SAFE PROBABILITY EDITING IS A RISK ALLOCATION PROBLEM

Anonymous authors
Paper under double-blind review

ABSTRACT

Under distribution shift, adaptation, calibration, and pooling can propose useful probability edits before target labels are available. The deployment question is how a fixed unlabeled cohort should spend a limited risk budget, not only whether every row passes the same pointwise check. Uniform clipping assigns every input the same allowance and therefore ignores large differences in candidate gain per unit risk. We show that an elementary worst-label identity has a consequential operational interpretation: all-outcome regret becomes a conditional-risk currency, turning label-free probability editing into hierarchical cohort allocation. Each row preserves the anchor action and obeys a local catastrophe cap, while a shared ledger distributes fixed worst-case Brier and log-regret budgets across the cohort. Because realized predictive utility is unavailable at edit time, the ledger minimizes intervention from an externally proposed candidate subject to this safety contract. The resulting convex program reduces to two risk prices that decouple the cohort into one-dimensional edits. On 10,000 ImageNetV2 predictions, the same mean budgets and a twofold local cap raise retained update from 19.9% to 30.0% and reduce total squared candidate distance from 176.1 to 123.4, with zero action changes. Across 20 frozen VLCS domain–seed conditions, retained update rises from 14.9% to 22.2% (paired gain 7.30 points; 95% CI 5.94–8.66). Under identical twofold local caps, the ledger achieves the smallest squared candidate-distance objective among five allocation rules. Exact pointwise safe bodies and sparse hulls provide supporting local geometry and scalable extensions. The guarantee is deterministic for the cohort in hand rather than a future-sample population claim. Together, the results show that safety need not mean uniform restraint: risk can be spent where it preserves the most candidate signal, without target labels or action changes.

1 INTRODUCTION

Probability estimates increasingly act as interfaces between learned models and downstream decisions. Under distribution shift, adaptation, calibration, ensemble, and pooling methods can all propose alternative probability vectors (Wang et al., 2021; Niu et al., 2022; 2023; Zhao et al., 2023; Alfarra et al., 2024). Deployment then faces a distinct question: before target labels arrive, how much of each proposed edit should be trusted? This question matters because an edit can preserve average accuracy while changing a deployed action or removing nearly all probability from an outcome that remains possible.

Existing safeguards provide two complementary but incomplete answers. Accuracy- and order-preserving calibration protects decisions or rankings (Rahimi et al., 2020; Esaki et al., 2024; Yang et al., 2023; Zhao et al., 2026), yet complete ranking preservation still permits unbounded log-loss deterioration. Calibration-set, conformal, and online risk controls provide statistical guarantees for future observations (Angelopoulos et al., 2024; Blot et al., 2025; Almeida et al., 2025), but they do not construct the pre-outcome edit applied to the probability vectors already in hand. A pointwise all-outcome cap closes this safety gap, but uniform clipping treats cohort heterogeneity as irrelevant. The same total risk can preserve substantially more candidate movement when it is allocated selectively, provided no individual edit can absorb catastrophic risk.

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.

The key step is conceptual rather than algebraically elaborate. For each row, worst-case conditional regret over all label distributions equals the regret of its worst label. For a fixed cohort, separability turns the average of these quantities into a shared risk budget with a precise conditional-risk meaning. The guarantee is therefore deterministic and cohort-specific: it protects the unlabeled probability vectors currently being edited, not future draws from an unknown population.

We realize this formulation with a hierarchical risk ledger (Figure 1). The Pathwise-Safe Prediction Layer (PSPL) first computes how far each anchor may move toward its candidate while preserving the anchor action and satisfying local Brier and log-loss caps. The ledger then spends a shared cohort budget across these admissible rays. At edit time, target labels are absent, so directly optimizing realized Brier, NLL, or accuracy would require label leakage or additional statistical assumptions. Once an external candidate generator has proposed a direction, squared candidate distance is instead a natural minimum-intervention objective: retain as much proposed signal as possible while satisfying the declared safety contract. Two dual risk prices reduce the convex cohort problem to independent scalar edits.

This work makes three contributions. First, we formulate pointwise safety as a local–cohort contract whose budget equals worst-case average conditional regret under arbitrary row-specific label distributions. Second, we derive an exact two-price allocator and show that it weakly dominates uniform clipping for the declared minimum-intervention objective. Third, matched-budget experiments on frozen ImageNetV2 and VLCS cohorts recover approximately 50% more candidate update with zero action changes; matched-cap comparisons and million-row audits isolate the allocation effect and verify the contract at scale. The exact Brier–log safe body, Pathwise-Safe Prediction Hulls (PSH), and adaptive projection are supporting local-geometric and scalable extensions of this main allocation principle.

