026
5	0.027	0.047
8	0.043	0.067

J.1 Two ablations: the reliability order and the selection split

Two design choices in the planner invite scrutiny, and we test both on the synthetic cells with the first-failure certifier of the main study.

The main text reports the central result of the order ablation: against twenty random orders, the frozen reliability order beats the median random order only 33% of the time and the best random order only 9%, with mean certified coverage 0.249 versus 0.282 for a random order. This experiment motivates the main-text interpretation that the dynamic program is useful as a population diagnostic, while candidate diversity and independent comparison carry most of the finite-data recovery.

The second choice is the single fifty-fifty split of the planning data into construction and selection. Holding the candidate pool fixed and varying only how much data scores it, selecting on all of the planning data, or on a five-fold cross-fit of it, does not beat selecting on one held-out half: the median change in realized coverage is zero and the mean is slightly negative, because scoring on data that also built the candidates reintroduces exactly the optimism the split was meant to remove. Cross-fitting is therefore not the missing ingredient; the disjoint split is already doing its job, and the remaining gap to the oracle lies in the objective and the budget, not in how the selection data is reused.

Table 13: Application contracts. Coverage is the served fraction of the stated candidate population, not necessarily the original dataset or deployment traffic.

Benchmark	Candidate population	Success event	Reporting unit
BFCL	Existing single-turn function-call records	Official AST check passes	Four recorded tool categories
Civil Comments	Comments predicted toxic, in a constructed class mixture	Toxicity label at least 0.5, irrespective of subtype correctness	Predicted policy subtype
DermaMNIST	Validation lesion images	Predicted lesion class is correct	Predicted class
MovieLens	Rated pairs with predicted liking probability at least 0.5	Observed rating at least four	First listed item genre

K Application Populations and Reproduction Details

A shared trace format encodes binary success as label equality for implementation convenience; this format does not make the underlying tasks identical. BFCL uses the official abstract syntax tree (AST) checker for a frozen Qwen2.5-14B single-turn trace (Patil et al., 2025). Civil Comments (Borkan et al., 2019) and MovieLens (Harper & Konstan, 2016) use prefiltered candidates. DermaMNIST (Yang et al., 2023) supplies a seven-class vision benchmark. After the deterministic cap, BFCL has 753 evaluation records, Civil Comments 8,000, DermaMNIST 1,003, and MovieLens 8,000. Certification and IID deployment are disjoint roles sampled from those pools.

K.1 Trace construction

The moderation source retains all toxic comments and samples non-toxic comments to produce a 300,000-row pool with 48.1% toxic records. A 60/20/20 split trains and evaluates a word-level term-frequency inverse-document-frequency (TF-IDF) logistic regression, while a ComplementNB classifier and a character-level TF-IDF logistic regression provide auxiliary signals. Non-toxic predictions fall outside the gate. A toxic comment counts as a justified block even when the predicted subtype is wrong, so this study does not certify subtype attribution.

The MovieLens model is a binary matrix-factorization classifier with 64-dimensional embeddings, trained for eight epochs with batch size 8,192, Adam learning rate 0.01, and dropout 0.2. An 80/10/10 random rated-pair split separates training, evaluation, and test sets. Item-kNN and eight dropout passes provide auxiliary scores. Because users and items can appear in multiple splits and the candidate population consists of observed ratings, interpretation as prospective recommendation is limited; no cold-start or user-independent deployment guarantee is established.

DermaMNIST uses 224-pixel images and a pretrained ResNet-18 fine-tuned for five epochs with AdamW, learning rate 10^{-4}, batch size 64, and weight decay 0.01. A separately fine-tuned ViT-B/16 supplies auxiliary confidence, and eight dropout passes supply disagreement scores. These constitute multiple model computations even though there is only one base-predictor instance in the reported frontier. Clinical triage decisions, patient outcomes, and patient-level generalization are outside the scope of this experiment.

L Reproducibility and Artifact Integrity

All reported results are tied to preserved experimental artifacts and a fully pinned software environment, so that every figure, table, and headline number can be regenerated deterministically from the recorded inputs and reconstructed dependencies.

The supplementary material contains the experiment implementations, their tests, the preregistration records, and machine-readable result tables, including raw outputs for the exactness and selection studies, the four application evaluations, and the model traces needed to replay them. An automated check recomputes every registered numerical claim and generated table from these canonical results and reports any discrepancy, establishing traceability from each reported value back to the split and

model that produced it. The accompanying documentation details the exact steps for reconstructing the environment and regenerating the figures and manuscript.