CONFIRM: EXACT POST-NOMINATION GATING FOR TEST-TIME MODEL REPLACEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Ranking adapted models does not determine whether the best-ranked candidate should replace the incumbent: every candidate may be worse, and labels used for adaptive nomination cannot be naively reused for authorization. We formulate limited-label replacement on an enumerated target pool and introduce CONFIRM, a selector-agnostic exact gate. Conditioning on the realized nomination history turns observed nominee-relevant outcomes into a fixed incumbent-relative offset; only a fresh uniform sample without replacement from unresolved disagreements remains random. The resulting paired finite-population p-value is conditionally valid, and Holm correction controls family-wise false replacement across predeclared deployment units. We characterize finite-budget certifiability through the offset, unresolved win, loss, and tie composition, and remaining budget, showing that equal accuracy gains can have different and even budget-reversing certifiability. We evaluate CONFIRM across TTA, CLIP adaptation, refusal-dominated transfer, and quantum error correction. Compared with valid finite-population alternatives, CONFIRM authorizes more beneficial replacements under partial, low-budget confirmation. In the predeclared DomainBed study, CONFIRM makes no harmful replacement across 174,080 repeated gated decisions; ungated harm ranges from 21.9% to 84.1%, including a random-nomination control.

1 INTRODUCTION

Test-time adaptation (TTA) produces candidate predictors from an unlabeled target stream and can improve accuracy under shift (Wang et al., 2021; Niu et al., 2022; 2023), but adaptation can also forget or collapse under changing streams (Zhao et al., 2023; Wang et al., 2022; Yuan et al., 2023). In replacement, these errors are asymmetric: declining an uncertain gain retains the incumbent, whereas authorizing a worse candidate degrades the deployed predictor. This leaves a distinct deployment question: a selector asks *which candidate looks best*, whereas a replacement gate asks *whether the nominee has enough evidence to displace the incumbent*. We study this authorization problem on a predeclared, enumerated target pool, retaining the incumbent unless strict improvement is established. Operationally, an online system can freeze an adapter checkpoint at a predeclared boundary, buffer an episodic pool, freeze all predictions, spend labels to nominate and confirm, and deploy the chosen fixed model for that episode. A later checkpoint or pool is a new unit. The guarantee ends at the current decision: it does not cover future drift, updates during confirmation, or optional stopping. Figure 1 illustrates how nomination and authorization can diverge on actual fresh-confirmation examples.

Existing label-efficient methods estimate target risk or nominate candidates (Kossen et al., 2021; Okanovic et al., 2025; Kim et al., 2024; Ashouritaklimi et al., 2026), while post-selection methods assess a data-dependent choice (Rink & Brannath, 2025; Farzaneh & Simeone, 2026). Neither endpoint gives exact paired authorization against an incumbent on an actively labeled finite pool. Nomination labels cannot generally be treated as a representative evaluation sample: discarding them wastes evidence, testing every candidate adds multiplicity, and naively pooling them with fresh labels can invalidate the test. We therefore condition nomination evidence into a fixed offset and obtain test randomness only from a fresh without-replacement draw over unresolved nominee disagreements. This is a statistical approval interface, not a new selector: the central contribution is the post-nomination finite-pool conditioning construction rather than the classical paired tail itself.

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

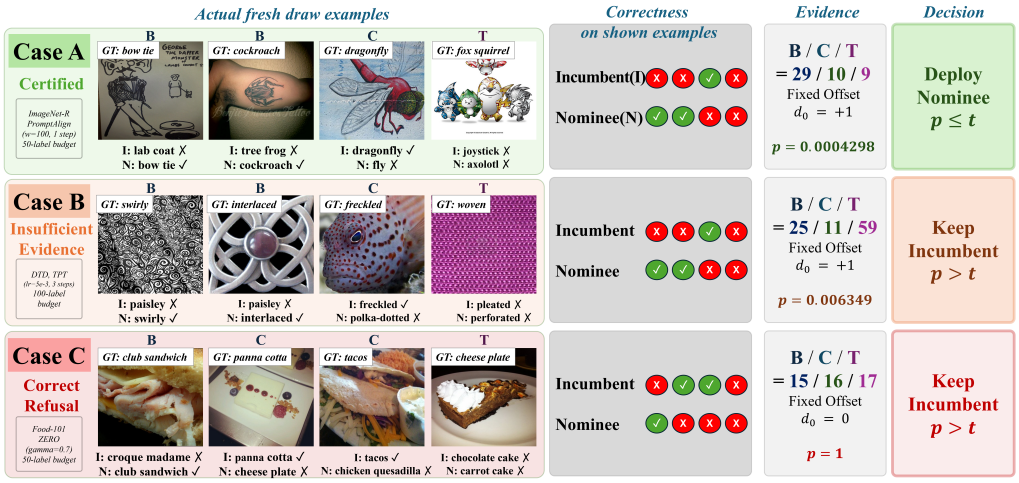

Figure 1: **Nomination does not imply replacement authorization.** Each row shows representative examples from an actual fresh confirmation draw, paired incumbent/nominee correctness, evidence, and the CONFIRM decision. B, C, and T denote nominee wins, incumbent wins, and paired ties (both wrong). Case A authorizes replacement; Case B retains the incumbent for insufficient evidence; and Case C illustrates correct refusal. Decisions use the full fresh draw plus the fixed Stage 1 offset d_0, not only the displayed examples; $t = 0.05/8$ is the prespecified threshold.

Contributions.

- C1. We formulate limited-label finite-pool replacement as an authorization problem distinct from ranking: nomination proposes a candidate; deployment requires incumbent-relative evidence.
- C2. Our main method contribution is CONFIRM: it conditions nominee-relevant nomination outcomes into their realized offset and obtains fresh randomness only from unresolved disagreements. It remains valid after arbitrary adaptive acquisition and nomination over a fixed prediction family, and Holm controls family-wise false replacement.
- C3. We derive exact finite-budget certifiability, including equal-gain non-ordering and budget-dependent reversals, and show useful low-budget authorization gains over separate marginal hypergeometric bounds and a direct unconditional paired exact gate at equal label cost and validity.

2 Related Work

Label-efficient evaluation and nomination. Prior work estimates target risk or accuracy (Kossen et al., 2021; Ashouritaklimi et al., 2026; Garg et al., 2022; Lee et al., 2024), selects candidates (Okanovic et al., 2025; Kay et al., 2025; Li et al., 2024; Kim et al., 2024; Danilowski et al., 2026), and stabilizes or changes adaptation (Yang et al., 2023; Liu et al., 2025). AETTA, MORPHEUS, and TACT respectively provide label-free accuracy estimation, geometry-based strategy selection, and candidate generation, rather than finite-pool incumbent-relative authorization. Rankings can depend on the selection rule (Cygert et al., 2026).

Risk control and post-selection inference. Baseline retention under certified improvement relates to high-confidence policy improvement and risk control (Thomas et al., 2015; Angelopoulos et al., 2025; Bates et al., 2021; Laufer-Goldshtein et al., 2023). Pareto Testing optimizes configurations and tests risk constraints on held-out calibration data, without reusing adaptive nomination evidence for paired finite-pool replacement. Post-selection inference uses splitting, simultaneous inference, conditioning, or selective randomization (Kuchibhotla et al., 2022; Fithian et al., 2014; Rink & Brannath, 2025; Farzaneh & Simeone, 2026; Freidling et al., 2024); transferability loss instead gives an unlabeled worst-case shift bound (Nikzad-Langerodi & Fonseca Diaz, 2026). Unlike selective

randomization inference, which uses treatment-assignment randomness, CONFIRM conditions on nomination history, retains observed nominee-relevant outcomes as a fixed incumbent-relative offset, and obtains randomness from fresh uniform sampling without replacement over unresolved disagreements.

Paired tests and certifiability. McNemar-type tests exploit discordant paired outcomes (McNemar, 1947), and their power depends on evidence frequency and direction (Lachin, 1992; Berger & Sidik, 2003). CONFIRM contributes a post-nomination finite-pool construction and its finite-budget certifiability analysis, not a new classical paired tail. Supplementary Table 3 compares these formulations.

3 Problem Setup: Finite-Pool Replacement

Fix an enumerated target pool $\mathcal{P} = \{(x_i, y_i)\}_{i=1}^{N}$ with hidden labels and predictions from an incumbent h_0 and $K-1$ frozen candidates. "Arbitrary adaptive nomination" means that queried labels may affect both acquisition and nomination within this fixed prediction family; candidate predictions are not retrained or regenerated during confirmation. Under 0–1 loss, candidate j has incumbent-relative risk difference

$$\Delta_j = \mathcal{R}(h_j) - \mathcal{R}(h_0) = \frac{1}{N}\sum_{i=1}^{N} Z_{ij}, \qquad Z_{ij} = \mathbb{I}[h_j(x_i) \neq y_i] - \mathbb{I}[h_0(x_i) \neq y_i]. \tag{1}$$

Let $\delta_j = -\Delta_j = \mathrm{Acc}(h_j) - \mathrm{Acc}(h_0)$. The replacement null is

$$H_j : \Delta_j \geq 0 \qquad \Longleftrightarrow \qquad H_j : \delta_j \leq 0.$$

Equality remains in the null, so replacement requires evidence of strict improvement.

The total budget b counts unique target labels across both stages. A selector may spend a subset adaptively before nominating $j^\star$; replacement occurs only after rejecting $H_{j^\star}$. The guarantee is transductive: Δ_j concerns this enumerated pool, not future stream examples.

When $h_j(x_i) = h_0(x_i)$, $Z_{ij} = 0$ for every label. Define

$$D_j = \{i : h_j(x_i) \neq h_0(x_i)\}, \qquad U = \bigcup_{j=1}^{K-1} D_j.$$

Only D_j affects candidate j; points outside U are uninformative. For $i \in D_j$,

$$Z_{ij} = \begin{cases} -1, & h_j \text{ correct (candidate win)}, \\ +1, & h_0 \text{ correct (incumbent win)}, \\ 0, & h_j \text{ and } h_0 \text{ both wrong (paired tie)}. \end{cases}$$

Because differing predictions cannot both be correct, this paired cancellation reduces authorization to win–loss imbalance (McNemar, 1947).

A deployment unit is one enumerated pool together with its frozen incumbent, candidate family, and one replacement decision. A deployment family is a predeclared set of such units sharing one error budget. A false replacement authorizes any nominee with $\Delta_{j^\star} \geq 0$; Holm controls the probability of at least one such decision under arbitrary dependence (Holm, 1979). A full-pool tie therefore counts as false replacement, whereas empirical harm requires $\Delta_{j^\star} > 0$. Arbitrary-dependence validity is robust but can be conservative for correlated units; this is the chosen safety–power trade-off, and exploiting dependence would require additional assumptions.

4 CONFIRM: The Confirmation Gate

Figure 2 summarizes the protocol and contributions C1–C3.

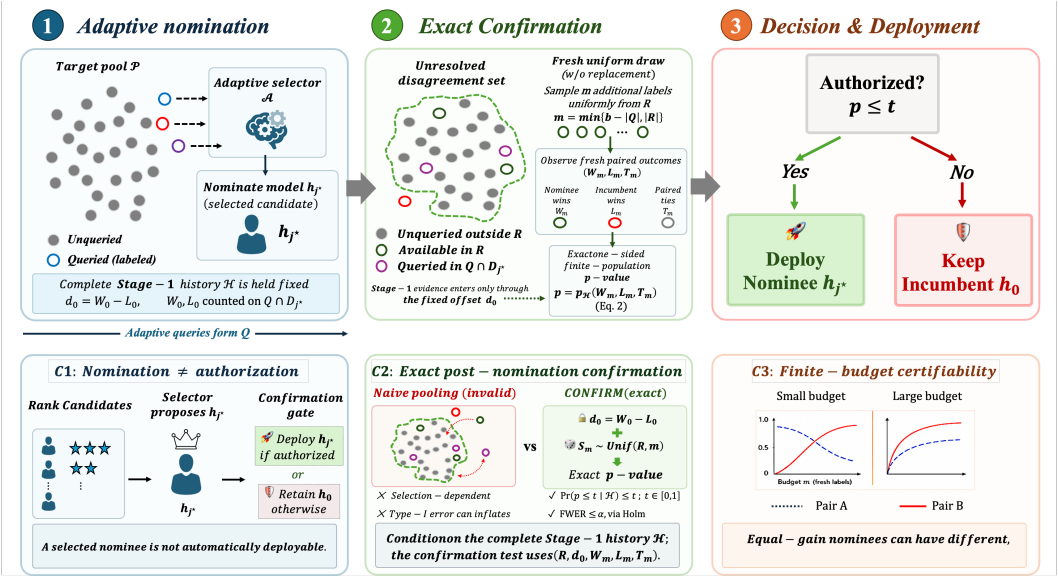

Figure 2: Overview of CONFIRM. Stage 1 nominates adaptively from a fixed prediction family. Stage 2 conditions on the nomination history, retains nominee-relevant Stage 1 label evidence as a fixed offset, and samples unresolved disagreements uniformly without replacement for exact confirmation. Stage 3 deploys only authorized nominees; otherwise it retains the incumbent. The lower panels summarize C1–C3: nomination versus authorization, exact post-nomination validity, and finite-budget certifiability.

4.1 Two-Stage Protocol

Let $\mathcal{A}$ return nominee $j^\star$ after adaptive Stage 1 acquisition. Its history $\mathcal{H}$ records the queried set Q, revealed labels, relevant internal randomization, and $j^\star$. After nomination, $j^\star$ and its pool predictions are fixed before Stage 2 labels are drawn. If W_0 and L_0 are nominee and incumbent wins in $Q \cap D_{j^\star}$, conditioning fixes the offset

$$d_0 = W_0 - L_0;$$

labels outside $D_{j^\star}$ do not affect the comparison.

For the unresolved set $R = D_{j^\star} \setminus Q$ with $P = |R|$, Stage 2 draws

$$m = \min\{b - |Q|, P\}$$

points uniformly without replacement. This draw supplies the test's only randomness; Stage-1 evidence enters through d_0.

4.2 Offset-Aware Exact Confirmation

Condition on $\mathcal{H}$. Stage 2 observes nominee wins W_m, incumbent wins L_m, and ties T_m, with $n_m = W_m + L_m$. Let M be the unknown number of informative outcomes among the P unresolved points and B their nominee-win count. The null requires

$$d_0 + 2B - M \le 0.$$

For fixed M, the largest null-compatible nominee-win count that also leaves the observed L_m incumbent wins is

$$B_M^{(0)} = \min\left\{M - L_m, \left\lfloor \frac{M - d_0}{2} \right\rfloor\right\}.$$

The compatible nuisance set is therefore

$$\mathcal{M} = \left\{M \in \mathbb{Z} : M \le P - T_m,\ B_M^{(0)} \ge W_m\right\}.$$

Conditional on n_m, the informative observations are uniform among the M informative outcomes. The finite-sample exact, generally conservative p-value is

$$p = \max_{M \in \mathcal{M}} \Pr\{X_M \geq W_m\}, \qquad X_M \sim \text{Hypergeom}\Big(M, B_M^{(0)}, n_m\Big). \tag{2}$$

Intuitively, the test asks whether the observed Stage 2 nominee wins remain surprising even under the null-compatible unresolved population that makes them easiest to explain. If $\mathcal{M}$ is empty, no compatible null population exists and $p = 0$. If $W_m = 0$ and at least one compatible null exists, $p = 1$, including the degenerate empty-informative-population case. Thus $d_0 < 0$, $d_0 = 0$, and $d_0 > 0$ enter the same rule as certification debt, no historical credit, and certification credit, respectively; no data-dependent early stopping is used. Unlike confidence-set maximization with a correction for excluded nuisance values (Berger & Boos, 1994), Eq. (2) maximizes over all compatible M.

Theorem 1 (Selector-agnostic finite-pool validity)**.** *For any adaptive acquisition-and-nomination rule $\mathcal{A}$ over a fixed candidate prediction family, fixed finite pool, and realized history $\mathcal{H}$ whose nominee satisfies $\Delta_{j^\star} \geq 0$,*

$$\Pr(p \leq t \mid \mathcal{H}) \leq t, \qquad t \in [0, 1],$$

provided confirmation is sampled uniformly without replacement from the nominee's unresolved disagreements. After joint conditioning on all unit histories, Holm controls the probability of at least one false replacement in any predeclared finite deployment family at level α, under arbitrary dependence among unit p-values.

Remark 1 (Exactness versus exact nominal calibration). Here "exact" means finite-sample validity under the stated without-replacement sampling design, without asymptotic or distributional approximations. Discreteness and maximization over compatible nuisance populations in Eq. (2) can make the p-value strictly super-uniform, so the null rejection probability need not equal the nominal level.