2 Related work

Adaptation and calibration. Tent minimizes entropy online (Wang et al., 2021); EATA adds filtering and anti-forgetting regularization (Niu et al., 2022); SAR targets reliability under severe shift (Niu et al., 2023). Large-scale evaluations document difficult unsupervised model selection and strong dependence on the starting model (Zhao et al., 2023); time-aware evaluation further favors lightweight methods in some regimes (Alfarra et al., 2024). Our framework begins after model selection: it accepts an anchor and candidate distribution and adds an $O(K)$ certified editing layer. Modern networks are often miscalibrated (Guo et al., 2017). Order-preserving calibration (Rahimi et al., 2020; Zhao et al., 2026) protects rankings but does not bound all-label proper-score regret. Recent accuracy-preserving approaches include Simplex Temperature Scaling (Esaki et al., 2024), semantic grouping (Yang et al., 2023), and Top-versus-All confidence calibration for many classes (Le Coz et al., 2024); LaSCal instead performs label-free target calibration under an explicit label-shift assumption (Popordanoska et al., 2024b). These methods learn calibration maps or scalar confidences. By contrast, PSPL edits between available full probability vectors and certifies anchor-relative harm for every hypothetical label. Because calibration scores alone can obscure generalization trade-offs, Chidambaram & Ge (2025) recommend reporting predictive metrics such as NLL alongside calibration metrics; accordingly, our primary utility measures are the proper Brier and log scores, while ECE remains descriptive.

Aggregation and proper-loss geometry. Probabilistic forecast pooling (Gneiting & Ranjan, 2010) and mixability (van Erven et al., 2011) motivate borrowing information from experts. Convex aggregation can improve robustness under misspecification (Marsden et al., 2021), and recent kernel-embedding pools optimize proper scores from past data (Allen et al., 2025); neither imposes per-input, anchor-relative constraints. The supplement of Yang et al. (2023) already notes that averaging accuracy-preserving group calibrators preserves the predicted class. Our hull theorem adds simultaneous convex-loss sublevel constraints, while Proposition 4 and Theorem 6 characterize strict off-ray enlargement and recovery of the global safe set. U-calibration controls regret for every bounded proper loss in an online sequence (Kleinberg et al., 2023; Luo et al., 2024), while decision calibration, omniprediction, and model reconciliation target unknown or specified downstream agents (Du et al., 2025; Gibbs & Tibshirani, 2026; Tang et al., 2025). These population or sequential objectives are complementary to our pre-outcome guarantee on one edited vector. Proper calibration errors admit Bregman-divergence structure (Popordanoska et al., 2024a), motivating our convex

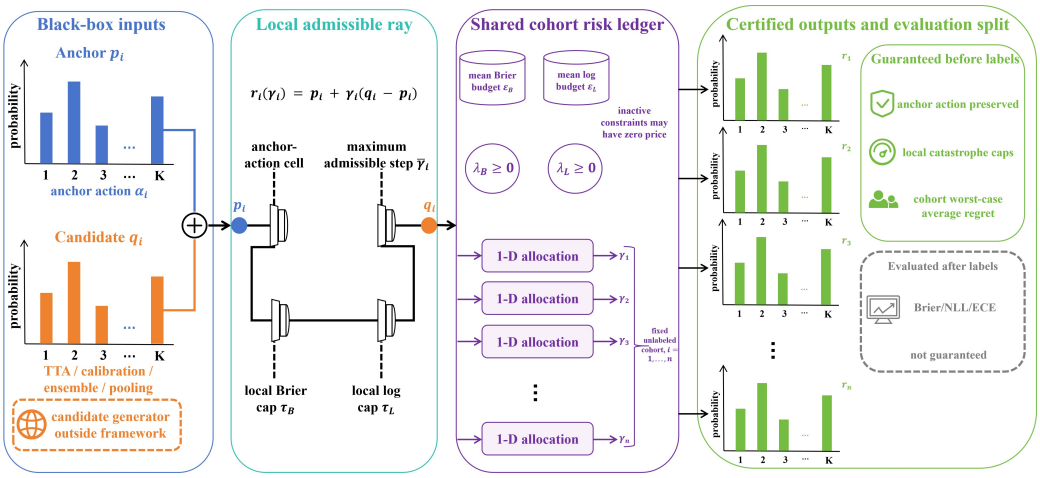

Figure 1: Hierarchical risk-ledger interface. A black-box candidate defines a direction from the anchor; local action and loss caps bound each row; two shared risk prices allocate a fixed cohort budget; labels enter only subsequent empirical evaluation.

prediction-space geometry. Post-hoc transformations can even reverse base-model rankings (Ranjan et al., 2024), reinforcing our separation of model selection from certified editing.

Sparse convex-hull optimization. Frank–Wolfe methods optimize over atomic convex hulls using linear minimization, sparse iterates, and primal–dual gap certificates (Jaggi, 2013). Recent work studies alternative extreme-point oracles and structured approximate oracles (Garber & Wolf, 2021; Zhou & Sun, 2022). Frank–Wolfe supplies the optimization scaffold; the new object is a prediction-specific safe atom family with closed-form simplex-vertex endpoints and a closed-form linear oracle. Coupling these atoms to all-label safe-body geometry and exact tie handling makes certified adaptive projection practical at $K = 1000$ with an $O(K + M)$ oracle.

Risk control and model selection. Conformal risk control bounds future expected losses from calibration data (Angelopoulos et al., 2024); adaptive conditional control learns input-dependent calibration classes (Blot et al., 2025), while importance weighting extends high-probability control under covariate shift (Almeida et al., 2025). These methods use labeled calibration evidence to make population-level statistical statements. The ledger occupies a complementary layer: it allocates a deterministic anchor-relative budget over the fixed unlabeled cohort being edited and remains valid for arbitrary row-specific conditional label distributions. Guided verification, online conformal ensembles, and certified UDA selection address complementary configuration or model-selection risks (Laufer-Goldshtein et al., 2024; Hajihashemi & Shen, 2024; Hu et al., 2024). Appendix Table 2 isolates the contract studied here from population, online, and calibration-map guarantees.

3 Safe Probability Editing as Cohort Risk Allocation

3.1 Outcome-wise Regret Yields Conditional Robustness

Let $p, q, r \in \Delta^{K-1}$ denote an anchor distribution, a black-box candidate, and the returned safe edit. For a predictive loss ℓ, define the outcome-wise regret vector $\delta(r, p) \in \mathbb{R}^K$ by $\delta_y = \ell(r, y) - \ell(p, y)$. The shield imposes $\delta_y \leq \epsilon$ for every outcome before its label is observed.

Theorem 1 (Conditional distributional robustness)**.** *For every p, r and loss ℓ,*

$$\sup_{\nu \in \Delta^{K-1}} \{\mathbb{E}_{Y\sim\nu}\ell(r, Y) - \mathbb{E}_{Y\sim\nu}\ell(p, Y)\} = \max_y \delta_y(r, p). \tag{1}$$

Hence the all-outcome budget is equivalent to an anchor-relative conditional-risk bound for every possible target label distribution ν, without an assumption on $P(Y \mid X = x)$.

The proof is immediate because a linear function on the simplex is maximized at a vertex. Its value is operational: the identity turns an outcome-wise checklist into a conditional-risk budget that can be compared and allocated across inputs.

The action constraint is equally general. Let a finite downstream system choose $a_p \in \arg\min_{a\in\mathcal{A}}\langle L_a, p\rangle$. Under anchor-favoring ties, preserving this anchor action requires

$$\langle L_{a_p} - L_a, r\rangle \le 0 \qquad \forall a \in \mathcal{A}, \tag{2}$$

and we denote this intersection of linear halfspaces by D_{a_p}; a deterministic rule that favors a challenger at equality replaces the corresponding boundary by a strict halfspace, realized numerically as in Section 3.2. Top-1 classification is the special case $L(a, y) = \mathbf{1}\{a \neq y\}$, which gives $r_c \ge r_j$ for anchor class c. Thus replacing top-1 by any finite cost-sensitive action rule leaves the convex safe-body, projection, and hull arguments unchanged.

Theorem 2 (Decision preservation is insufficient)**.** *For any strictly positive $p \in \Delta^{K-1}$, $K \ge 2$, and $M > 0$, there exists q with the same unique top class and the same complete strict probability ordering as p, yet $-\log q_j + \log p_j > M$ for some outcome j.*

Choose the lowest-probability coordinate j, set $0 < q_j < p_j e^{-M}$, and transfer the removed mass to the top coordinate. The ordering is preserved while the hypothetical-label log regret exceeds M. This separates the proposed outcome-robust contract from accuracy- or order-preserving calibration.

3.2 A HIERARCHICAL LEDGER SPENDS RISK ACROSS THE COHORT

Uniform pointwise budgets protect every input identically, even when candidate benefit per unit risk is highly heterogeneous. For a cohort of n edits, define $\rho_{i,s}(r_i; p_i) = \max_y\{\ell_s(r_i, y) - \ell_s(p_i, y)\}$. We impose both a local catastrophe cap $\rho_{i,s} \le \tau_s$ and a cohort budget $n^{-1}\sum_i \rho_{i,s} \le \epsilon_s$, with $\tau_s \ge \epsilon_s$.

Theorem 3 (Cohort outcome robustness)**.** *For arbitrary, row-specific conditional label distributions $\nu_i \in \Delta^{K-1}$,*

$$\sup_{\nu_1,\dots,\nu_n} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{Y_i\sim\nu_i}[\ell_s(r_i, Y_i) - \ell_s(p_i, Y_i)] = \frac{1}{n}\sum_{i=1}^{n} \rho_{i,s}(r_i; p_i). \tag{3}$$

Thus the cohort budget is a deterministic worst-case average conditional-regret guarantee, while τ_s prevents risk concentration on one input.

The objective separates over i, and each linear maximization over ν_i is attained at its worst label; no exchangeability or target-shift model is used.