Theorem 1 is proved in Appendix A.

Proposition 1 (Naive post-nomination pooling can be anti-conservative)**.** *There exists a fixed finite pool and a frozen candidate family in which all candidates exactly tie the incumbent, yet an adaptive nomination rule followed by a pooled exact test that treats nomination and confirmation outcomes as one uniform sample rejects with probability $6095/72576 \approx 0.084$ at level $\alpha = 0.05$.*

The failure is purely post-selection: although the pool and predictions are fixed, Stage 1 outcomes helped select the nominee and are therefore not a uniform test sample conditional on nomination. CONFIRM conditions on them and obtains its randomness only from Stage 2.

4.3 Finite-Budget Certifiability

Theorem 1 controls false replacement; certifiability asks how often a genuine improvement is authorized with m remaining labels. Let the unresolved population contain

$$B \text{ nominee wins}, \qquad C \text{ incumbent wins}, \qquad T \text{ ties}, \qquad P = B + C + T.$$

Then

$$\delta_{j^\star} = \frac{d_0 + B - C}{N}, \tag{3}$$

and, for $0 \leq m \leq P$,

$$(W_m, L_m, T_m) \sim \text{MultiHypergeom}(P; B, C, T; m).$$

Write $p_{\mathcal{H}}(w, l, r)$ for Eq. (2) evaluated at the corresponding confirmation outcome.

Definition 1 (Conditional certifiability profile)**.** For threshold $t \in [0, 1]$,

$$\Pi_{\mathcal{H}}(m, t) = \Pr[p_{\mathcal{H}}(W_m, L_m, T_m) \leq t \mid \mathcal{H}]. \tag{4}$$

For a single test, $t = \alpha$. Under Holm, the actual authorization probability also depends on the other unit p-values; the following corollary relates it directly to the single-unit certifiability profile.

Theorem 2 (Exact finite-budget certifiability)**.** *For fixed $\mathcal{H}$, unresolved composition (B, C, T), budget $m \leq P$, and threshold t,*

$$\Pi_{\mathcal{H}}(m, t) = \sum_{\substack{w,l,r \geq 0 \\ w+l+r=m}} \frac{\binom{B}{w}\binom{C}{l}\binom{T}{r}}{\binom{P}{m}} \mathbb{I}\{p_{\mathcal{H}}(w, l, r) \leq t\}. \tag{5}$$

Equivalently, $N_m = W_m + L_m \sim \mathrm{Hypergeom}(P, B + C, m)$ *and* $W_m \mid N_m = n \sim \mathrm{Hypergeom}(B + C, B, n)$*: the first determines the amount of informative evidence and the second its direction.*

Corollary 1 (Holm family-wise certifiability bounds)**.** *Consider a predeclared deployment family of* L *units and condition jointly on all Stage-1 histories* $\mathbf{H}$*. For a genuinely improving unit* ℓ*, let*

$$\Pi_\ell(m_\ell, t) = \Pr(p_\ell \le t \mid \mathbf{H})$$

denote its conditional certifiability profile under this joint conditioning. If Holm's procedure at family level α *is applied to the* L *confirmation* p*-values, then*

$$\Pi_\ell\Big(m_\ell, \frac{\alpha}{L}\Big) \le \Pr(\ell \textit{ is authorized by Holm} \mid \mathbf{H}) \le \Pi_\ell(m_\ell, \alpha). \tag{6}$$

The bounds hold without independence among unit p*-values.*

For $B + C > 0$, define decisiveness $q = (B + C)/P$ and directional edge $\eta = (B - C)/(B + C)$. Equation (3) becomes

$$\delta_{j^\star} = \frac{d_0 + Pq\eta}{N}.$$

For fixed N, d_0, and P, gain retains only $q\eta$, whereas Eq. (5) retains evidence frequency and direction separately.

Proposition 2 (Global gain is insufficient)**.** *Nominees can have the same* N*,* P*, offset* d_0*, and positive full-pool gain* δ*, yet different conditional certifiability profiles. Equal gain therefore need not imply equal evidence requirements.*

Proposition 3 (Budget-dependent certifiability ordering)**.** *There exist equally beneficial nominees* j, k *and budgets* $m_1 < m_2$ *such that*

$$\Pi_j(m_1, t) > \Pi_k(m_1, t), \qquad \Pi_j(m_2, t) < \Pi_k(m_2, t).$$

Thus no ranking based only on global gain represents exact certifiability at every budget.

Proposition 4 (Historical evidence as certification credit)**.** *For a fixed unresolved population and confirmation outcome,* $d_0' \ge d_0$ *implies*

$$p_{d_0'}(W_m, L_m, T_m) \le p_{d_0}(W_m, L_m, T_m), \qquad \Pi(m, t \mid d_0') \ge \Pi(m, t \mid d_0).$$

Favorable nomination evidence is deterministic certification credit; an unfavorable offset is certification debt.

A fixed-population witness makes Propositions 2–3 explicit. With $N = P = 100$, $d_0 = 0$, and $t = 0.05$, populations $(B, C, T) = (60, 40, 0)$ and $(20, 0, 80)$ have identical gain 0.2, yet certifiability is 0.154 versus 0.025 at $m = 10$ and 0.224 versus 0.365 at $m = 20$, reversing their ordering (Appendix A).

Appendix A shows that the synthetic reversal persists for $L \in \{1, 2, 4, 8, 16, 32\}$, so it is not a single-threshold or near-census artifact. Because (B, C, T) are unknown at deployment, certifiability is a design diagnostic, not an oracle policy: historical or simulated compositions can guide budgets and expose when Stage 1 evidence changes the remaining burden. It does not prescribe maximizing d_0, since acquisition also changes the unresolved population.

5 Experiments

The experiments follow C1–C3: PACS isolates the inference mechanism, a PACS-based DomainBed study tests heterogeneous replacement, and CLIP, CINIC-10, and QEC probe offset reuse, refusal, and transfer across prediction systems. Here K includes the incumbent; "harm" means full-pool $\Delta_j > 0$, whereas the guarantee controls the broader null $\Delta_j \ge 0$.

5.1 Mechanism audit

Design. On eight frozen 1,024-example PACS streams, we audit $K \in \{16, 32\}$, budgets 50–400, and 20 query seeds. A disclosed post-hoc ablation fixes the Model Selector nominee, queried labels, and budget, comparing CONFIRM's offset reuse with *adversarial-history exact*, *Naive reuse*, and *All-candidate exact*. All valid methods test the same full-pool null; their exact definitions appear in Appendices C.1 and B.2.

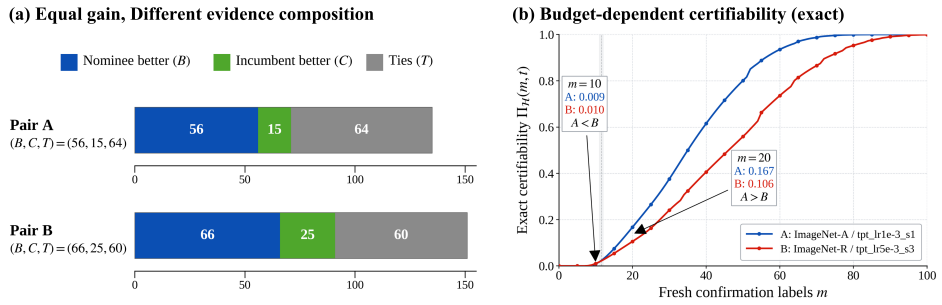

Figure 3: **Frozen-data equal-gain certifiability reversal.** (a) Frozen post-hoc CLIP pairs have equal full-pool gain ($+41/1024 = 4.00$ pp) but compositions $(B, C, T) = (56, 15, 64)$ and $(66, 25, 60)$. (b) At $t = \alpha/8 = 0.00625$, their exact certifiability order reverses between $m = 11$ and $m = 12$ (B >A at $m = 10$; A >B at $m = 20$). Exact nested-hypergeometric curves illustrate, rather than prove, the fixed-population result.

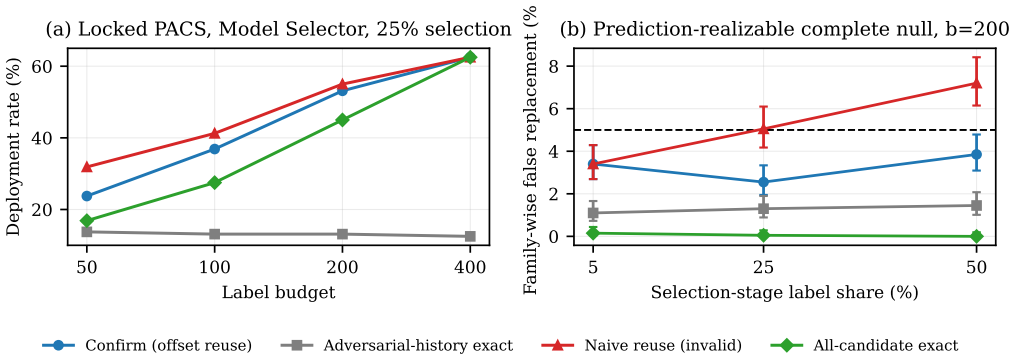

Figure 4: Matched inference ablation. Left: deployment on locked PACS with Model Selector, $K = 16$, and a 25% nomination share. Right: family-wise false replacement under a complete null; error bars are Wilson 95% intervals and the dashed line is the 5% target.

Power from the offset. For Model Selector at $K = 16$, a 25% nomination share yields 10–50 deployment points (Figure 4, left) and 1.20–3.94 accuracy points over adversarial-history exact. At the frozen 5% share (2–20 Stage 1 labels), maximum gains are 3.125 and approximately 0.0824 points, respectively (Appendix C.1, Table 5). Power depends on nominee-relevant history; CLIP provides a stronger 5%-share case (Section 5.3).

Null calibration. Across 2,000 prediction-realizable complete-null families, Naive reuse rejects at 7.20% (Wilson 95% interval $[6.15, 8.42]$ (Wilson, 1927)) against a 5% target, versus 3.85% $[3.09, 4.79]$ for CONFIRM; every candidate ties the incumbent under one jointly realizable label vector. The latter is consistent with Remark 1; a selected-unit diagnostic appears in Appendix D.2.

5.2 Main heterogeneous benchmark

Design. Our predeclared prediction-blind PACS-based DomainBed study (Gulrajani & Lopez-Paz, 2021) uses eight ERM checkpoints on eight 1,024-example streams (64 jointly Holm-controlled units). The $K = 16$ heterogeneous TTA portfolio and $K = 32$ cross-checkpoint extension were frozen before candidate accuracy was computed. Because PACS labels entered the earlier audit, the study is prediction-blind, not label-blind (Appendix D.1).

Safety and recovery. Across 174,080 repeated gated decisions (136 cells, each with 20 query repetitions on 64 fixed pools) we observe no harm, versus 21.9–84.1% harmful ungated decisions; the 84.1% endpoint is the $K = 32, b = 100$ random control in Appendix D.1. At $K = 16, b = 400$,

Table 1: Predeclared prediction-blind PACS-based DomainBed study at $K = 16$ (1,280 decisions per cell). All gated harm counts are zero; bold marks the descriptive maximum among non-oracle selectors. "Oracle" is the full-pool-best *non-incumbent* Stage-1 survivor, so it can still be worse than the incumbent. $K = 32$, cluster-uncertainty, and missed-gain diagnostics appear in Appendix D.

	+CONFIRM: accuracy↑ (all harm 0)				Selector alone: accuracy↑ (harm↓)			
Stage-1 selector	b=50	b=100	b=200	b=400	b=50	b=100	b=200	b=400
Retain incumbent	88.985	88.985	88.985	88.985	88.985 (0)	88.985 (0)	88.985 (0)	88.985 (0)
SHR (hypergraph)	89.065	89.292	89.590	89.711	89.581 (561)	89.825 (469)	89.944 (425)	89.965 (420)
Model Selector	89.160	89.385	89.451	89.739	89.655 (518)	89.858 (428)	89.959 (420)	89.966 (420)
Active Testing-LURE	89.105	89.338	89.530	89.629	89.336 (634)	89.632 (551)	89.836 (488)	89.928 (445)
PPAT	89.212	89.371	89.492	89.563	89.075 (704)	89.322 (640)	89.610 (560)	89.812 (486)
HILTTA-FP	89.164	**89.576**	**89.839**	**89.867**	88.843 (820)	88.824 (880)	88.824 (880)	88.797 (880)
TTALine ALine-D	**89.250**	89.407	89.598	89.601	88.239 (680)	88.239 (680)	88.239 (680)	88.239 (680)
Uniform labels	89.079	89.273	89.468	89.528	88.787 (758)	89.213 (667)	89.585 (562)	89.806 (493)
Random nominee	89.084	89.197	89.356	89.361	80.905 (1042)	79.011 (1065)	81.699 (1036)	81.067 (1060)
Oracle non-source survivor	89.557	90.009	90.085	90.085	89.966 (420)	89.966 (420)	89.966 (420)	89.966 (420)

Table 2: Disclosed matched valid-gate comparison on seven unaffected CLIP datasets, aggregated equally over four budgets, seven fixed datasets, and 20 query repetitions (560 repeated decisions). Useful is authorization conditional on a beneficial nominee; missed is mean forfeited positive gain per unit, in accuracy points. False includes full-pool ties; "Family false" is the number of budget–repetition families with at least one false replacement, out of 80. The five gate rows share 131.9 labels, mean fresh fraction $m/P = 68.5\%$, and 44.1% confirmation censuses.

Method	Acc.	Δsrc	Deploy (%)	Useful (%)	Missed (pts)	Harm/false	Family false	Labels
Retain incumbent	60.435	0.000	0.0	0.0	4.653	0/0	0/80	0.0
Stage-1 nominee, no gate	65.050	+4.614	100.0	100.0	0.000	48/65	44/80	8.3
Adversarial-history exact	64.539	+4.103	61.8	69.9	0.549	0/0	0/80	131.9
$d_{\mathrm{abl}} = \min(d_0, 0)$	64.622	+4.187	66.2	74.9	0.466	0/0	0/80	131.9
Bonferroni marginal bounds	64.358	+3.922	63.9	72.3	0.731	0/0	0/80	131.9
Unconditional paired exact	64.311	+3.876	62.7	70.9	0.777	0/0	0/80	131.9
CONFIRM	**64.635**	**+4.199**	**67.1**	**76.0**	**0.453**	0/0	0/80	131.9

SHR recovers 66.0% of source-to-oracle headroom, leaving 0.374 accuracy points unrealized, while HILTTA-FP improves from 88.797 to 89.867 accuracy and eliminates 880 harms (Table 1). These empirical zeros are diagnostic, not the source of validity (Theorem 1); small selector gaps are descriptive because cluster-resampled contrasts are unresolved in most cells. Additional diagnostics appear in Appendices D.1 and D.2.

5.3 SEALED FOUNDATION-MODEL AUDIT

Before label reveal, we froze a CLIP ViT-B/16 incumbent (Radford et al., 2021) and 16 adaptation candidates (Shu et al., 2022; Samadh et al., 2023; Karmanov et al., 2024; Farina et al., 2024) across eight 1,024-example datasets. A post-reveal PromptAlign class-order defect excludes ImageNet-V2, while the remaining seven retain the conservative $0.05/8$ threshold. A disclosed frozen-prediction comparison fixes the nominee, Stage 1 history, Stage 2 draw, realized label count, and error allocation for every valid gate (Table 2). Relative to a Bonferroni combination of exact marginal hypergeometric bounds and a direct unconditional paired exact gate, CONFIRM gains 0.277/0.324 accuracy points and 3.64/5.05 points of beneficial authorization, respectively. All three make zero observed false replacements. At $b = 50$, the gains are 1.090/1.069 accuracy points and 14.7/14.7 useful-authorization points; five of seven dataset means are positive and two tie, but the dataset-cluster 95% t intervals include zero ($[-0.30, 2.48]$/$[-0.34, 2.48]$). Repetitions thus measure query-draw variability, not 560 independent domains.

The advantage is not a census artifact. Among 313 partial confirmations ($m < P$), CONFIRM gains 0.496/0.579 accuracy points and 6.45/8.96 useful-authorization points over the two comparators; among 247 censuses all three decisions coincide. Thus this audit supports a modest low-budget matched-validity advantage, not universal accuracy superiority (Appendix E).