The ledger chooses how much of each candidate ray to retain. Let $r_i(\gamma_i) = p_i + \gamma_i(q_i - p_i)$ and let $\bar{\gamma}_i$ be the maximum step satisfying the action and local caps. For Brier and log loss, write $d_i = q_i - p_i$ and

$$b_i(\gamma) = a_i\gamma + c_i\gamma^2, \tag{4}$$

$$a_i = \frac{2(\langle p_i, d_i\rangle - \min_y d_{iy})}{K}, \tag{5}$$

$$c_i = \frac{\|d_i\|_2^2}{K}, \tag{6}$$

$$l_i(\gamma) = -\log(1 - s_i\gamma), \tag{7}$$

$$s_i = \max_y \frac{(p_{iy} - q_{iy})_+}{p_{iy}}. \tag{8}$$

The minimum-intervention ledger is defined by the objective

$$\min_{\gamma_1,\dots,\gamma_n} \sum_i \|d_i\|_2^2(1 - \gamma_i)^2, \tag{9}$$

subject to the separately numbered feasibility conditions

$$\frac{1}{n}\sum_i b_i(\gamma_i) \le \epsilon_B, \tag{10}$$

$$\frac{1}{n}\sum_i l_i(\gamma_i) \le \epsilon_L, \tag{11}$$

$$0 \le \gamma_i \le \bar{\gamma}_i \quad (i = 1, \dots, n). \tag{12}$$

Proposition 1 (Convex allocation, decomposition, and dominance)**.** *Program 9 is convex and has a unique edited probability for every nonzero candidate direction. Strong duality holds for positive budgets. For dual prices $\lambda_B, \lambda_L \ge 0$, its rows independently minimize*

$$\|d_i\|_2^2(1-\gamma)^2 + \lambda_B b_i(\gamma)/\epsilon_B + \lambda_L l_i(\gamma)/\epsilon_L, \qquad 0 \le \gamma \le \bar{\gamma}_i, \tag{13}$$

whose monotone derivative has a scalar root. If $\tau_s \ge \epsilon_s$, uniform pointwise clipping is feasible for the ledger, so the ledger's total squared candidate distance cannot be larger.

Convexity follows from Equations 4–7; the anchor is a Slater point for positive budgets. The implementation optimizes only the two nonnegative prices and solves all row roots by vectorized bisection, followed by a deterministic feasibility repair. Its cost is $O(nK + nT)$ for T inner iterations, labels are absent from the interface, and setting $\tau_s = \epsilon_s$ recovers the pointwise shield.

3.3 Closed-form pathwise realization

Let the anchor $p \in \mathrm{int}(\Delta^{K-1})$ satisfy $p_j > 0$ for every class, let $q \in \Delta^{K-1}$, define $d = q - p$, and let $r(\gamma) = p + \gamma d$ for $\gamma \in [0, 1]$. For budgets $\epsilon_B, \epsilon_L \ge 0$, require for every label y:

$$\mathcal{B}(r(\gamma), y) - \mathcal{B}(p, y) \le \epsilon_B, \tag{14}$$

$$-\log r_y(\gamma) + \log p_y \le \epsilon_L, \tag{15}$$

where $\mathcal{B}(s, y) = K^{-1}\sum_j (s_j - \mathbf{1}\{j = y\})^2$. The implementation must also predict the same class from $r(\gamma)$ and p. Strict positivity makes anchor-relative log regret well-defined. Equivalently, the log constraint requires $r_y \ge \exp(-\epsilon_L)p_y$ for every hypothetical label: at $\epsilon_L = .1$, no class probability may fall below 90.48% of its anchor value. Likewise, $K\epsilon_B = .05$ caps the increase in squared distance to every one-hot outcome by $.05$. The implementation validates and renormalizes inputs in float64 and applies a documented positive floor before renormalization; guarantees are relative to this returned, possibly floored anchor rather than the unprocessed input.

3.4 Brier endpoint

Expanding the Brier constraint gives

$$\mathcal{B}(p + \gamma d, y) - \mathcal{B}(p, y) = \gamma\frac{2(\langle d, p\rangle - d_y)}{K} + \gamma^2\frac{\|d\|_2^2}{K}. \tag{16}$$

Thus the worst label yields

$$A = \frac{2(\langle d, p\rangle - \min_j d_j)}{K}, \tag{17}$$

$$B = \frac{\|d\|_2^2}{K}. \tag{18}$$

and worst regret $A\gamma + B\gamma^2$. If $A + B \le \epsilon_B$, set $\gamma_B = 1$; otherwise use the stable root

$$\gamma_B = \frac{2\epsilon_B}{A + \sqrt{A^2 + 4B\epsilon_B}}, \tag{19}$$

with the zero-B case treated linearly.

3.5 LOG-LOSS AND DECISION ENDPOINTS

Only coordinates with $d_j < 0$ can worsen log loss, giving

$$\gamma_L = \min\left\{1, \min_{j:d_j<0} \frac{p_j(1-\exp(-\epsilon_L))}{-d_j}\right\}. \tag{20}$$

Let $c = \arg\max_j p_j$ under a fixed tie rule. For each challenger with $d_j > d_c$, decision preservation requires

$$\gamma \le \frac{p_c - p_j}{d_j - d_c}. \tag{21}$$

The minimum of these ratios and one is γ_D. If ties favor c, the interval is closed. Under NumPy's lowest-index rule, equality with a lower-index challenger is excluded; our implementation returns the largest float64 step whose constructed output retains c, using an adjacent-float correction and rare-row binary search.

Theorem 4 (Pathwise-safe endpoint)**.** *For fixed p, q, the linear path, budgets, and anchor-favoring ties,*

$$\gamma^\star = \min\{1, \gamma_B, \gamma_L, \gamma_D\} \tag{22}$$

is the maximum feasible step. For a deterministic rule excluding a boundary tie, this is the supremum and the implementation returns its largest machine-representable feasible predecessor.

Proof. The worst-label Brier regret is nondecreasing because $A, B \ge 0$; the log constraint is monotone on each decreasing coordinate. Every pairwise decision margin is affine, so each anchor-containing feasible set is an interval. Their intersection begins at zero and ends at the minimum endpoint. Any larger step violates an active constraint; an excluded tie makes only that boundary open. □

3.6 EXACT GLOBAL BODY AND MINIMUM-INTERVENTION PROJECTION

The endpoint is useful but restricted to one declared direction. We first characterize the complete feasible body, independently of rays. Let $\{\ell_s\}_{s=1}^S$ be any finite family of losses for which $\ell_s(\cdot, y)$ is convex on the simplex for every label y, with budgets $\epsilon_s \ge 0$. Along any affine candidate ray, each composition $\gamma \mapsto \ell_s(p + \gamma(q-p), y)$ is convex. Because its budget sublevel set contains $\gamma = 0$, intersecting all label, loss, and decision constraints yields an anchor-containing interval; its maximal endpoint is therefore obtainable by one-dimensional root finding even when no analytic formula is available. The Brier/log instance above is distinguished by closed-form endpoints rather than by a loss-specific safety principle. Using the anchor-action cell D_{a_p} from Equation 2, define

$$\mathcal{C}(p) = \left\{r \in \Delta^{K-1} : r \in D_{a_p}, \quad \ell_s(r, y) - \ell_s(p, y) \le \epsilon_s \ \forall s, y\right\}, \tag{23}$$

with strict boundary exclusions when required by the deterministic action tie rule.

For the Brier–log instance, write $D_c = \{r : r_c \ge r_j \ \forall j\}$ and $\alpha = \exp(-\epsilon_L)$. The abstract safe set then has an explicit geometry.

Proposition 2 (Exact compact Brier–log safe body)**.** *For anchor-favoring ties,*

$$\mathcal{C}_{BL}(p) = \Delta^{K-1} \cap D_c \cap \{r : r \ge \alpha p\} \cap \left\{r : \|r\|_2^2 - \|p\|_2^2 - 2\min_j(r_j - p_j) \le K\epsilon_B\right\}. \tag{24}$$

Equivalently, the log-safe set is the affine contamination simplex

$$\{r : r \ge \alpha p,\ \mathbf{1}^\top r = 1\} = \alpha p + (1-\alpha)\Delta^{K-1}, \tag{25}$$

and the Brier condition has the one-cone representation

$$t \le r_j - p_j \qquad \forall j, \tag{26}$$

$$\|r\|_2^2 \le \|p\|_2^2 + K\epsilon_B + 2t. \tag{27}$$

Thus the exact formulation uses one quadratic/rotated-SOC constraint plus linear inequalities, rather than K one-hot-centered cones.

The proof and equivalence to the K one-hot-centered balls are in Appendix A.4. For any declared candidate q, the global projection $\pi_{\mathcal{C}}(q) = \arg\min_{r \in \mathcal{C}_{BL}(p)} \|r - q\|_2$ therefore exists uniquely and is a label-free minimum-intervention repair. For log safety alone it is closed form:

$$\pi_{\mathcal{C}_L}(q) = \alpha p + (1-\alpha)\Pi_\Delta\left((q - \alpha p)/(1-\alpha)\right). \tag{28}$$

3.7 CERTIFIED INNER APPROXIMATIONS

The exact body provides the reference solution; sparse safe hulls make that geometry deployable per input. For declared directions $q_1, \ldots, q_M$, let $r_m^\star$ be the maximal safe ray endpoint and define

$$h(w) = w_0 p + \sum_{m=1}^{M} w_m r_m^\star, \tag{29}$$

$$\mathcal{P}(p) = \operatorname{conv}\{p, r_1^\star, \ldots, r_M^\star\}. \tag{30}$$

Theorem 5 (Anchor-augmented safe polytope)**.** *For losses convex in the predictive distribution and endpoints certified against the same anchor, budgets, and tie rule,*

$$\bigcup_{m=1}^{M} [p, r_m^\star] \subseteq \mathcal{P}(p) \subseteq \mathcal{C}(p). \tag{31}$$

Thus projection onto $\mathcal{P}(p)$ is a label-free, minimum-intervention inner approximation.

The result follows from convexity of the simplex, action cell, and loss sublevel sets; Appendix B reports multi-direction diagnostics.