5.4 SEALED REFUSAL TEST

Because PACS labels entered development, we separately sealed a label-blind CINIC-10 audit (Darlow et al., 2018), removing exact CIFAR-10 pixel matches (Krizhevsky, 2009) before freezing 32 disjoint 1,024-example units, 15 candidates, budgets, seeds, and the Holm family. The incumbent is uniquely optimal on 31 units; one Tent candidate wins the remaining unit by one error. At $b = 400$, all five gated selectors avoid harm, and Model Selector+CONFIRM replaces only that unit in all 20 runs, versus 46 and 38 harms/640 for ungated PPAT and Uniform. This is candidate scarcity, not confirmation failure (Appendix F).

5.5 HETEROGENEOUS-DECODER GENERALITY AUDIT

On the public Willow archive (Sivak et al., 2026), a disclosed post-hoc all-source rotation fixed before alternative-source gate results evaluates Sparse Blossom, Tesseract, and AlphaQubit-family predictions (Higgott & Gidney, 2025; Aghababaie Beni et al., 2025; Bausch et al., 2024). Across 84 replaceable and 28 non-replaceable tasks, ungated selection makes 1,402–1,660 harms/5,600 decisions; CONFIRM makes none and never replaces a non-replaceable incumbent. Authorization rises from 32.6% at 250 labels to 96.8% at 4,000, with 10.52% error reduction versus 10.65% for the oracle. This supports transfer across prediction systems, not future-device safety (Appendices G.1 and G.2).

6 WHAT THE EVIDENCE ESTABLISHES

C1–C2: nomination proposes; conditional inference authorizes. Ungated nomination causes harm across audits, including refusal-dominated CINIC-10. Proposition 1 and the complete-null audit show why nomination labels cannot generally be treated as fresh test data. Theorem 1 establishes conditional validity through a fixed history offset and fresh uniform sampling from unresolved disagreements. In CLIP, capping favorable d_0 at the 5% share changes accuracy by only 0.013 points; comparison with a direct paired gate sharing the offset and outcomes isolates conditional inference as the source of low-budget gain. PACS at 25% shows when offset can matter.

C3: certifiability is evidence- and budget-specific. Theorem 2 and Propositions 2–3 show that global gain does not determine finite-budget certifiability. Figure 3 and Supplementary Figure 5 relate authorization to evidence composition, Stage 1 history, and budget; Corollary 1 extends the profile to deployment families.

7 LIMITATIONS AND CONCLUSION

Limitations. Predeclared DomainBed is prediction-blind and CINIC-10 label-blind; the PACS mechanism audit, frozen-data pair, and QEC are post-hoc, while CLIP comparisons are post-label. CLIP gains concentrate at low budgets and partial confirmation; wide seven-dataset cluster intervals preclude broad superiority claims. Repetitions are not independent populations. The ternary construction targets 0–1 classification; other losses require different tests. Finite-population benefits can diminish at small sampling fractions. At high label fractions, supervised updating is an alternative; CONFIRM targets settings where updating is costly, disallowed, or separate from authorization. AETTA/MORPHEUS nomination and TACT candidates remain future evaluations. Validity covers only the enumerated pool and frozen predictions; updates require new decisions, with no claim of future-drift safety or universal no-harm.

Conclusion. CONFIRM formulates finite-pool model replacement as an authorization problem distinct from adaptive nomination. Its central contribution is a post-nomination conditioning construction that turns nominee-relevant Stage 1 evidence into a fixed offset and obtains fresh randomness only from unresolved disagreements, enabling selector-agnostic finite-sample-valid inference under the stated sampling design. We further characterize finite-budget certifiability, showing that it depends on evidence composition, nomination history, and remaining label budget rather than accuracy gain alone. Together, these results provide an authorization procedure for determining whether an adaptively nominated model has sufficient incumbent-relative evidence for replacement.

AI USE STATEMENT

We acknowledge the use of Large Language Models (LLMs) to assist in the preparation of this manuscript. Specifically, LLMs assisted with drafting and revising text, literature discovery, experimental code development and debugging, and organizing and analyzing results. The authors carefully reviewed and edited all AI-assisted outputs and take full responsibility for the final content and scientific integrity of this work.

ETHICS STATEMENT

The study uses public image and quantum-device data, with no human subjects or private information. We disclose the PACS mirror's unknown license and do not treat repository timestamps as acquisition dates. The guarantee is not assurance for arbitrary drift or autonomous safety-critical use.

REPRODUCIBILITY STATEMENT

The appendices give proofs, protocols, baseline scope, seeds, and disclosed implementation deviations and valid-gate comparisons; their experiment-specific effects are localized where reported and do not alter the formal validity result in Theorem 1. Appendix H summarizes reproducibility structure and the corrected boundary-case implementation.

REFERENCES

Laleh Aghababaie Beni, Oscar Higgott, and Noah Shutty. Tesseract: A search-based decoder for quantum error correction. *arXiv preprint arXiv:2503.10988*, 2025. URL https://arxiv.org/abs/2503.10988.

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, 2025. doi: 10.1214/24-AOAS1998.

Kianoosh Ashouritaklimi, Valentin Kilian, Daolang Huang, Tom Rainforth, and François Caron. Prediction-powered active testing. *arXiv preprint arXiv:2607.08347*, 2026. URL https://arxiv.org/abs/2607.08347.

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

Johannes Bausch, Andrew W. Senior, Francisco J. H. Heras, et al. Learning high-accuracy error decoding for quantum processors. *Nature*, 635:834–840, 2024. doi: 10.1038/s41586-024-08148-8.

Roger L. Berger and Dennis D. Boos. p values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, 89(427):1012–1016, 1994. doi: 10.1080/01621459.1994.10476836.

Roger L. Berger and K. Sidik. Exact unconditional tests for a 2×2 matched-pairs design. *Statistical Methods in Medical Research*, 12(2):91–108, 2003. doi: 10.1191/0962280203sm312ra.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101: Mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. URL https://data.vision.ee.ethz.ch/cvl/datasets_extra/food-101/.

Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8344–8353, 2022.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. URL https://www.robots.ox.ac.uk/~vgg/data/dtd/.

Sebastian Cygert, Damian Sójka, Tomasz Trzciński, and Bartłomiej Twardowski. Realistic evaluation of test-time adaptation: Unsupervised model selection. In *International Conference on Computer Vision Theory and Applications*, pp. 75–86, 2026. doi: 10.5220/0014320100004084.

Michal Danilowski, Alexander Murphy, Young D. Kwon, and Abhirup Ghosh. MORPHEUS: Meta test-time adaptation via neural collapse geometry. In *ICLR Workshop on Test-Time Updates*, 2026. URL https://openreview.net/forum?id=YfyApJH1Rj.

Luke N. Darlow, Elliot J. Crowley, Antreas Antoniou, and Amos J. Storkey. CINIC-10 is not ImageNet or CIFAR-10. Technical Report EDI-INF-ANC-1802, University of Edinburgh, 2018. URL https://arxiv.org/abs/1810.03505.

Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. Frustratingly easy test-time adaptation of vision-language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4099.

Amirmohammad Farzaneh and Osvaldo Simeone. Post-selection distributional model evaluation. *arXiv preprint arXiv:2603.23055*, 2026. URL https://arxiv.org/abs/2603.23055.

William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

Tobias Freidling, Qingyuan Zhao, and Zijun Gao. Selective randomization inference for adaptive experiments. *arXiv preprint arXiv:2405.07026*, 2024. URL https://arxiv.org/abs/2405.07026.

Saurabh Garg, Sivaraman Balakrishnan, Zachary C. Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *International Conference on Learning Representations*, 2022.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019. URL https://github.com/phelber/EuroSAT.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *IEEE/CVF International Conference on Computer Vision*, 2021a. URL https://github.com/hendrycks/imagenet-r.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021b. URL https://github.com/hendrycks/natural-adv-examples.

Oscar Higgott and Craig Gidney. Sparse blossom: Correcting a million errors per core second with minimum-weight matching. *Quantum*, 9:1600, 2025. doi: 10.22331/q-2025-01-20-1600.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 2427–2440, 2021.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14162–14171, 2024.

Justin Kay, Grant Van Horn, Subhransu Maji, Daniel Sheldon, and Sara Beery. Consensus-driven active model selection. In *IEEE/CVF International Conference on Computer Vision*, pp. 4594–4604, 2025.

Eungyeup Kim, Mingjie Sun, Christina Baek, Aditi Raghunathan, and J. Zico Kolter. Test-time adaptation induces stronger accuracy and agreement-on-the-line. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *International Conference on Machine Learning*, pp. 5753–5763, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Arun K. Kuchibhotla, John E. Kolassa, and Todd A. Kuffner. Post-selection inference. *Annual Review of Statistics and Its Application*, 9:505–527, 2022. doi: 10.1146/annurev-statistics-100421-044639.

John M. Lachin. Power and sample size evaluation for the McNemar test with application to matched case-control studies. *Statistics in Medicine*, 11(9):1239–1251, 1992. doi: 10.1002/sim.4780110909.

Bracha Laufer-Goldshtein, Adam Fisch, Regina Barzilay, and Tommi Jaakkola. Efficiently controlling multiple risks with Pareto Testing. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=cyg2YXn_BqF.

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop on Challenges in Representation Learning*, 2013.

Taeckyung Lee, Sorn Chottananurak, Taesik Gong, and Sung-Ju Lee. AETTA: Label-free accuracy estimation for test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28643–28652, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Lee_AETTA_Label-Free_Accuracy_Estimation_for_Test-Time_Adaptation_CVPR_2024_paper.html.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *International Conference on Computer Vision*, pp. 5542–5550, 2017a.

Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. In *ICLR Workshop*, 2017b. URL https://arxiv.org/abs/1603.04779.

Yushu Li, Yongyi Su, Xulei Yang, Kui Jia, and Xun Xu. Exploring human-in-the-loop test-time adaptation by synergizing active learning and model selection. *Transactions on Machine Learning Research*, 2024.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6028–6039, 2020. URL https://proceedings.mlr.press/v119/liang20a.html.

Yingnan Liu, Rui Qiao, Mong Li Lee, and Wynne Hsu. Test-time adaptation by causal trimming. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2510.11133.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Ramin Nikzad-Langerodi and Valeria Fonseca Diaz. Transferability loss for safe model selection under domain shift. In *Workshop on Catch, Adapt, and Operate: Monitoring ML Models Under Drift (CAO) at ICLR*, 2026. URL https://openreview.net/forum?id=09D1j9dCrZ.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International Conference on Machine Learning*, pp. 16888–16905, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations*, 2023.

Patrik Okanovic, Andreas Kirsch, Jannes Kasper, Torsten Hoefler, Andreas Krause, and Nezihe Merve Gürel. All models are wrong, some are useful: Model selection with limited labels. In *International Conference on Artificial Intelligence and Statistics*, pp. 2035–2043, 2025.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. URL https://www.robots.ox.ac.uk/~vgg/data/pets/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400, 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

Ants Remm, Nathan Lacroix, Lukas Bödeker, Elie Genois, Christoph Hellings, François Swiadek, Graham J. Norris, Christopher Eichler, Alexandre Blais, Markus Müller, Sebastian Krinner, and Andreas Wallraff. Experimentally informed decoding of stabilizer codes based on syndrome correlations. *Physical Review Research*, 8:013044, 2026. doi: 10.1103/z1ng-wg3k.

Pascal Rink and Werner Brannath. Post-selection confidence bounds for prediction performance. *Machine Learning*, 114(3), 2025. doi: 10.1007/s10994-024-06632-w. Article 82.

Jameel Abdul Samadh, Mohammad Hanan Gani, Noor Hussein, Muhammad Uzair Khattak, Muhammad Muzammal Naseer, Fahad Shahbaz Khan, and Salman H. Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-3525.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1038.

Volodymyr Sivak, Alexis Morvan, Michael Broughton, et al. Reinforcement learning control of quantum error correction. *Nature*, 655:879–884, 2026. doi: 10.1038/s41586-026-10759-2.

Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *International Conference on Machine Learning*, pp. 2380–2388, 2015.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

Haohan Wang, Songwei Ge, Zachary C. Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://github.com/HaohanWang/ImageNet-Sketch.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.

Xu Yang, Yanan Gu, Kun Wei, and Cheng Deng. Exploring safety supervision for continual test-time domain adaptation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 1649–1657, 2023. doi: 10.24963/ijcai.2023/183.

Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15922–15932, 2023. doi: 10.1109/CVPR52729.2023.01528.

Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. In *International Conference on Machine Learning*, pp. 42058–42080, 2023.

SUPPLEMENTARY OVERVIEW

Appendix guide. Appendix A gives all proofs; Appendix B specifies the algorithm and baseline scope; Appendix C contains the PACS mechanism, robustness, and post-hoc stress diagnostics; Appendix D reports the predeclared DomainBed study and gate-power diagnostics; Appendices E and F give the sealed CLIP and label-blind CINIC-10 audits; Appendix G gives the QEC generality and deviation audits; and Appendix H summarizes reproducibility and implementation notes.

Table 3: Closest formulations and their inferential targets.

Method	Selection	Inferential target	Evidence / guarantee
Exact paired tests	Prespecified	Paired difference	Exact per-test inference
Data carving	Adaptive	Selected parameter	Conditional reuse; selective Type-I
Selective RI	Adaptive	Selected effect	Conditional randomization inference
Post-selection bounds	Adaptive	Selected-model performance	Reuse; simultaneous coverage
Pareto Testing	Split optimize/test	Configuration risk constraints	Held-out tests; finite-sample control
PS-DME	Adaptive	KPI distribution	Full-sample reuse; FCR
TR-LOSS	Rule-specific	Source-relative bound	Unlabeled worst-case
Finite-population LCB	Post-nomination	Incumbent-relative gain	Separate marginal limits; finite-sample
CONFIRM	**Adaptive (fixed family)**	**Incumbent-relative gain**	**Offset + fresh draw; exact FWER**

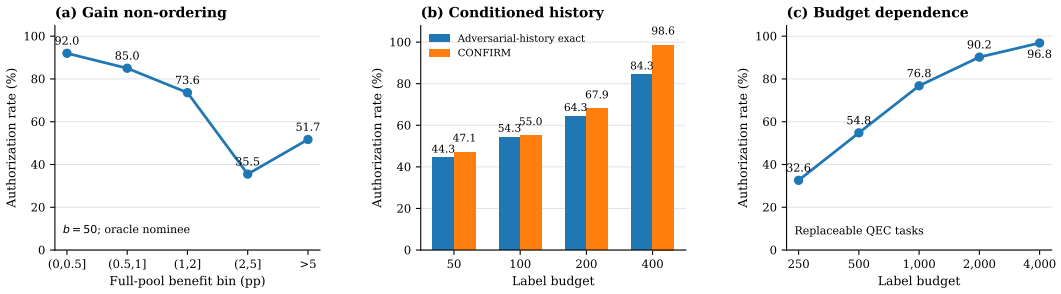

Figure 5: **Empirical factors governing finite-budget certifiability.** (a) In a disclosed post-hoc DomainBed oracle-nominee diagnostic at $b = 50$, authorization is not ordered by full-pool benefit bins. (b) On the seven unaffected frozen CLIP datasets, CONFIRM and adversarial-history exact share the nominee, history, fresh draw, labels, and threshold; conditioning on the realized history yields additional authorizations, especially at the largest budget. (c) On replaceable QEC tasks, authorization rises with the confirmation budget. These descriptive summaries illustrate the composition, history, and budget factors in Theorem 2.

A COMPLETE PROOFS

A.1 PROOF OF THEOREM 1

For a family of L units let $\mathbf{H} = (\mathcal{H}_1, \ldots, \mathcal{H}_L)$ collect every stage-1 history, and condition jointly on $\mathbf{H}$. This fixes each nomination and the set $\mathcal{T}_0 = \{\ell : \Delta_{j_\ell^\star} \geq 0\}$ of true deployment nulls.

The sampling design requires each unit's stage-2 sample to remain uniform without replacement conditional on $\mathbf{H}$; separate preassigned random seeds satisfy this condition, while independence among units is unnecessary.

Fix $\ell \in \mathcal{T}_0$. Its explored offset d_0 and unresolved population are now fixed. Let the unresolved population contain B nominee wins, C incumbent wins, and $P - B - C$ ties, and write $M = B + C$. The true null gives $d_0 + 2B - M \le 0$. For a realized confirmation outcome (W_m, L_m, T_m) with $n_m = W_m + L_m$, the actual M satisfies

$$n_m \le M \le P - T_m, \qquad B \le \left\lfloor \frac{M - d_0}{2} \right\rfloor, \qquad B \le M - L_m.$$

Because also $B \ge W_m$, the actual M belongs to the compatible set $\mathcal{M}$ in Section 4.2, and $B \le B_M^{(0)}$. Conditional on n_m,

$$W_m \sim \mathrm{Hypergeom}(M, B, n_m).$$

The hypergeometric upper tail is nondecreasing in its population success count. Therefore, pointwise in the realized outcome,

$$p \ge \Pr\Big\{\mathrm{Hypergeom}(M, B_M^{(0)}, n_m) \ge W_m\Big\} \ge \Pr\{\mathrm{Hypergeom}(M, B, n_m) \ge W_m\}.$$

The last quantity is the ordinary exact upper-tail p-value under the fixed true population and is super-uniform conditional on n_m. Since this finite-population test is discrete and the reported p-value further maximizes over compatible nuisance populations, the inequality can be strict, as noted in Remark 1. Hence $\Pr(p_\ell \le t \mid \mathbf{H}) \le t$ after averaging over n_m.