Adaptive PSH avoids enumerating the full hull. With atom set $\mathcal{A}(p)$ containing the anchor, model-ray endpoints, and the K closed-form simplex-vertex endpoints, Frank–Wolfe uses

$$s_t \in \arg\min_{s \in \mathcal{A}(p)} \langle x_t - q, s \rangle, \tag{32}$$

$$\eta_t = \Pi_{[0,1]} \frac{\langle q - x_t, s_t - x_t \rangle}{\|s_t - x_t\|_2^2}, \tag{33}$$

$$x_{t+1} = (1 - \eta_t) x_t + \eta_t s_t. \tag{34}$$

Proposition 3 (Adaptive PSH certificate)**.** *Every iterate is safe, decreases squared candidate distance, uses at most $t+1$ atoms, and has computable gap $G_t = \langle x_t - q, x_t - s_t \rangle$ satisfying $0 \le f(x_t) - f^\star \le G_t$. The exact linear oracle costs $O(K + M)$ per iteration.*

Scope of the approximation. Two linearly independent safe directions strictly enlarge either ray alone; we record this elementary fact as Proposition 4 for reference.

Proposition 4 (Strict off-ray enlargement)**.** *If $a - p$ and $b - p$ are linearly independent safe directions, every strict mixture of a and b is safe but lies on neither ray.*

For a unique anchor action and positive budgets, radial endpoints recover the compact safe body as directions become dense.

Theorem 6 (Finite-direction approximation)**.** *Nested dense direction sets U yield inner hulls $\mathcal{P}_U$ with $d_H(\mathcal{P}_U, \mathcal{C}(p)) \to 0$.*

Proposition 5 (Candidate-projection gap)**.** *For every finite inner hull,*

$$0 \le \operatorname{dist}(q, \mathcal{P}_U) - \operatorname{dist}(q, \mathcal{C}(p)) \le d_H(\mathcal{P}_U, \mathcal{C}(p)). \tag{35}$$

Proofs and finite-precision tie handling are in Appendix A.3–A.5. This separation makes the interface modular: certification determines admissible movement, while any adaptation, calibration, or pooling system may supply the anchor and candidate.

4 EVIDENCE FOR HIERARCHICAL RISK ALLOCATION

4.1 THE SAFETY CONTRACT HOLDS AT SCALE

We first test whether the analytic contract survives finite precision and deployment-scale stress. Development, post-hoc, and frozen-confirmatory evidence are kept separate. The main log budget is $\epsilon_L = .1$, corresponding to at least 90.48% retention of every hypothetical-label probability. ImageNetV2 uses $\epsilon_B = .05/K$; VLCS uses $\epsilon_B = .005$. Appendix Figure 4 visualizes the log

interpretation, and every table states its dataset-specific Brier setting. Primary contract metrics are worst-outcome Brier/log regret, action violations, candidate distance, and exact-global capture. Brier, NLL, accuracy, and ECE are secondary realized-utility measures; labels never enter the shield, and ECE is descriptive rather than guaranteed (Chidambaram & Ge, 2025). One million random 10-class pairs produced zero decision changes and zero Brier or log-budget violations above 10^{-10}; maximum excesses were 7.75×10^{-17} and 3.33×10^{-16}. An additional million-row augmented-polytope stress test spans 3–65 classes, 2–8 candidates, near-zero probabilities, adversarial candidates, and random per-example anchor/endpoint weights; it records zero decision violations and maximum Brier/log harm of .00499938 and .09999627 at budgets .005 and .1. Ten API tests include anchor, ray, and endpoint-hull containment, random convex weights, strict lower-index tie handling, and large-class float32 normalization. Across 180,000 anchor-selection simulations over $K \in \{3, 10, 100\}$ and $n \in \{50, 200, 1000\}$, the finite-validation bound had zero violations, confirming implementation-wide feasibility across class counts and cohort sizes.

4.2 Risk allocation recovers safe candidate movement

Risk allocation recovers candidate movement that uniform safety budgets discard. We compare uniform pointwise clipping with the risk ledger at identical cohort budgets; the ledger allows at most twice each budget on one row. Figure 2a reports two frozen, label-free editing studies. On all 10,000 ImageNetV2 predictions, allocation raises retained update from 19.9% to 30.0%—a 50.7% relative gain—and increases squared-distance reduction from the anchor by 59.2%. The mean worst-label log regret remains .1, the local cap is .2, and NLL improves by .00251 with no decision change. Across 20 VLCS domain–seed conditions, retained update rises from 14.9% to 22.2%; the paired gain is 7.30 points (95% CI 5.94–8.66), and squared candidate distance falls by 25.98 (CI 19.44–32.52). Candidate fidelity under the deterministic contract is the primary endpoint because target labels are unavailable when the edit is constructed. Realized Brier and NLL are reported as downstream diagnostics; they move favorably here, although their 20-condition intervals do not resolve a nonzero mean change.

To separate budget relaxation from allocation quality, we also compare five rules under the *same* twofold local caps and cohort budgets: equal cap fraction, confidence priority, candidate-distance priority, marginal efficiency, and the exact ledger solution. On ImageNetV2 the ledger has the smallest total squared candidate distance (123.39); the closest marginal-efficiency heuristic reaches 123.91, a 0.42% gap. Across the 20 VLCS conditions, the paired ledger-minus-marginal difference is -0.346 (95% CI $[-0.532, -0.161]$). The same heuristic retains 0.11 percentage points more Euclidean update on average. This metric split is informative: the ledger optimizes squared candidate distance, whereas retained Euclidean update answers a different geometric question.

4.3 Local safety remains expressive

A useful allocator needs a local contract that rules out failures while leaving room to move. On all 10,000 ImageNetV2 predictions, decision-only projection retains 94.1% of the proposed update and preserves every top-1 action, yet its 95th/99th/99.9th percentiles of worst-outcome log regret are 1.015/1.061/1.085 and its NLL worsens from 1.362 to 1.371. Pointwise Brier–log clipping caps every row at .1, showing that action preservation cannot substitute for outcome safety.

The exact body and sparse hulls then test whether safety is geometrically expressive as well as protective. On 10,000 balanced VLCS predictions, a seven-direction hull captures 96.0% of global safe-projection distance reduction; 32-step Adaptive PSH captures 95.8% with 1.89 active directions on average. On one frozen ImageNetV2 prediction per class ($K = n = 1000$), 736 exact solutions are closed-form log-simplex projections and 264 use the compact one-cone program; Adaptive PSH capture rises from 86.9% at 16 steps to 94.9% at 256, with zero violations. Appendix Tables 1 and 4 report the full diagnostics.

Across 45 CIFAR-C corruption–architecture conditions, NLL improves in 45/45 and Brier in 44/45 under the pointwise shield, with every decision preserved. OfficeHome and VLCS seed-level means move in the same direction, while the wide four-domain intervals position these cross-domain results as supporting evidence; Appendix B reports complete baselines, intervals, ECE, efficiency, and candidate-hull ablations.

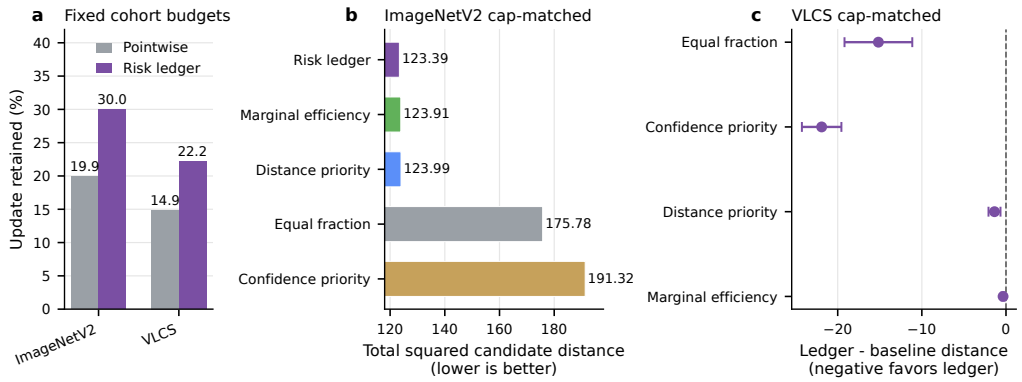

Figure 2: Hierarchical allocation results. (a) The ledger retains more candidate update than uniform pointwise clipping at fixed cohort budgets. (b,c) All allocation rules use identical twofold local caps and cohort budgets and have zero action changes. Panel b reports ImageNetV2 squared candidate distance (lower is better); panel c reports paired VLCS ledger-minus-baseline differences (negative favors the ledger), with 95% t intervals over 20 conditions. The marginal-efficiency heuristic is close but remains suboptimal for the declared objective.

5 Discussion

The contribution is a problem reformulation: the worst-label identity supplies a conditional-risk currency, and cohort separability makes that currency allocatable. Local caps block catastrophic row-level edits; two global prices then spend the remaining budget where it preserves the most candidate movement. The approximately 50% ImageNetV2 and VLCS gains, together with the matched-cap comparisons, show that allocation—not looser safety—recovers movement discarded by uniform clipping.

Candidate fidelity is the strongest label-free edit-time objective: the candidate proposes a direction and the ledger returns its closest certified edit. Anchor and candidate quality still govern realized utility, as the 1,000-class Adaptive PSH diagnostic shows; safe bodies and sparse hulls instead enrich the local geometry. The contract is deterministic and anchor-relative for the fixed unlabeled cohort; guarantees for future draws or candidate selection require a separate statistical layer.

6 Conclusion

Safe probability editing becomes less conservative when safety is treated as allocatable risk. A worst-label identity supplies the currency, and the two-price ledger spends it across a fixed unlabeled cohort. At matched budgets and caps, it retains about 50% more candidate update on two distribution-shift families without target labels or action changes.