Conditional on $\mathbf{H}$, the true-null set is fixed and every true-null p-value is marginally super-uniform. Holm's step-down argument therefore controls the conditional FWER under arbitrary dependence (Holm, 1979); averaging over $\mathbf{H}$ gives the unconditional claim.

Subject to the fixed-prediction setup, nothing in the argument restricts how $\mathcal{H}$ was generated. In particular, an acquisition rule may deliberately query atypical indices or choose the nominee using any function of the revealed labels and internal randomization. Those labels enter as the exact offset and never as a representative sample, so adaptive acquisition changes the conditioning set, not the sampling law of the fresh confirmation draw. Exhaustive small-population repository tests enumerate every compatible population and sample outcome, and confirm super-uniformity numerically for arbitrary acquisition offsets.

A.2 Proof of Proposition 1

Construct four frozen candidates with pairwise disjoint disagreement sets $D_1, \dots, D_4$, each of size 10. On every D_j, exactly five points are candidate wins and five are incumbent wins, with no ties; outside D_j, candidate j agrees with the incumbent. Thus

$$\Delta_j = 0, \qquad j = 1, \dots, 4,$$

so every candidate lies on the replacement null boundary. The construction is jointly realizable by one fixed label vector.

In Stage 1, sample three points uniformly without replacement from each D_j, independently across the four disjoint sets, and let W_j be the number of candidate wins. Nominate a candidate attaining $\max_j W_j$, with any deterministic tie rule. For each candidate,

$$\Pr(W_j = 3) = \frac{\binom{5}{3}}{\binom{10}{3}} = \frac{1}{12}.$$

Hence

$$\Pr\left(\max_j W_j = 3\right) = 1 - \left(\frac{11}{12}\right)^4.$$

After nomination, draw one additional point uniformly from the seven unqueried disagreements of the selected candidate. Conditional on its Stage 1 outcome being three candidate wins, two candidate-win points and five incumbent-win points remain, so the probability that the fresh observation is also a candidate win is $2/7$.

Now consider a naive pooled analysis that treats the selected candidate's three Stage 1 observations and one Stage 2 observation as if they formed a uniform sample of four points from its ten disagreements. Four candidate wins give the exact one-sided finite-population value

$$p_{\text{pool}} = \frac{\binom{5}{4}}{\binom{10}{4}} = \frac{1}{42} < 0.05,$$

whereas three candidate wins give $p_{\text{pool}} = 11/42 > 0.05$. Therefore the naive procedure rejects exactly when the selected candidate has three Stage 1 wins and the fresh Stage 2 point is also a win. Its true complete-null rejection probability is

$$\left[1 - \left(\frac{11}{12}\right)^4\right] \frac{2}{7} = \frac{6095}{72576} \approx 0.08398 > 0.05.$$

Thus an exact hypergeometric calculation does not restore validity when selection-dependent Stage 1 outcomes are incorrectly treated as fresh random test observations.

A.3 Proof of Theorem 2

Conditional on $\mathcal{H}$, the unresolved population and its composition (B, C, T) are fixed. A uniform sample without replacement of size m has multivariate hypergeometric mass

$$\Pr(W_m = w, L_m = l, T_m = r \mid \mathcal{H}) = \frac{\binom{B}{w}\binom{C}{l}\binom{T}{r}}{\binom{P}{m}}$$

for nonnegative (w, l, r) summing to m, and zero otherwise. Summing this mass over the rejection region $\{(w, l, r) : p_{\mathcal{H}}(w, l, r) \le t\}$ gives Eq. (5).

For the equivalent two-stage representation, set $N_m = W_m + L_m$. Choosing which sampled points are informative gives

$$N_m \sim \text{Hypergeom}(P, B + C, m).$$

Conditional on $N_m = n$, the informative observations are a uniform without-replacement sample of size n from the $B + C$ informative points, so

$$W_m \mid N_m = n \sim \text{Hypergeom}(B + C, B, n).$$

Equivalently, the factorization follows from

$$\frac{\binom{B}{w}\binom{C}{n-w}\binom{T}{m-n}}{\binom{P}{m}} = \frac{\binom{B+C}{n}\binom{T}{m-n}}{\binom{P}{m}} \frac{\binom{B}{w}\binom{C}{n-w}}{\binom{B+C}{n}}.$$

A.4 Proof of Corollary 1

Let $p_{(1)} \le \cdots \le p_{(L)}$ denote the ordered confirmation p-values, and suppose unit ℓ has rank r. If $p_\ell \le \alpha/L$, then for every $k \le r$,

$$p_{(k)} \le p_\ell \le \frac{\alpha}{L} \le \frac{\alpha}{L - k + 1}.$$

Hence Holm reaches rank r and rejects unit ℓ. Conversely, if Holm rejects unit ℓ, then

$$p_\ell = p_{(r)} \le \frac{\alpha}{L - r + 1} \le \alpha.$$

Therefore, pointwise in the realized family of p-values,

$$\{p_\ell \le \alpha/L\} \subseteq \{\ell \text{ is rejected by Holm}\} \subseteq \{p_\ell \le \alpha\}.$$

Taking conditional probabilities given $\mathbf{H}$ yields Eq. (6). No independence assumption is used.

A.5 Proof of Proposition 2

It suffices to construct two populations. Take $N = P = 100$, $d_0 = 0$, and

$$(B, C, T) = (60, 40, 0) \quad \text{and} \quad (B, C, T) = (20, 0, 80).$$

Both have full-pool gain $\delta = (d_0 + B - C)/N = 0.2$. At threshold $t = 0.05$ and budget $m = 10$, exhaustive evaluation of Eq. (5) gives respectively

$$\Pi_1(10, 0.05) = 0.1538068, \qquad \Pi_2(10, 0.05) = 0.0254645.$$

Thus equal N, P, d_0, and positive gain do not determine the certifiability profile.

A.6 Proof of Proposition 3

Use the same two populations as in the preceding proof. At $t = 0.05$, exact enumeration gives

$$\Pi_1(10, 0.05) = 0.1538068 > 0.0254645 = \Pi_2(10, 0.05),$$

whereas

$$\Pi_1(20, 0.05) = 0.2236791 < 0.3646830 = \Pi_2(20, 0.05).$$

Hence their certifiability ordering reverses between $m_1 = 10$ and $m_2 = 20$, despite identical positive full-pool gain.

The conclusion is robust to the multiplicity threshold in a controlled exact calculation. For $L \in \{1, 2, 4, 8, 16, 32\}$ and $t = 0.05/L$, Population 1 is more certifiable at $m = 10$, while Population 2 is more certifiable at $m = 25$. Both budgets are far from the $P = 100$ census. Discreteness can create short intermediate plateaus or counter-reversals, but the stated ordering at the two fixed budgets holds for every tested family size.

A.7 Proof of Proposition 4

Fix the unresolved population and a confirmation outcome (w, l, r). For each candidate nuisance value M, write

$$B_M^{(0)}(d) = \min\left\{ M - l, \left\lfloor \frac{M - d}{2} \right\rfloor \right\}.$$

If $d_0' \geq d_0$, then $B_M^{(0)}(d_0') \leq B_M^{(0)}(d_0)$ for every M. Moreover, the compatibility condition $B_M^{(0)}(d) \geq w$ implies

$$\mathcal{M}(d_0') \subseteq \mathcal{M}(d_0).$$

For every $M \in \mathcal{M}(d_0')$, stochastic monotonicity of the hypergeometric distribution in its success count gives

$$\Pr\left\{ \mathrm{Hypergeom}(M, B_M^{(0)}(d_0'), w + l) \geq w \right\} \leq \Pr\left\{ \mathrm{Hypergeom}(M, B_M^{(0)}(d_0), w + l) \geq w \right\}.$$

Taking maxima, first over the smaller compatible set and then over the larger one, yields $p_{d_0'}(w, l, r) \leq p_{d_0}(w, l, r)$. The conclusion also holds when $\mathcal{M}(d_0')$ is empty because the convention is then $p_{d_0'} = 0$.

Finally, the confirmation-outcome distribution for a fixed unresolved population does not depend on the conditioned offset. The rejection indicator is pointwise nondecreasing in d_0, so taking its expectation gives

$$\Pi(m, t \mid d_0') \geq \Pi(m, t \mid d_0).$$

B Algorithm and Baseline Implementation

B.1 Algorithm and Implementation Details

For a predeclared multi-unit family, run Algorithm 1 once per unit, apply Holm to the returned p-values at family level α, and replace only rejected unit nulls.

Algorithm 1 CONFIRM with an arbitrary adaptive selector $\mathcal{A}$ over fixed predictions for one deployment unit

Require: frozen predictions of $h_0, \dots, h_{K-1}$ on N points; budget b; nomination share ρ; level α; selector $\mathcal{A}$

1: **for** $j = 1, \dots, K-1$: $D_j \leftarrow \{i : h_j(x_i) \neq h_0(x_i)\}$; $U \leftarrow \bigcup_{j=1}^{K-1} D_j$ ▷ no labels
2: $b_{\text{real}} \leftarrow \min(b, |U|)$; $b_{\text{nom}} \leftarrow \text{round}(\rho b_{\text{real}})$
3: $(Q, j^\star) \leftarrow \mathcal{A}(b_{\text{nom}})$ ▷ $|Q| \le b_{\text{nom}}$
4: $W_0, L_0 \leftarrow$ nominee and incumbent wins in $Q \cap D_{j^\star}$
5: $R \leftarrow D_{j^\star} \setminus Q$; $m \leftarrow \min\{b_{\text{real}} - |Q|, |R|\}$
6: sample m indices uniformly without replacement from R
7: compute p by Eq. (2)
8: **return** $(j^\star, p)$

Stage-1 front end and overlap accounting. Our benchmark uses SHR as one adaptive nomination front end for exercising the selector-agnostic confirmation layer. For $i \in U$, let $\mathcal{C}_i = \{j : i \in D_j\}$. SHR partitions U by common incidence signatures, estimates incumbent-relative gaps, queries the stratum expected to reduce uncertainty most near the current leader, and nominates the smallest estimated gap. Its nomination share $\rho = 0.05$ and one-nominee rule were fixed on development data before PACS labels were opened; the safety result does not depend on this front end. Overlap only affects realized label reuse: for nonempty Q, if $n_j(Q) = |Q \cap D_j|$ and $r(Q) = |Q|^{-1} \sum_{i \in Q} |\mathcal{C}_i|$, then $\sum_j n_j(Q) = r(Q)|Q|$. We use this identity descriptively, not as a sample-complexity claim.

Worked exact-test example. Suppose Stage 1 queries 8 of 120 disagreements and observes $(W_0, L_0) = (2, 3)$, so $d_0 = -1$. Stage 2 samples 20 of the remaining $P = 112$ and observes $(W_m, L_m, T_m) = (12, 6, 2)$, so $n_m = 18$ and $18 \le M \le 110$. Maximizing with $B_M^{(0)} = \min\{M - 6, \lfloor (M+1)/2 \rfloor\}$ gives $p = 0.106$. Historical evidence shifts the null but is not modeled as a random sample.

The SHR acquisition score. For stratum s, let N_s be its size, n_s its queried count, and $\widehat{v}_{sj}$ candidate j's paired-outcome variance, floored at 0.05 (and set to 1 when $n_s \le 1$). The current width and one-query reduction are

$$w_j = \sum_{s : j \in \mathcal{C}_s} \frac{N_s}{N} \sqrt{\frac{\widehat{v}_{sj}}{n_s + 1}}, \qquad \eta_{sj} = \frac{N_s}{N} \sqrt{\widehat{v}_{sj}} \left(\frac{1}{\sqrt{n_s + 1}} - \frac{1}{\sqrt{n_s + 2}} \right).$$

With τ the median positive w_j, SHR scores stratum s by

$$\sum_{j \in \mathcal{C}_s} \left(0.25 + \exp[-(\widehat{\Delta}_j - \min_\ell \widehat{\Delta}_\ell)/\tau]\right) \max(w_j, 10^{-8}) \eta_{sj}.$$

The next index in the highest-scoring stratum comes from a seed-fixed random permutation.

When the observable disagreement union contains fewer than b examples, the realized budget is capped at $|U|$ before applying the 5% split. This is why the largest PACS budgets use between 9 and 20 acquisition labels rather than always 20. Within a chosen stratum the next index is drawn from a seed-fixed random permutation, making each Stage 1 acquisition reproducible from its seed. The validity result imposes no representativeness requirement on Stage 1 acquisition; uniform without-replacement sampling is required only for the Stage 2 confirmation draw. Every run has exactly one survivor. Nominal budgets 50 and 100 use 2 and 5 acquisition labels; budgets 200 and 400 use 9 to 10 and 9 to 20.

Acquisition-fraction sensitivity. Sweeping the acquisition share offline on the frozen predictions gives, at $b = 100$ and $K = 16$, accuracies 57.216, 56.984, 57.098 and 56.269 for $\rho = 0.05, 0.10, 0.20, 0.40$, with deployment rates 44.4%, 43.1%, 43.8% and 35.6%; at $b = 400$ the same sweep gives 58.154, 58.155, 58.192 and 58.204 at a constant 62.5% deployment. Every row has zero observed harm, consistent with Theorem 1 permitting arbitrary acquisition. A multi-survivor prototype ($b = 100$ accuracies 57.216, 57.217, 57.230 for one, two and three survivors;

$b = 400$ accuracies 58.154, 58.156, 58.156) is reported only here because its frozen implementation deploys the lowest-index survivor instead of performing a fair second-stage comparison, and supports no novelty claim.

B.2 Baseline Implementation Scope

Matched authorization baselines. All matched gates below test the original full-pool null $H_{j^\star}$: $\Delta_{j^\star} \geq 0$, use the same nominee and Stage 1 history, and draw the same m labels uniformly from $R = D_{j^\star} \setminus Q$. They differ only in how they use the fixed Stage 1 outcomes and the fresh paired counts. The method stored under the historical implementation key `split_exact` is called *adversarial-history exact* here: it discards the values observed in $Q \cap D_{j^\star}$ but not the fact that those indices were removed from R, and evaluates Eq. (2) with $(W_0, L_0) = (0, |Q \cap D_{j^\star}|)$. Assigning every discarded outcome to the incumbent is worst-case for the full-pool null, hence valid but potentially conservative; it is not an independent test split. The historical-credit ablation instead sets

$$d_{\mathrm{abl}} = \min(d_0, 0)$$

in Eq. (2). It preserves unfavorable historical evidence and removes only favorable credit. Proposition 4 implies its p-value is no smaller than CONFIRM's on the same draw.

The *Bonferroni marginal bounds* comparator uses the same observed offset d_0. Let $X_b \sim \mathrm{Hypergeom}(P, b, m)$ and $Y_c \sim \mathrm{Hypergeom}(P, c, m)$. Its exact one-sided inversions are

$$\underline{B}_t = \begin{cases} 0, & W_m = 0, \\ \min\{b \in [W_m, P - m + W_m]_{\mathbb{Z}} : \Pr(X_b \geq W_m) > t/2\}, & W_m > 0, \end{cases}$$

and

$$\overline{C}_t = \begin{cases} P, & L_m = m, \\ \max\{c \in [L_m, P - m + L_m]_{\mathbb{Z}} : \Pr(Y_c \leq L_m) > t/2\}, & L_m < m. \end{cases}$$

Thus $m = 0$ yields $(\underline{B}_t, \overline{C}_t) = (0, P)$, while a census yields (W_m, L_m). The gate authorizes only if

$$d_0 + \underline{B}_t - \overline{C}_t > 0.$$

The union bound gives at least $1 - t$ simultaneous coverage for the two counts, so false authorization is at most t conditionally on the history. The CLIP comparison uses $t = 0.05/8$, hence the same finite-sample family error allocation as CONFIRM. Unlike Eq. (2), this baseline does not exploit the joint win–loss–tie constraint. We make no claim against finite-population confidence procedures as a class.

The *unconditional paired exact* comparator directly tests the ternary paired difference, following the nuisance-maximized exact matched-pairs construction (Berger & Sidik, 2003). For $s_m = W_m - L_m$, define

$$p_{\mathrm{unc}}(s_m) = \max_{\substack{B,C \geq 0,\ B+C \leq P \\ d_0 + B - C \leq 0}} \Pr_{B,C,P-B-C}(W'_m - L'_m \geq s_m)\,,$$

where (W'_m, L'_m, T'_m) is multivariate hypergeometric. This comparator uses the same offset, paired outcomes, fresh draw, labels, and threshold as CONFIRM, but it does not condition on the sampled informative count n_m. Nuisance maximization makes its upper-tail p-value finite-sample valid: under any true null composition it upper-bounds that composition's ordinary right-tail p-value. Exhaustive enumeration over all null compositions for small P checks the implementation. Retain-incumbent and Stage 1-no-gate are decision controls rather than matched valid gates; naive reuse remains an intentionally invalid diagnostic. All boundary cases are explicit: an empty compatible null gives $p = 0$, while $W_m = 0$ with a compatible null gives $p = 1$.

Paired MABT is our incumbent-relative adaptation of the Rink–Brannath post-selection bound (Rink & Brannath, 2025): bootstrap tilting and a max-T multiplicity correction are applied to paired candidate minus incumbent correctness. It is a distributional post-selection comparator, not an exact finite-population gate.

Model Selector uses its published stochastic acquisition with $\epsilon = 0.47$. Active Testing uses 5% uniform support and the LURE estimator. Our PPAT implementation follows equations 8, 12 and 13 of the expanded July 2026 arXiv version. It shares a source-proxy proposal across the

model family, uses committee soft probabilities as the prediction-powered control variate, and applies the exact residualized LURE formula with coefficient one; this is a disclosed multi-model extension because PPAT is formulated for risk estimation rather than family-wise safe selection. Our HILTTA-FP reduction uses source margins and frozen source features for released K-margin and K-center acquisition, then ranks non-source candidates by its anchor MSE plus labeled cross-entropy components; it omits gradient updates from queried labels, so direct comparison is restricted to the fixed-prediction selection subproblem. TTALine implements ALine-D Algorithm 2 using aligned Photo ID logits saved before target labels; its all-source result occurs because candidate agreement leaves the fitted line unable to separate the stream-wise specialists. All comparisons are fixed-pool, not claims that our gate supersedes the baselines' original tasks.

For completeness, the conservatism knob that a practitioner might use instead of a gate was also swept: each baseline runs at its full budget with no two-stage split and deploys only when its own estimated margin over the source clears a threshold τ, with τ swept over every realized margin value. Choosing τ post hoc so that observed harm is exactly zero is an optimistic bound with no guarantee of any kind, and it is not available in advance to a practitioner. Under it, at $b = 100$ Active Testing-LURE reaches 58.016 and at $b = 400$ it reaches 58.258, while HILTTA-FP with a constructed knob falls to 56.262 and 56.274. We report this because it is the strongest thing that can be said for an ungated selector, and it still requires knowledge nobody has at deployment time.

C PACS Mechanism Audit and Robustness Diagnostics

Guide to the PACS diagnostics. The locked protocol establishes the original evaluation setting. The matched replay then isolates the effect of adding CONFIRM, the sub-pool and destructive-family studies stress candidate harm at finer granularity and under stronger regressions, and the final diagnostics examine Stage 1 acquisition and tie-handling behavior. Post-hoc analyses are labeled explicitly and do not modify the locked benchmark.

C.1 Locked Protocol and Integrity

We use PACS (Li et al., 2017a), fixing an immutable 9,991-example parquet before reading target labels. A frozen ImageNet ResNet-18 backbone (He et al., 2016) and linear head are trained on a deterministic half of Photo. The remaining Photo examples and Art, Cartoon, and Sketch form eight 1,024-example streams: three stationary targets, a balanced mixture, an abrupt recurrent stream, and three gradual Photo-to-target transitions. Each candidate starts from the same source checkpoint, with predictions recorded before each online update on batches of 64.

The $K = 32$ family contains the source, AdaBN (Li et al., 2017b), and 30 Tent, EATA, and SAR configurations spanning five learning rates and continual or episodic reset (Wang et al., 2021; Niu et al., 2022; 2023); $K = 16$ is a predeclared balanced prefix. We evaluate budgets 50, 100, 200, and 400 over 20 seeds at joint $\alpha = 0.05$. Nomination baselines are Uniform, Model Selector (Okanovic et al., 2025), Active Testing-LURE (Kossen et al., 2021), PPAT (Ashouritaklimi et al., 2026), HILTTA-FP—a fixed-pool use of HILTTA acquisition without its supervised online update (Li et al., 2024)—and TTALine ALine-D (Kim et al., 2024). Protocol, prediction, evaluation, data, and checkpoint artifacts are frozen and integrity-checked; replaying all 1,440 stored repetitions exactly reproduces the stored accuracy field. The prediction artifact has shape $(8, 32, 1024, 7)$ and contains finite logits with aligned IDs and features.

This benchmark isolates inference and selector wrapping rather than a new TTA ranking. Figure 6 gives the ungated accuracy and harm curves; Appendix C.6 explains why this family audits the gate but weakly separates selectors.

The source initializes from the official ImageNet-1K ResNet-18 checkpoint. The backbone is frozen; a seven-class linear head is trained for 100 epochs on half of Photo with SGD (learning rate 0.05, momentum 0.9, weight decay 10^{-4}, batch size 128, cosine decay). All image preprocessing is fixed: RGB, short-side resize 256, center crop 224, ImageNet normalization.

The eight streams are stationary Art, Cartoon and Sketch; a shuffled 256-per-domain balanced mixture; eight abrupt 128-example blocks in the order Photo, Art, Photo, Cartoon, Photo, Sketch, Cartoon, Art; and three gradual Photo-to-Art/Cartoon/Sketch streams whose target fraction in block b

is $b/7$. All permutations use seed 20260806 plus unit index. The 32 candidates are source; AdaBN; ten Tent, ten EATA and ten SAR candidates from learning rates $\{6.25\times10^{-5}, 1.25\times10^{-4}, 2.5\times10^{-4}, 5\times10^{-4}, 10^{-3}\}$ crossed with continual and episodic reset. EATA uses Fisher coefficient 2,000, redundancy margin 0.05 and reliability margin $0.4\log 7$; SAR uses SAM radius 0.05 (Foret et al., 2021), reliability margin $0.4\log 7$ and recovery threshold 0.2; SGD momentum is 0.9. This construction, and in particular the sweep of a single backbone over learning rate and reset policy, is the standard way the TTA literature defines a candidate family, and Appendix C.6 shows what it costs as a selection benchmark.

Table 4: Per-stream source and full-label oracle accuracy for $K = 16$ ($K = 32$ is identical), with the smallest source-relative accuracy change available in the locked pool. Appendix C.3 corrects the reading of the last column: the zeros are a property of the 1,024-example pooling, not of the stream.

Stream	Source	Oracle	Oracle model	Min. acc. change	Worst candidate
Stationary Art	56.15	56.15	Source	-1.074	AdaBN
Stationary Cartoon	17.87	45.51	Tent continual, 10^{-3}	0.000	Source
Stationary Sketch	28.52	42.09	SAR continual, 10^{-3}	0.000	Source
Balanced mixed	50.00	50.00	Source	-1.660	Tent continual, 10^{-3}
Abrupt recurrent	57.32	65.04	AdaBN	0.000	Source
Gradual Photo to Art	76.76	76.76	Source	-3.125	AdaBN
Gradual Photo to Cartoon	56.93	64.94	AdaBN	0.000	Source
Gradual Photo to Sketch	63.38	65.62	Tent continual, 6.25×10^{-5}	0.000	Source

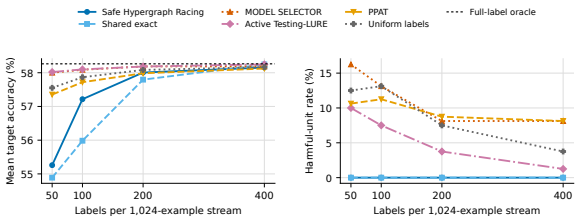

Figure 6: Locked PACS, $K = 16$, mean over 20 seeds and eight streams: the ungated contrast that motivates the gate. Left, accuracy approaches the full-label oracle; right, ungated selectors trade higher low-budget accuracy for source-worse choices. Under the matched protocol of Section 5.1 the right panel collapses to zero for *every* selector, which is the point of this paper.

Matched offset-reuse ablation (post hoc). Table 5 supplies the numerical comparisons reported in Section 5.1. Within each setting, CONFIRM and adversarial-history exact share the Model Selector nominee, nomination history, fresh confirmation draw, label budget, and deployment-level correction; they differ in how the observed nomination outcomes enter the test (Appendix B.2). These are post-hoc comparisons on frozen predictions, not new adaptation runs. Figure 4 (left) shows only deployment rates for $K = 16$ at the 25% share; the table also reports accuracy gains, the frozen 5% share, and the $K = 32$ extension.

C.2 Matched-Safety Replay (Post Hoc)

The disclosed post-hoc matched harness replays the frozen artifacts only; no TTA was re-run and nothing was trained. SHR re-run inside the harness reproduces the frozen evaluation accuracy exactly in all eight (K, b) cells. Every arm receives $B_{\text{tot}} = \min(b, |U|)$ unique labels, stage 1 spending $E = \min(\text{round}(0.05B_{\text{tot}}), B_{\text{tot}} - 1)$ of them.

The same terminal audit used for CINIC-10 changes only the locked-PACS ungated Model Selector cell at $b = 400$: six units exhaust their disagreement pool before 400 labels, so the corrected census is 58.264 with $0/160$ harm, not the frozen 58.188 with $13/160$. The gated nomination is intentionally taken at the smaller exploration budget and is unaffected. DomainBed is also unaffected because every $K = 16$ union contains 830–1,001 examples.

Table 5: Matched offset-reuse gains on frozen PACS predictions, using Model Selector nomination. Each entry is CONFIRM minus adversarial-history exact in percentage points (pp), averaged over eight fixed 1,024-example streams and 20 query seeds (160 repeated decisions per setting, not 160 independent populations). b is the total label budget; K includes the incumbent. Deployment gains are rounded to three decimals and accuracy gains to four; differences are computed before rounding.

		5% nomination share		25% nomination share	
K	b	Deploy. gain (pp)	Acc. gain (pp)	Deploy. gain (pp)	Acc. gain (pp)
16	50	0.625	0.0824	10.000	1.1963
16	100	0.625	0.0824	23.750	2.1765
16	200	3.125	0.0702	40.000	3.5059
16	400	0.000	0.0000	50.000	3.9429
32	50	0.000	0.0000	10.000	1.1597
32	100	0.000	0.0000	25.000	2.3413
32	200	3.750	0.0842	39.375	3.6334
32	400	0.625	0.0140	50.000	3.9429

Matched protocol. Each selector is wrapped by Algorithm 1 under the same label budget. Let $b_{\mathrm{tot}} = \min(b, |U|)$ and

$$b_{\mathrm{nom}} = \min\{\mathrm{round}(0.05 b_{\mathrm{tot}}), b_{\mathrm{tot}} - 1\}\,.$$

Stage 1 spends at most b_{nom} unique labels under the selector's own acquisition rule; Stage 2 uses the remainder. The primary *forced* variant nominates the selector's best non-incumbent model, whereas a verbatim variant permits the selector to retain the incumbent without testing. Random- and oracle-survivor controls nominate, respectively, a random candidate and the truly best non-incumbent candidate. LURE, PPAT, HILTTA-FP, and Uniform query the full 1,024-example pool and may spend most Stage-1 labels outside the final nominee's disagreement set; such labels simply do not enter its paired offset.

Result. Across the five wrapped baselines, every (K, b) cell has $0/160$ harmful deployments: 6,400 gated decisions in the forced variant and another 6,400 in the verbatim variant. SHR also records zero harm in its eight cells, whereas ungated harm reaches 37.5%. These decisions repeatedly sample the same eight streams and are not 12,800 independent safety trials. The oracle-survivor control makes one harmful replacement at $K = 16, b = 100$, losing 1.074 accuracy points. This confirms that the implementation can reject a true null; a single event neither contradicts the level-α guarantee nor estimates calibration.

Table 6: Disclosed post-hoc replay on locked PACS at $K = 16$ under forced nomination (160 decisions per cell). Entries are accuracy (harm/160); bold marks the descriptive gated maximum among non-oracle selectors.

	+CONFIRM: accuracy↑ (harm↓)				Selector alone: accuracy↑ (harm↓)			
Stage-1 selector	50	100	200	400	50	100	200	400
SHR front end	55.259 (0)	**57.216 (0)**	58.006 (0)	58.154 (0)	not evaluated			
Model Selector	**56.106 (0)**	57.084 (0)	**58.068 (0)**	**58.264 (0)**	58.001 (26)	58.088 (21)	58.188 (13)	58.264 (0)
Active Testing-LURE	55.864 (0)	56.988 (0)	57.960 (0)	58.186 (0)	58.021 (16)	58.099 (12)	58.184 (6)	58.247 (2)
PPAT	55.911 (0)	57.097 (0)	57.652 (0)	58.171 (0)	57.352 (17)	57.719 (18)	57.985 (14)	58.124 (13)
HILTTA-FP	55.432 (0)	56.940 (0)	57.881 (0)	58.179 (0)	57.544 (60)	57.544 (60)	57.544 (60)	57.556 (60)
Uniform	55.452 (0)	56.982 (0)	57.940 (0)	58.182 (0)	57.552 (20)	57.867 (21)	58.083 (12)	58.168 (6)
Random survivor	55.850 (0)	57.136 (0)	57.943 (0)	58.170 (0)	not applicable			
Oracle survivor	55.900 (0)	57.178 (1)	57.926 (0)	58.264 (0)	not applicable			

No ungated PACS SHR run was frozen. Model Selector at $b = 400$ uses the disclosed post-reveal repair; its pre-repair result is 58.188 with 13 harms. Realized label counts agree through $b = 200$ up to one 0.2-label difference; at $b = 400$, SHR uses 0.1–0.34% fewer labels than its comparators. No confirmation draw is empty, and valid finite-population censuses occur in at most 67/640 runs per method. Because the incumbent is oracle on three of eight streams, each forced arm contains 60 harmful nominations per cell. The gate rejects all of them for every deployable selector, apart from the single oracle-control event above.

Table 7: Matched safety, $K = 32$, forced variant. Accuracy points with harmful deployments out of 160 in parentheses. Every cell is $0/160$.

Stage-1 selector	b=50	b=100	b=200	b=400
SHR (hypergraph)	55.177 (0)	56.970 (0)	58.006 (0)	58.154 (0)
Model Selector	55.833 (0)	57.057 (0)	57.913 (0)	58.261 (0)
Active Testing-LURE	56.102 (0)	56.730 (0)	57.838 (0)	58.198 (0)
PPAT	55.850 (0)	56.967 (0)	57.869 (0)	58.174 (0)
HILTTA-FP	55.688 (0)	56.891 (0)	57.759 (0)	58.179 (0)
Uniform labels	56.071 (0)	56.526 (0)	57.889 (0)	58.175 (0)
Random survivor	55.601 (0)	56.913 (0)	57.690 (0)	58.176 (0)
Oracle survivor	56.154 (0)	57.271 (0)	58.011 (0)	58.264 (0)

Table 8: Verbatim variant, $K = 16$: the selector's own top-1 over all models is used, and if it names the source no test is run. Accuracy points with harmful deployments out of 160 and realized unique labels per seed over the eight streams. Every cell is $0/160$; the same holds at $K = 32$.

	Accuracy (harmful /160)				Realized unique labels			
Selector	50	100	200	400	50	100	200	400
Model Selector	56.085 (0)	57.330 (0)	58.026 (0)	57.897 (0)	380.8	695.5	1296.4	2194.9
Active Testing-LURE	55.320 (0)	56.575 (0)	56.971 (0)	57.878 (0)	306.4	538.8	1016.9	2042.0
PPAT	53.738 (0)	54.714 (0)	55.839 (0)	56.543 (0)	203.2	396.2	934.0	1635.9
HILTTA-FP	55.432 (0)	56.940 (0)	57.881 (0)	58.179 (0)	400.0	800.0	1572.0	2673.0
Uniform labels	55.516 (0)	56.706 (0)	57.261 (0)	57.673 (0)	354.4	652.8	1205.5	2167.9
Random survivor	55.850 (0)	57.136 (0)	57.943 (0)	58.170 (0)	400.0	800.0	1572.0	2673.0
Oracle survivor	55.900 (0)	57.185 (0)	57.898 (0)	58.264 (0)	256.0	515.0	1029.0	1868.0

Draw bookkeeping. Over the 640 cells per method per K at $K = 16$, draws truncated to a census of the survivor's entire unexplored disagreement set number 61 for SHR, 22 for Model Selector, 67 for Active Testing-LURE and 0 for PPAT, HILTTA-FP, Uniform and both controls; empty draws ($C = 0$) number 0 for every method. In the verbatim variant stage 1 names the source in 97, 198, 305, 0 and 126 cells for Model Selector, Active Testing-LURE, PPAT, HILTTA-FP and Uniform, and in 240 cells for the oracle control. Stage-1 nomination of the true best non-source candidate at $b = 400$ is 62.5% for SHR, 24.4% for Model Selector, 8.8–10.0% for the remaining selectors and the random control, 0.0% for HILTTA-FP and 25.0% for the oracle control. This column uses a strict index-match convention that does not credit ties in the ground truth, which is why even the oracle control scores 25.0%; Appendix C.6 uses a tie-crediting convention under which the same control scores 100%.

Robustness to the acquisition share. Applying acquisition shares 0.25 and 0.50 identically to every arm leaves every gated cell at zero harm except two isolated events (Uniform at $\rho = 0.25$, $b = 100$, and the random-survivor control at $\rho = 0.50$, $b = 100$), and does not change the ranking of methods: at $\rho = 0.25$ the $K = 16$ accuracies at $b = 50/100/200/400$ are $55.144/56.320/57.591/58.193$ for SHR against $55.228/56.543/57.828/58.264$ for Model Selector plus gate, and at $\rho = 0.50$ they are $54.686/56.409/57.502/58.231$ against $54.575/56.299/57.672/58.264$.

Contrasts under the matched gate, locked family, $K = 16$. SHR minus each gated comparator, with descriptive stream-resampled 95% intervals: at $b = 50$, -0.847 $[-1.801, -0.099]$ against Model Selector, -0.591 $[-1.204, -0.092]$ against random survivor and -0.641 $[-1.264, -0.129]$ against oracle survivor; at $b = 100$ every contrast contains zero (for example $+0.132$ $[-0.195, +0.654]$ against Model Selector); at $b = 200$ only PPAT is resolved, $+0.354$ $[+0.040, +0.765]$; at $b = 400$, -0.110 $[-0.242, -0.012]$ against Model Selector and against oracle survivor. These descriptive intervals condition on the stored confirmation-seed stream.

C.3 Granularity Stress Test

Each stream is cut into four contiguous 256-example sub-pools, giving 32 deployment units with Holm across all 32 at joint $\alpha = 0.05$ and $\lfloor b/4 \rfloor$ labels per sub-pool, so only $b = 50$ differs from the 8-unit study (48 rather than 50 labels per stream). The four sub-pools of a stream share one adaptation

trajectory and are not independent clusters; the effective cluster count remains eight and no interval narrows.

Table 9: Smallest source-relative accuracy change available to an adversary at $K = 16$, on the full pool and each 256-example sub-pool (points). Positive means no candidate is worse than source. "New" marks streams that look structurally harm-free at 1,024 examples but are not at 256.

Stream	Full pool	q1	q2	q3	q4	New
Stationary Art	-1.074	$+0.781$	-1.562	$+1.172$	-5.078	no
Stationary Cartoon	$+27.246$	$+30.078$	$+22.266$	$+29.297$	$+27.344$	no
Stationary Sketch	$+13.086$	$+10.156$	$+14.844$	$+13.672$	$+13.281$	no
Balanced mixed	-1.660	-2.734	-3.516	-1.562	$+1.172$	no
Abrupt recurrent	$+7.520$	$+0.781$	$+7.031$	$+4.688$	$+17.578$	no
Gradual Photo to Art	-3.125	-4.297	-1.953	-6.250	$+0.000$	no
Gradual Photo to Cartoon	$+7.910$	-2.344	-0.391	$+8.594$	$+25.781$	**yes**
Gradual Photo to Sketch	$+2.148$	-6.250	-3.906	$+3.516$	$+15.234$	**yes**

Table 10: 32-unit decision outcomes at $K = 16$: harmful deployments out of 640 for the ungated per-sub-pool protocol against the gated one, with the ungated worst single-deployment loss in points. Every gated cell is $0/640$, at $K = 32$ as well.

	Ungated harmful /640				Gated harmful /640				Ungated
Selector	50	100	200	400	50	100	200	400	worst
Model Selector	73	40	10	0	0	0	0	0	-6.250
Active Testing-LURE	50	33	21	3	0	0	0	0	-6.250
PPAT	81	64	51	47	0	0	0	0	-6.250
HILTTA-FP	240	240	240	240	0	0	0	0	-6.250
Uniform labels	40	42	35	23	0	0	0	0	-6.250
SHR	not evaluated				0	0	0	0	n/a

At 256-example granularity harm is no longer confined to the three source-is-oracle streams: at $b = 50$ the ungated Model Selector's 73 harmful decisions fall on stationary Art (11), balanced mixed (35), gradual Photo-to-Art (12), gradual Photo-to-Cartoon (12) and gradual Photo-to-Sketch (3), and the pattern persists at every budget for every ungated selector.

C.4 Destructive-Candidate Stress Test (Post Hoc)

Status. Disclosed post-hoc exploratory stress test. The family was constructed after the locked protocol was frozen and after target labels were opened. It is not part of the locked benchmark. The locked artifacts were read only and their integrity records remain unchanged. Re-running `source`, `adabn` and `tent_continual_lr0.001` through the new code reproduces the frozen artifact bitwise in float16 on every checked cell, so the destructive candidates come out of the identical stream construction, batching, pre-update prediction convention, preprocessing and frozen source checkpoint. The post-hoc prediction artifacts are separately integrity-recorded.

Definitions. The two `D_tent_continual` candidates are released Tent, continual, SGD (momentum 0.9) at lr 10^{-2} and 3×10^{-2}. `D_sar_continual_lr0.01` is released SAR, continual, SAM (lr 10^{-2}, $\rho = 0.05$); its recovery rule fires repeatedly and keeps resetting the model, which is why it does not collapse. `D_eata_continual_lr0.01` is released EATA at lr 10^{-2} with the locked family's Fisher anchor, and the anchor is what stops it collapsing. `D_adabn_frozen_first_batch` gives the BatchNorm layers momentum 1.0 and a single train-mode forward pass on the first batch, then freezes them; `D_adabn_single_image` is identical but takes statistics from one image; `D_adabn_bs1` normalizes every test example using itself alone, with no gradient update; and `D_tent_bs1_lr0.001` keeps batches of 64 for prediction but applies the entropy update 64 times per batch on single examples. All eight are label-free at generation time; only the *choice* of which configurations to keep was informed by a one-stream pilot after labels were open, which is part of why this is exploratory.

Worst attainable harm per stream. Identically for $K = 16 + 8$ and $K = 32 + 8$, the smallest source-relative accuracy change available moves from the locked values -1.074, $+0.000$,

Table 11: Per-stream accuracy change against the source, in points, for the eight destructive candidates.

Candidate	St. Art	St. Cart.	St. Sk.	Bal. mix	Abrupt	Gr. Art	Gr. Cart.	Gr. Sk.	Pooled
`D_tent_continual_lr0.01`	+0.49	+29.49	+14.26	−3.61	+8.30	−2.25	+8.11	+0.68	+6.934
`D_tent_continual_lr0.03`	+2.25	+31.45	+16.21	−6.54	+8.20	−1.56	+7.81	−3.22	+6.824
`D_sar_continual_lr0.01`	−1.07	+29.00	+13.96	−1.27	+7.71	−3.12	+8.01	+2.15	+6.921
`D_eata_continual_lr0.01`	−1.07	+27.64	+14.06	−1.37	+7.62	−3.03	+8.01	+2.25	+6.763
`D_adabn_frozen_first_batch`	−0.49	+28.61	+13.87	−4.10	−6.54	−11.23	−7.32	−2.64	+1.270
`D_adabn_single_image`	−32.42	+4.98	−3.91	−17.38	−24.80	−31.64	−12.21	−27.83	−18.152
`D_adabn_bs1`	−37.21	−2.83	−8.89	−35.35	−42.68	−61.43	−43.26	−48.63	−35.034
`D_tent_bs1_lr0.001`	−6.05	+25.68	+4.49	−2.25	+5.08	−5.27	+5.96	−1.47	+3.271

+0.000, −1.660, +0.000, −3.125, +0.000, +0.000 to the augmented values −37.207, −2.832, −8.887, −35.352, −42.676, −61.426, −43.262, −48.633 points on stationary Art, Cartoon and Sketch, balanced mixed, abrupt recurrent and the three gradual streams respectively; the worst candidate is `D_adabn_bs1` on all eight. The count of candidates worse than source rises from 15/0/0/15/0/15/0/0 to 21/1/2/23/3/23/3/5 at $K = 16 + 8$.

Table 12: Destructive family, $K = 16 + 8 = 24$: harmful deployments out of 160. "Worst" is the largest single-deployment loss in accuracy points over all budgets. Every gated cell is $0/160$; the same is true at $K = 32 + 8 = 40$ and in the verbatim variant.

	Ungated (no gate)					Gated				
Selector	50	100	200	400	Worst	50	100	200	400	Worst
Model Selector	29	22	13	0	−1.562	0	0	0	0	n/a
Active Testing-LURE	24	15	13	7	−3.125	0	0	0	0	n/a
PPAT	29	19	14	5	−4.102	0	0	0	0	n/a
HILTTA-FP	60	60	60	40	−2.637	0	0	0	0	n/a
Uniform labels	16	21	10	3	−5.273	0	0	0	0	n/a
Random survivor			not evaluated			0	0	0	0	n/a
Oracle survivor			not evaluated			0	0	0	0	n/a

Table 13: Augmented family, $K = 16 + 8 = 24$: gated accuracy in points, all cells $0/160$ harmful. Source is 50.867 and the oracle 59.436. SHR here is the shipped acquisition inside the same gate.

Stage-1 selector	b=50	b=100	b=200	b=400
SHR (hypergraph)	55.530	56.888	57.992	58.352
Model Selector	55.247	57.625	57.921	58.853
Active Testing-LURE	55.375	56.592	57.770	58.402
PPAT	53.522	56.481	57.559	58.077
HILTTA-FP	55.615	56.561	57.103	57.056
Uniform labels	55.331	56.658	57.979	58.491
Random survivor	55.446	56.031	57.271	57.802
Oracle survivor	56.755	58.262	59.099	59.436

At $K = 32 + 8 = 40$ the pattern is the same, with gated accuracies at $b = 400$ of 58.444 for SHR, 58.781 for Model Selector, 58.542 for Active Testing-LURE, 58.032 for PPAT, 57.056 for HILTTA-FP, 58.508 for Uniform, 57.546 for the random control and 59.436 for the oracle control, and every cell at $0/160$. Draw bookkeeping over 640 cells per method shows 124 to 160 census truncations and $C = 0$ in zero cells, for both augmented families.

C.5 Stage-1 Acquisition Contrasts

This disclosed post-hoc comparison holds the confirmation gate fixed and asks whether SHR's Stage-1 front end separates from random-survivor and Model Selector nomination on the augmented candidate family.

C.6 Selector Degeneracy and Tie Handling (Post Hoc)

Status. Disclosed post-hoc diagnostic. Every tie-aware number here comes from an algorithm written after the locked protocol was frozen and after the target labels were opened, and is not part of the locked result. The locked pipeline is not modified; every arm uses the identical stage-2 gate and label budget.

Table 14: Post-hoc Stage-1 contrasts on the augmented family: SHR minus comparator in accuracy points, with descriptive 95% stream-resampled percentile intervals (8 clusters, 10,000 draws). Positive favors SHR; $\star$ marks intervals excluding zero.

		vs. random survivor + gate		vs. Model Selector + gate	
Family	b	Difference	95% interval	Difference	95% interval
K=16+8	50	$+0.084$	$[-0.297, +0.558]$	$+0.284$	$[-0.444, +1.008]$
K=16+8	100	$+0.856^\star$	$[+0.074, +1.805]$	-0.737	$[-1.680, +0.031]$
K=16+8	200	$+0.721^\star$	$[+0.076, +1.487]$	$+0.071$	$[-0.726, +1.176]$
K=16+8	400	$+0.551^\star$	$[+0.092, +1.041]$	$-0.500^\star$	$[-0.745, -0.258]$
K=32+8	50	$+0.306$	$[-0.267, +0.886]$	-0.258	$[-0.750, +0.049]$
K=32+8	100	$+0.114$	$[-0.385, +0.574]$	-0.195	$[-0.853, +0.370]$
K=32+8	200	$+0.302^\star$	$[+0.045, +0.584]$	$+0.591$	$[-0.468, +2.160]$
K=32+8	400	$+0.897^\star$	$[+0.134, +1.745]$	-0.337	$[-0.739, +0.048]$

Table 15: Structural degeneracy measured on the frozen artifacts. $|U|$ is the source-candidate disagreement union; $|P|$ the pairwise-separating set; "agree" the fraction of U on which all incident candidates agree with each other; "spread" the ground-truth accuracy range of the 15 non-source candidates in points.

Stream	$\lvert U\rvert$	$\lvert P\rvert$	$\lvert P\rvert/\lvert U\rvert$	Agree	Spread
Stationary Art	326	5	1.53%	100.0%	0.00
Stationary Cartoon	703	9	1.28%	99.7%	0.39
Stationary Sketch	763	17	2.23%	99.3%	0.49
Balanced mixed	350	16	4.57%	98.0%	0.78
Abrupt recurrent	371	4	1.08%	99.5%	0.20
Gradual Photo to Art	172	2	1.16%	100.0%	0.00
Gradual Photo to Cartoon	302	1	0.33%	100.0%	0.10
Gradual Photo to Sketch	352	5	1.42%	99.4%	0.10

Diagnosis. Table 15 summarizes the structural degeneracy of the locked candidate family. Two more are worth recording. First, the acquisition concentrates on 1.0 to 1.02 distinct strata of the 2 to 11 available, which is maximal concentration, but the dominant stratum holds 97.4–99.7% of U, so concentration costs no coverage and the consensus rate under the hypergraph rule (99.1–99.6%) is indistinguishable from uniform sampling (99.1–99.8%). Second, using the best possible statistic on the purchased labels, the mean Spearman correlation with the true gap is 0.019 at $b = 50$ and only 0.269 at $b = 400$ ($K = 16$), with a top-1 hit rate of 53.3% against a chance rate of 51.1%. There is almost nothing in the labels to rank with.

The patch. Stage 1 first censuses the pairwise-separating set P in greedy set-cover order over unresolved candidate pairs (label-free selection); guarantees every candidate a coverage floor of incident labels; and ranks by $\widehat{\Delta}_j \pm z\, w_j$ with $z = 1$. Specifically, if S_j is the known paired-loss sum, M_j the unresolved disagreement mass, and the uniform phase provides n_j^B incident outcomes with mean $\widehat{r}_j$ and variance $\widehat{v}_j = \max(s_j^2, 0.05)$, then $\widehat{\Delta}_j = (S_j + M_j\widehat{r}_j)/N$ and $w_j = (M_j/N)\sqrt{\widehat{v}_j/(n_j^B + 1)}$. Thus unobserved mass is never scored as zero. The patch nominates the candidate the gate is most likely to certify, using the projected stage-2 p-value with a continuous sign-test tie-break; breaks remaining ties by a seeded random draw rather than by candidate index; and, in the budget-adaptive variant, spends $\mathrm{round}(0.05b) + |P|$ acquisition labels so that the separating set is bought outright.

Over all eight (K, b) cells, i.e. $1{,}280$ gated unit decisions per arm, the total harmful deployments are 0 for the shipped rule, 0 for the random tie-break, 0–1 for the census variants, 1 for the matched random control and 0 for Model Selector and the oracle control. These observed harm counts are consistent with Theorem 1, which applies separately to each Stage 1 rule; the theorem does not require identical empirical outcomes across tie-handling variants.

Attribution is measurable. Isolating the acquisition change from the tie-break change, the separation census is worth $+0.08$ to $+0.42$ points and excludes zero in seven of the eight cells, while the tie-break change alone has mixed signs with five of eight intervals containing zero, which is what a fair tie-break applied to an all-tied vector should look like. Against the matched random control the

Table 16: Degeneracy before and after, $K = 16$. "Tied" is the mean number of candidates tied at the minimum of that arm's ranking statistic out of 15; "all" the fraction of races where every candidate ties; "adabn" the fraction nominating the lowest-index candidate; "#surv" the number of distinct candidates ever nominated over 160 races; "optimal" the fraction nominating a truly optimal candidate.

b	Arm	Tied	All	adabn	#surv	Optimal
50	Shipped (locked)	15.00	100.0%	100.0%	1	62.5%
50	Random tie-break only	15.00	100.0%	6.2%	15	63.7%
50	Separation census + UCB	7.45	0.0%	5.6%	15	68.1%
50	Budget-adaptive census	12.14	63.1%	5.0%	15	100.0%
100	Shipped (locked)	14.91	98.8%	99.4%	2	61.9%
100	Separation census + UCB	7.26	12.5%	2.5%	15	90.6%
100	Budget-adaptive census	11.33	50.0%	3.8%	15	100.0%
200	Shipped (locked)	14.82	97.5%	98.8%	3	61.9%
200	Separation census + UCB	9.51	25.0%	3.8%	15	98.1%
200	Budget-adaptive census	12.01	61.3%	4.4%	15	100.0%
400	Shipped (locked)	14.75	96.2%	98.1%	4	62.5%
400	Separation census + UCB	9.50	25.0%	4.4%	15	100.0%
400	Budget-adaptive census	13.81	89.4%	7.5%	15	100.0%

Table 17: Re-measurement under the matched gate, $K = 16$, accuracy points with harmful deployments out of 160 in parentheses. All arms in this run share one confirmation-draw seed stream, so contrasts isolate stage 1; the shipped column differs from Table 6 because this diagnostic uses a different fixed confirmation-seed stream.

Arm	b=50	b=100	b=200	b=400
Shipped (locked)	55.762 (0)	57.086 (0)	57.912 (0)	58.154 (0)
Random tie-break only	55.604 (0)	57.023 (0)	57.845 (0)	58.173 (0)
Separation census, point statistic	55.759 (0)	57.441 (0)	57.925 (0)	58.264 (0)
Separation census + UCB	55.745 (1)	57.443 (0)	57.925 (0)	58.264 (0)
Budget-adaptive census	55.788 (0)	57.352 (0)	57.941 (0)	58.264 (0)
Random survivor + gate	55.778 (1)	57.238 (0)	57.761 (0)	58.175 (0)
Model Selector + gate	55.570 (0)	57.599 (0)	57.899 (0)	58.264 (0)
Oracle survivor + gate	55.865 (0)	57.598 (0)	57.870 (0)	58.264 (0)

census variants exclude zero in four to five of eight cells at 95% and three to four after a Bonferroni correction across the eight cells; against Model Selector plus gate they are indistinguishable in seven to eight of eight. The tie-aware variant draws level with the strongest baseline; it does not pass it. The budget-adaptive census buys the complete separating set in 100% of races at every budget for a mean of 7.38 extra labels, and is 0.669 points behind the random tie-break arm at $K = 32$, $b = 50$, interval $[-1.368, -0.071]$, the one cell where the census is not worth its price.

D DomainBed Benchmark and Certifiability Diagnostics

D.1 Protocol and Cross-Domain Dominance

The predeclared prediction-blind study uses four leave-one-domain-out PACS incumbents, two training seeds per held-out domain and eight streams per incumbent, for 64 deployment units. The $K = 16$ prefix comprises the incumbent and its adaptations: BatchNorm-statistic mixtures, Tent/EATA/SAR, LAME (Boudiaf et al., 2022), T3A (Iwasawa & Matsuo, 2021), SHOT-IM (Liang et al., 2020), and hard pseudo-labeling (Lee, 2013) variants. SHOT-IM uses the information-maximization objective, not the full SHOT pipeline. The $K = 32$ extension also includes the alternate source checkpoint and its BN-test adaptation. All checkpoints, candidate predictions and stream indices were frozen before held-out accuracies were computed. The full-label analysis below is evaluation-only and was never available to a selector or to the confirmation gate.

The eight streams are `stationary_target`, `stationary_source`, `balanced_quad`, `target_heavy`, `source_heavy`, `abrupt_recurrent`, `gradual_source_to_target` and `gradual_target_to_source`. Their target-domain shares are $100\%, 0\%, 25\%, 75\%, 25\%$, alternating recurrent blocks, a gradual increase and a gradual decrease, respectively; every stream has 1,024 examples.

Table 18: Cross-domain oracle analysis. Accuracy is averaged over the 16 units for each held-out domain (two incumbents × eight streams). The identity of the best model changes with the domain; no adapter is uniformly best.

Held-out domain	$K = 16$ pure TTA		$K = 32$ heterogeneous	
	Oracle model	Acc.	Oracle model	Acc.
Art Painting	LAME-kNN5	89.758	LAME-kNN5	89.758
Cartoon	BN-mix $\alpha = .75$	86.877	BN-mix $\alpha = .75$	86.877
Photo	Source	95.044	T3A-M100	95.172
Sketch	BN-mix $\alpha = .75$	87.091	BN-mix $\alpha = .75$	87.091

The domain averages do not conceal a single unit-level winner. At $K = 16$, unique oracle wins across the 64 units are distributed among Source (21), LAME-kNN5 (14), BN-mix $\alpha = .75$ (10), T3A-M20 (5), BN-mix $\alpha = .25$ (4) and BN-mix $\alpha = .50$ (3), with seven tied units. The best average-rank candidate, BN-mix $\alpha = .75$, has domain ranks $(4, 1, 4, 1)$ for (Art, Cartoon, Photo, Sketch), so it is not universally dominant. At $K = 32$, unique wins remain dispersed: alternate source (19), Source (14), BN-mix $\alpha = .75$ (8), LAME-kNN5 (6), T3A-M100 (6), BN-mix $\alpha = .25$ (4), and three other candidates (3 total), with four tied units. This heterogeneity is precisely why nomination and source-relative confirmation are separate decisions.

The predeclared contrast summary also supports the scope statement in the main text. Across the eight (K, b) cells, SHR is ahead of the gated random survivor in six and statistically unresolved in two under the disclosed two-stage domain/unit bootstrap; against gated Model Selector it is ahead in zero, behind in three and unresolved in five. These are descriptive contrasts over four held-out-domain clusters, not an independent-selector leaderboard.

The maximum ungated harm rate is $1076/1280 = 84.06\%$ for the random-nomination control at $K = 32$ and $b = 100$, across 64 fixed pools and 20 query repetitions. Excluding random and oracle controls, ungated harm ranges from 21.9% to 68.8%.

D.2 Gate Power, Utility, and Null Calibration (Post Hoc)

Figure 7 isolates three reviewer-relevant questions using frozen predictions and stored selection arrays: gate power conditional on a good nominee, safety–utility trade-offs, and selected-unit null calibration. First, an oracle nominee isolates gate power from nomination quality: among 740 repeated positive nominations, deployment is 68.6%, 93.9%, 100% and 100% as b increases, while all 540 nonpositive nominations are rejected at every budget. Table 19 gives the exact effect-bin counts and deployment rates. Their nonmonotonicity reflects disagreement structure: full-pool benefit is normalized by N, whereas the test samples within the nominee's disagreement set. The rates are descriptive because the same 64 units recur. Second, at DomainBed $K = 16, b = 400$, the harm-cost break-even values are 0.774, 0.692, 0.860, 0.658 and 0.721 for SHR, Model Selector, LURE, PPAT and Uniform. HILTTA-FP and TTALine have higher gated accuracy even before adding a harm penalty. Third, in 5,000 independent selected-unit complete-null simulations, the empirical rejection probability at nominal 0.05 is 3.40% for CONFIRM and 7.42% for naive reuse. The former is consistent with the conservative super-uniformity described in Remark 1, whereas the latter demonstrates anti-conservative reuse after adaptive nomination. This panel is a marginal p-value diagnostic, distinct from the family-wise simulation in Figure 4.

The generating script, JSON, seeds, and input integrity records are preserved in the research archive, not all in the compact supplement. No result in this section changes a frozen success criterion.

E Sealed CLIP Adapter Audit

Design and sealing. The incumbent is OpenAI CLIP ViT-B/16 with the single prompt `a photo of a [class]`. Four variants each of TPT, PromptAlign, TDA and ZERO form 16 adapted candidates. Variants are the published default plus prespecified perturbations of learning rate or step count, alignment weight, cache scale or confidence fraction; no value was selected from target accuracy. For each of ImageNet-A (Hendrycks et al., 2021b), ImageNet-R (Hendrycks et al.,

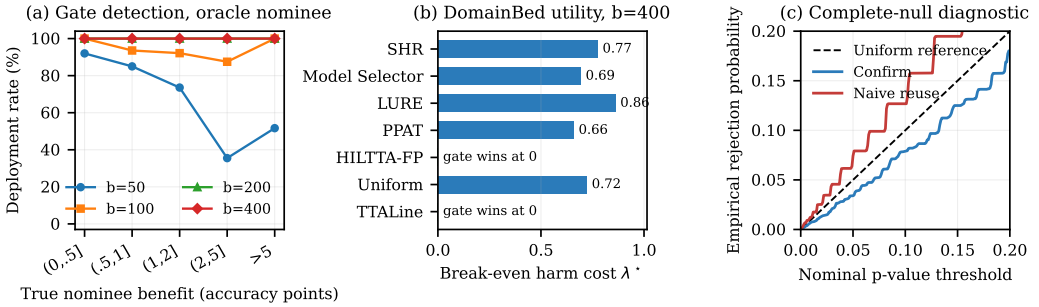

Figure 7: Disclosed post-hoc diagnostics from frozen predictions. (a) Deployment versus full-pool benefit for an oracle nominee. (b) Harm cost at which gated utility weakly exceeds ungated utility; methods labeled "gate wins at 0" already improve at zero cost. (c) Selected-unit null p-value ECDF.

Table 19: Oracle-nominee deployment by full-pool accuracy benefit. Entries are deployment rates (%); the n row gives repeated nominations in each fixed bin and is shared across budgets.

b	≤ 0	$(0, 0.5]$	$(0.5, 1]$	$(1, 2]$	$(2, 5]$	> 5	All > 0
n	540	200	140	140	200	60	740
50	0.0	92.0	85.0	73.6	35.5	51.7	68.6
100	0.0	100.0	93.6	92.1	87.5	100.0	93.9
200	0.0	100.0	100.0	100.0	100.0	100.0	100.0
400	0.0	100.0	100.0	100.0	100.0	100.0	100.0

2021a), ImageNet-Sketch (Wang et al., 2019), ImageNet-V2 matched-frequency (Recht et al., 2019), DTD (Cimpoi et al., 2014), Food-101 (Bossard et al., 2014), Oxford-IIIT Pets (Parkhi et al., 2012) and EuroSAT (Helber et al., 2019), the 1,024 smallest deterministic hash ranks of a fixed salt and relative image path define the pool. Candidates use 64 augmented views and one common seed within family. The evaluator uses $b \in \{50, 100, 200, 400\}$, 20 query seeds, a 5% selection share and level $0.00625 = 0.05/8$ per dataset. No-gate, adversarial-history exact (the archived `split_exact` key), paired MABT and CONFIRM use the same frozen candidate matrix. Adversarial-history exact and CONFIRM also share the nominee, query history, fresh draw, and realized label count, so their difference isolates historical evidence use.

Before label opening, 128 candidate files passed a target-absence audit and all 17 prediction rows per dataset used identical subset indices. Candidate files and pre-label artifacts were recorded in a frozen integrity manifest. After opening, eight label arrays and eight $(17, 1024)$ artifacts were created. Re-running the evaluator reproduces every selected candidate, decision and summary; MABT bounds differ by at most 4.44×10^{-16} across BLAS implementations and no decision changes.

What is and is not established. The seven no-gate harms at $b = 100$ are seven query realizations on one ImageNet-Sketch pool, not seven independent domain failures. Each costs $9/1024 = 0.879$ accuracy points; no gate nevertheless has the highest accuracy at every budget and harm only in this one cell. Across budgets, CONFIRM makes 30 deployments that adversarial-history exact withholds, never makes the reverse decision, and every addition improves on source (conditional mean gain 1.790 points). Averaged over all 560 decisions, the accuracy difference is 0.096 points. MABT is more accurate than CONFIRM by 1.179 points at $b = 200$ and 0.075 at $b = 400$, but uses 18.5% and 43.9% more realized labels. Thus matched-cost accuracy against MABT is unresolved; adversarial-history exact is the worst-case history-discard comparison, while MABT is a distributional post-selection comparator rather than an exact finite-population guarantee.

Matched valid-gate diagnostic (post hoc). Table 2 adds two gates on the identical nominee, history, and draw. Capping d_0 at zero changes aggregate accuracy by only 0.013 points, so favorable

Table 20: Seven-dataset sealed aggregate on unaffected datasets. ImageNet-V2 is excluded for the representation defect below, while the conservative eight-unit level is retained. Accuracy and deployment are percentages; source delta and oracle regret are in accuracy points. Harm is a count out of 140 dataset–repetition decisions per budget and method. Oracle regret is the mean accuracy gap to the full-pool-best model. No gate uses the full nomination budget, unlike the shared Stage-1 nominee in Table 2.

b	Method	Accuracy	Source delta	Deploy	Harm	Oracle regret	Labels
50	No gate	65.361	+4.925	100.0	0	1.213	50.0
	Adv.-history exact	63.295	+2.860	44.3	0	3.278	45.1
	Paired MABT	62.928	+2.492	34.3	0	3.646	50.0
	CONFIRM	63.400	+2.965	47.1	0	3.174	45.1
100	No gate	65.438	+5.003	100.0	7	1.136	100.0
	Adv.-history exact	64.943	+4.508	54.3	0	1.630	91.6
	Paired MABT	64.624	+4.189	50.0	0	1.950	100.0
	CONFIRM	64.962	+4.527	55.0	0	1.611	91.6
200	No gate	66.450	+6.015	100.0	0	0.123	195.1
	Adv.-history exact	64.494	+4.058	64.3	0	2.080	164.7
	Paired MABT	65.724	+5.289	73.6	0	0.850	195.1
	CONFIRM	64.545	+4.109	67.9	0	2.029	164.7
400	No gate	66.571	+6.136	100.0	0	0.003	325.7
	Adv.-history exact	65.423	+4.987	84.3	0	1.151	226.4
	Paired MABT	65.707	+5.272	72.1	0	0.866	325.7
	CONFIRM	65.632	+5.197	98.6	0	0.942	226.4

historical credit has a small marginal effect at this 5%-share. Against the Bonferroni marginal bounds gate, CONFIRM's accuracy advantage is 1.090 points at $b = 50$, 0.019 at $b = 100$, and zero at $b = 200, 400$; useful-authorization gains are 14.7, 0.9, 0, and 0 points. Against unconditional paired exact, the corresponding accuracy gains are 1.069, 0.175, 0.051, and 0 points. All three gates make zero observed false replacements with the same labels. This is an observed low-budget advantage, not evidence of universal dominance.

Table 21: Dataset-wise paired accuracy differences, CONFIRM minus each valid comparator, at $b = 50$. Brackets are paired 95% t intervals over 20 query repetitions within the fixed dataset; C/B counts CONFIRM-only/baseline-only authorizations. These intervals quantify draw variability, not generalization to new datasets.

	Bonferroni marginal bounds			Unconditional paired exact		
Dataset	Δacc.	95% interval	C/B	Δacc.	95% interval	C/B
ImageNet-A	3.247	$[0.55, 5.95]$	5/0	2.598	$[0.10, 5.09]$	4/0
ImageNet-R	0.742	$[-0.14, 1.62]$	3/0	0.894	$[-0.01, 1.80]$	4/0
ImageNet-Sketch	0.142	$[-0.15, 0.44]$	1/0	0.142	$[-0.15, 0.44]$	1/0
DTD	0.000	$[0.00, 0.00]$	0/0	0.000	$[0.00, 0.00]$	0/0
Food-101	0.229	$[-0.14, 0.59]$	2/0	0.063	$[-0.07, 0.20]$	1/0
Oxford Pets	0.000	$[0.00, 0.00]$	0/0	0.000	$[0.00, 0.00]$	0/0
EuroSAT	3.267	$[0.87, 5.66]$	6/0	3.784	$[1.31, 6.26]$	7/0

Five of seven dataset means are positive and two are zero against either comparator. Treating the seven datasets, rather than 140 repetitions, as the generalization units gives mean differences 1.090 and 1.069 points, medians 0.229 and 0.142, and 95% t intervals $[-0.30, 2.48]$ and $[-0.34, 2.48]$. Leave-one-dataset-out means remain positive (0.727–1.271 and 0.616–1.247), but the wide cluster intervals preclude a population-level significance claim.

All 18 CONFIRM-only decisions relative to Bonferroni marginal bounds and all 25 relative to unconditional paired exact occur under $m < P$; no comparator-only decision occurs. The partial-confirmation gains are 0.496/0.579 accuracy points and 6.45/8.96 useful-authorization points, whereas every census decision coincides.

Table 22: Budget-wise matched decisions. C is CONFIRM; B is the named baseline. "Both/C/B/Neither" partitions all 140 decisions. Useful and accuracy columns report $C - B$ in percentage points.

b	Census	m/P	Baseline	Both	C only	B only	Neither	Δacc.	Δuseful
50	19.3	41.9	Bonferroni marginal bounds	49	17	0	74	1.090	14.7
			Unconditional paired exact	49	17	0	74	1.069	14.7
100	22.9	53.3	Bonferroni marginal bounds	76	1	0	63	0.019	0.9
			Unconditional paired exact	74	3	0	63	0.175	2.8
200	36.4	79.0	Bonferroni marginal bounds	95	0	0	45	0.000	0.0
			Unconditional paired exact	90	5	0	45	0.051	3.7
400	97.9	99.8	Bonferroni marginal bounds	138	0	0	2	0.000	0.0
			Unconditional paired exact	138	0	0	2	0.000	0.0

Table 23: Partial confirmation versus census. Useful is conditional on a beneficial nominee; missed is forfeited positive gain per decision. All rows use identical labels within a stratum and have zero harm and false replacement.

Stratum	Method	Decisions	Acc.	Useful	Missed	Labels
$m < P$	Bonferroni marginal bounds	313	56.374	50.9	1.307	113.3
	Unconditional paired exact	313	56.290	48.4	1.390	113.3
	CONFIRM	313	56.869	57.3	0.811	113.3
$m = P$	Bonferroni marginal bounds	247	74.475	100.0	0.000	155.6
	Unconditional paired exact	247	74.475	100.0	0.000	155.6
	CONFIRM	247	74.475	100.0	0.000	155.6

Disclosed post-label sensitivities. Neither analysis below changes the sealed seven-dataset result. First, replacing the prespecified Bonferroni rule by Holm at joint level $0.04375 = 7(0.05/8)$ preserves the original error allocation over the seven retained datasets and increases CONFIRM's matched-history additions from 30 to 35, with no reverse or harmful decision. Second, increasing the selection share from $\rho = 0.05$ to 0.25 after seeing the primary result makes the offset more consequential, as predicted by Section 5.1, but is exploratory rather than prespecified confirmatory evidence.

Table 24: Post-label sensitivity on seven unaffected CLIP datasets. Accuracy and deployment rate compare label-matched adversarial-history exact/CONFIRM; Extra counts CONFIRM-only deployments out of 140 decisions. Every row has zero harm and zero reverse decisions.

Rule	b	History acc.	CONFIRM acc.	History/CONFIRM deploy	Extra
Holm, $\rho = .05$	50	63.391	63.421	46.4/47.9	2
	100	64.981	65.000	55.7/56.4	1
	200	64.494	64.611	64.3/72.9	12
	400	65.423	65.632	84.3/98.6	20
Bonferroni, $\rho = .25$	50	61.470	62.869	20.7/43.6	32
	100	63.181	64.907	43.6/67.1	33
	200	62.115	65.092	14.3/85.0	99
	400	62.285	65.477	14.3/96.4	115

Table 25: Candidate availability after the metadata-only ImageNet-V2 alignment. Counts are adapted candidates better/equal/worse than source. The oracle column is the best of source and 16 candidates.

Dataset	Source	Oracle	Better/equal/worse
ImageNet-A	44.63	57.91	16/0/0
ImageNet-R	74.12	80.66	14/1/1
ImageNet-Sketch	47.85	51.07	9/0/7
ImageNet-V2	62.79	69.24	16/0/0
DTD	44.14	46.78	16/0/0
Food-101	83.30	87.11	10/1/5
Oxford Pets	88.18	90.62	8/0/8
EuroSAT	40.82	51.86	10/1/5

Disclosed ImageNet-V2 representation defect. TorchVision `ImageFolder` orders directories named `0,1,...,999` lexicographically. TTA-VLM permutes its classifier into that order; Promp-

tAlign's classifier remained in numeric ImageNet order while its custom loader used the lexicographic order. Collection therefore misinterpreted the four PromptAlign output indices. The inverse of the pre-existing `imagenet_v_mask` is a deterministic permutation derived from class metadata without labels. It changes the four apparent accuracies from 1.367% to 69.043%, 69.043%, 69.238% and 68.457%. Because the defect was discovered after labels were opened and changes the evaluated prediction representation, we report this correction only as sensitivity analysis and exclude ImageNet-V2 from Table 20. The sensitivity retains all query seeds and decisions and leaves the qualitative safety result unchanged, but it is not a clean eight-dataset confirmatory experiment.

F LABEL-BLIND CINIC-10 AUDIT

Separation and freeze. CINIC-10 had not entered any preceding method development. The generation command projected only the parquet image column, decoded RGB pixels, removed all hashes found in the complete 60,000-image CIFAR-10 train/test reference, and obtained the expected 70,000 ImageNet-contributor images. NumPy `default_rng(20260825)` selected and partitioned 32,768 images. The source, normalization, online batch size 64, 16-model prefix, four budgets, 20 seeds, 5% exploration share, Holm family at joint level 0.05, success criteria, and a 10,000-draw unit bootstrap with seed 20260925 were written before reveal. EATA's Fisher anchor used only 2,000 source-reference images and source pseudo-labels. The prediction artifact contained logits, source features, image indices and model names, all finite, with no target labels or derived performance field. The reveal command verified the frozen integrity manifest before first accessing the label column.

Candidate and unit structure. The source scores 65.7135%. The 15 non-source accuracies span 62.7991–63.0188%; continual Tent at 10^{-3} is the best fixed adapter. It also wins unit 23 by one example (357 rather than 358 errors), while source is the unique oracle on every other unit. Every unit contains a harmful candidate, with its worst available loss ranging from 0.391 to 4.590 points. The union disagreement contains 236–296 examples (mean 264.4), the mean pairwise disagreement is 34.9, and a unit contains 4–10 distinct prediction functions (mean 6.03). Source dominance is therefore empirical, not a consequence of identical predictions.

Table 26: First-reveal summary across budgets. Gated entries report the range of mean deployment rates over the five selectors; all 12,800 repeated gated unit decisions are non-harmful. Ungated harm is the range of mean harmful decisions per 32-unit run over LURE, PPAT and Uniform; the defective frozen Model Selector terminal row is excluded.

b	Gated deploy range	Gated harm	Ungated harm/run	Worst ungated loss
50	0.00–0.00%	0	5.90–12.20	−4.395
100	0.00–0.00%	0	3.00–9.65	−4.590
200	0.00–0.00%	0	0.50–6.30	−4.590
400	0.00–3.12%	0	0.10–2.30	−3.711

Table 27: CINIC-10 at $b = 400$, averaged over 20 seeds. Harm is out of 640 repeated decisions; source accuracy is 65.714. Model Selector's ungated entry is the disclosed terminal-census repair, and SHR has no separate ungated implementation.

	Gated			Ungated		
Selector	Acc.	Deploy	Harm	Acc.	Deploy	Harm
Model Selector	65.717	3.12%	0	65.717	3.12%	0
Active Testing-LURE	65.714	0.78%	0	65.713	2.34%	2
PPAT	65.714	0.00%	0	65.571	8.75%	46
Uniform	65.714	0.00%	0	65.607	7.34%	38
SHR	65.714	0.00%	0	not evaluated		

At $b = 400$, unit-resampled 95% descriptive intervals for missed oracle gain are $[0, 0]$ points for gated Model Selector, $[0, 0.0069]$ for gated LURE, $[0, 0.0092]$ for gated PPAT, Uniform and SHR, $[0.0986, 0.1959]$ for ungated PPAT, and $[0.0694, 0.1512]$ for ungated Uniform. Units, not the 20 acquisition repetitions, are resampled.

Predeclared failure and post-reveal repair. Three criteria pass: ungated harm is observed, all gated selectors have zero harm, and no frozen object changes. Two fail: there is only one rather than four replaceable units, and not every gated selector deploys on 25% of units. Because source is uniquely best on 31 units, no zero-harm rule can exceed 3.125% deployment without selecting a harmful model; the failed threshold is not redefined after seeing this fact.

The frozen Model Selector helper recorded snapshots only when a requested budget was reached. For $b = 400 > |U|$, its fallback copied the earlier exploration snapshot rather than the terminal census. This affected only the ungated Model Selector row: the frozen JSON reports mean accuracy 64.787 and 10.65 harmful decisions per run, both excluded from the paper. A disclosed post-reveal terminal census selects the exact oracle on every unit, giving 65.7166%, 3.125% deployment and zero harm. Gated Model Selector was specified to nominate at the exploration snapshot and is unchanged, as are LURE, PPAT, Uniform, SHR, the frozen artifacts and the pass/fail verdict.

G QEC Generality Audits

G.1 Heterogeneous QEC Source Rotation

The released Willow archive contains 100,000 shots at each of seven evaluation durations for 14 spatial layouts under traditional calibration and the same 14 under calibration plus RL fine-tuning (Sivak et al., 2026). We concatenate durations within a condition-layout deployment unit rather than treating them as independent, and pair the two conditions when resampling the 14 layouts. The four-model portfolio consists of Sparse Blossom with the SI1000 prior, Tesseract with SI1000 and frequency-calibrated priors, and AlphaQubit2. Every source rotation uses the other three models as candidates.

The Sparse-Blossom-source protocol was hashed before its label-dependent run, but all three candidates then proved better on all 28 units. We preserve that result. The all-source factorial was written and hashed afterward, before any alternative-source gate result, and is labeled post hoc throughout. It creates 112 source-unit tasks, of which 84 have an improving candidate and 28 do not. For every budget and Monte Carlo repetition Holm is applied once across all 112 nominee tests; All-candidate exact corrects 336 tests. Seed repetitions are averaged within task before any layout resampling.

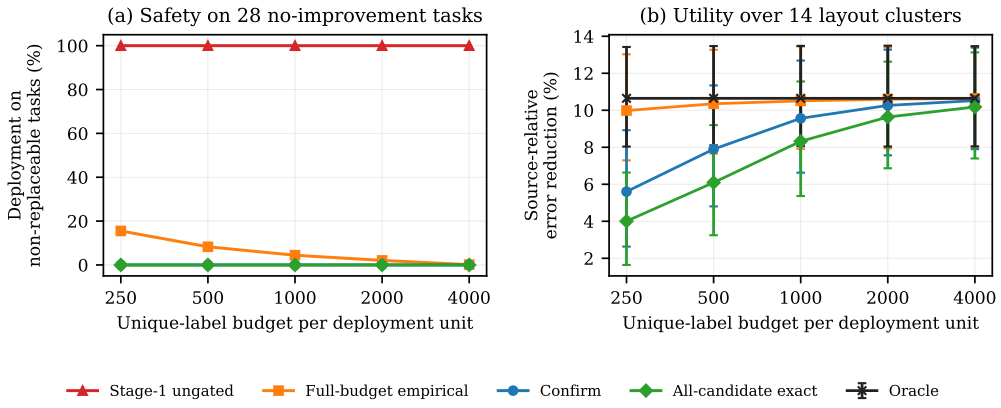

Figure 8: All-source heterogeneous QEC rotation. Left: replacement on the 28 tasks whose incumbent is already full-pool optimal. Right: mean utility with 95% layout-cluster bootstrap intervals. Every point averages 50 repetitions over 112 source-unit tasks; the same four models rotate through source and candidate roles.

At $b = 250$, CONFIRM's layout-cluster interval for relative error reduction is $[2.63, 8.92]\%$; at $b = 4{,}000$ it is $[7.90, 13.39]\%$. Source-wise, CONFIRM's deployment at the two budgets is respectively $50.3 \to 100.0\%$ for Sparse Blossom, $35.7 \to 99.6\%$ for Tesseract-SI1000, $11.8 \to 90.6\%$ for Tesseract-frequency, and $0 \to 0\%$ for AlphaQubit2, with zero harm in every source stratum. Thus the result is not driven only by the weakest incumbent, while the strongest incumbent is retained.

Table 28: Heterogeneous QEC source rotation. Harm is a decision whose selected decoder has higher full-pool shot error than its incumbent, out of 5,600 decisions per cell. Red. is mean source-relative error reduction.

b	Method	Deploy	Harm	Red.	Repl. deploy	Non-repl. deploy
250	Stage-1 ungated	100.0%	1,660	6.74%	100.0%	100.0%
250	Full-budget empirical	75.0%	251	9.98%	94.8%	15.5%
250	CONFIRM	24.4%	0	5.60%	32.6%	0.0%
250	All-candidate exact	14.4%	0	4.01%	19.2%	0.0%
500	Stage-1 ungated	100.0%	1,530	7.70%	100.0%	100.0%
500	CONFIRM	41.1%	0	7.90%	54.8%	0.0%
1,000	Stage-1 ungated	100.0%	1,444	8.23%	100.0%	100.0%
1,000	CONFIRM	57.6%	0	9.56%	76.8%	0.0%
2,000	Stage-1 ungated	100.0%	1,409	8.50%	100.0%	100.0%
2,000	CONFIRM	67.6%	0	10.26%	90.2%	0.0%
4,000	Stage-1 ungated	100.0%	1,402	8.63%	100.0%	100.0%
4,000	Full-budget empirical	75.0%	2	10.64%	99.9%	0.14%
4,000	CONFIRM	72.6%	0	10.52%	96.8%	0.0%
4,000	Oracle	75.0%	0	10.65%	100.0%	0.0%

The source-rotation protocol, implementation, frozen result, and evaluation manifest are integrity-recorded; full digests and the structurally trivial Sparse-Blossom-source run are preserved in the research archive, not bundled in the compact supplement. The public dataset contains multiple layouts, distances, durations and control conditions, but it does not convert them into independent future devices or acquisition dates.

G.2 ETH QEC Protocol, Scope, and Deviation Audit

The external syndrome-correlation data archive accompanies Remm et al. (2026); its DOI is `https://doi.org/10.3929/ethz-b-000733332`. The data archive, frozen protocol, corrected result, and author graph are integrity-recorded.

The first reveal exposed a global logical-frame sign error: the parser assumed bare product parity zero represented the positive encoded eigenstate, whereas all four one-cycle prepared states require the opposite convention. This decoder-independent parser correction changed only the common outcome mapping and was applied identically to both arms; it therefore cannot bias the comparison toward either decoder. Features, predictions, canary indices, $K = 2$ and $\alpha = 0.025$ were retained; the last is the Bonferroni allocation of family level 0.05 across the predeclared X- and Z-basis decisions. PACS, DomainBed and CINIC-10 instead use family level 0.05 directly. Both pre-correction and corrected outputs are preserved. With the author graph, source-only/adapted-only counts remain favorable: 730/445 for Z ($p = 4.19 \times 10^{-17}$) and 716/478 for X ($p = 2.97 \times 10^{-12}$). The NPZ files contain only labels, rounds and data; ZIP packaging on 2025-04-23 and repository deposit on 2025-05-02 are not acquisition dates, so the experiment establishes transfer to an external device and campaign, not generalization across dates. It also uses a declared phenomenological matching support rather than the source paper's circuit-level correlated decoder.

H Artifact and Implementation Notes

Reproducibility structure. Prediction generation, label reveal, and evaluation are separated wherever required by the frozen protocols. Locked and post-hoc analyses consume the same frozen prediction artifacts without modifying the locked package. Exact-test validation enumerates small finite populations and arbitrary acquisition offsets, while selector-level tests check deterministic and valid budgeted selections. The reported PACS, post-selection, QEC, bootstrap, and acquisition-share analyses are auditable from the preserved outputs. The anonymized supplement contains the method implementation, the compact CLIP prediction and label inputs, the frozen protocols, and the tests needed to reproduce the main-text CLIP table and certifiability figure. The package is a compact statistical replay, not a self-contained end-to-end reproduction of every benchmark: full neural and QEC regeneration requires external data, weights, and dependencies, while replay of historical sealed runs also requires the corresponding prediction artifacts and freeze records.

Corrected boundary case. An audit found that the original confirmation routine could return NaN when the fresh draw contained no nominee win and the compatible informative population was degenerate. Mathematically, if a compatible null exists then $\Pr(X \geq 0) = 1$; if none exists, the p-value is zero. The submitted routine now implements these two branches explicitly. Exhaustive small-population tests cover $W_m = 0$, $P = 0$, censuses, both offset signs, and absence of a compatible null. They separately validate the marginal-bound coverage and the unconditional paired exact comparator by enumerating every small finite-population null. The corrected NaN branch occurred zero times in the 1,280 shipped races and 29,440 matched-harness calls; replay leaves every reported decision unchanged. Original frozen outputs are retained for provenance, while the supplement ships the corrected routine.