Reproducibility Statement

The appendices provide complete derivations, tie handling, frozen protocols, corruption-level tables, budget grids, and implementation audits. The anonymous supplement includes the PSPL, risk-ledger, and Adaptive PSH implementations; unit tests; processed source tables; pinned dependencies; figure scripts; checksums; and commands that reproduce Figures 2–4. Raw public datasets and pretrained checkpoints are not redistributed, and the README identifies the external inputs required for full end-to-end reruns.

Ethics Statement

Medical results are methodological stress tests, not clinical validation or diagnostic claims. Clinical use would require privacy governance, subgroup analysis, and prospective validation.

AI USE STATEMENT

Generative AI was used solely for language polishing and improving the clarity and readability of the manuscript. The authors take full responsibility for the content of the submission.

REFERENCES

Motasem Alfarra, Hani Itani, Alejandro Pardo, Shyma Yaser Alhuwaider, Merey Ramazanova, Juan Camilo Perez, Zhipeng Cai, Matthias Müller, and Bernard Ghanem. Evaluation of test-time adaptation under computational time constraints. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 976–991, 2024. URL https://proceedings.mlr.press/v235/alfarra24a.html.

Sam Allen, David Ginsbourger, and Johanna Ziegel. Efficient pooling of predictions via kernel embeddings. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=dji9MfONNP.

Duarte C. Almeida, João Bravo, Jacopo Bono, Pedro Bizarro, and Mário A. T. Figueiredo. High probability risk control under covariate shift. In *Proceedings of the Fourteenth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 266 of *Proceedings of Machine Learning Research*, pp. 133–152, 2025. URL https://proceedings.mlr.press/v266/almeida25a.html.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/f3549ef9b5ff520a7e41ff3cc306ab2b-Abstract-Conference.html.

Vincent Blot, Anastasios Nikolas Angelopoulos, Michael Jordan, and Nicolas J.-B. Brunel. Automatically adaptive conformal risk control. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 19–27, 2025. URL https://proceedings.mlr.press/v258/blot25a.html.

Muthu Chidambaram and Rong Ge. Reassessing how to compare and improve the calibration of machine learning models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/9a30247fcca95299c2fd48d4667282de-Abstract-Conference.html.

Ally Du, Dung Daniel Ngo, and Zhiwei Steven Wu. Reconciling model multiplicity for downstream decision making. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/948552777302d3abf92415b1d7e9de70-Abstract-Conference.html.

Yasushi Esaki, Akihiro Nakamura, Keisuke Kawano, Ryoko Tokuhisa, and Takuro Kutsuna. Accuracy-preserving calibration via statistical modeling on probability simplex. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1666–1674, 2024. URL https://proceedings.mlr.press/v238/esaki24a.html.

Dan Garber and Noam Wolf. Frank–wolfe with a nearest extreme point oracle. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 2103–2132, 2021. URL https://proceedings.mlr.press/v134/garber21a.html.

Isaac Gibbs and Ryan J. Tibshirani. Sample-efficient omniprediction for proper losses. In *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 2679–2719, 2026. URL https://proceedings.mlr.press/v336/gibbs26a.html.

Tilmann Gneiting and Roopesh Ranjan. Combining probability forecasts. *Journal of the Royal Statistical Society: Series B*, 72(1):71–91, 2010. doi: 10.1111/j.1467-9868.2009.00726.x. URL https://academic.oup.com/jrsssb/article/72/1/71/7076442.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Erfan Hajihashemi and Yanning Shen. Multi-model ensemble conformal prediction in dynamic environments. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3769.

Dapeng Hu, Mi Luo, Jian Liang, and Chuan-Sheng Foo. Towards reliable model selection for unsupervised domain adaptation: An empirical study and a certified baseline. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4316. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f50cebc22663df45ce619645bfabb3b3-Abstract-Datasets_and_Benchmarks_Track.html.

Martin Jaggi. Revisiting Frank–Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 427–435, 2013. URL https://proceedings.mlr.press/v28/jaggi13.html.

Bobby Kleinberg, Renato Paes Leme, Jon Schneider, and Yifeng Teng. U-calibration: Forecasting for an unknown agent. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 5143–5145, 2023. URL https://proceedings.mlr.press/v195/kleinberg23a.html.

Bracha Laufer-Goldshtein, Adam Fisch, Regina Barzilay, and Tommi Jaakkola. Risk-controlling model selection via guided bayesian optimization. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=numGBcElus.

Adrien Le Coz, Stéphane Herbin, and Faouzi Adjed. Confidence calibration of classifiers with many classes. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2469. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/8df80d7115a55eb81010c967a247b1ae-Abstract-Conference.html.

Haipeng Luo, Spandan Senapati, and Vatsal Sharan. Optimal multiclass U-calibration error and beyond. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0241. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/0e4d695de4c606494ba9b0f3dac3b57a-Abstract-Conference.html.

Annie Marsden, John Duchi, and Gregory Valiant. Misspecification in prediction problems and robustness via improper learning. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2161–2169, 2021. URL https://proceedings.mlr.press/v130/marsden21a.html.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16888–16905, 2022. URL https://proceedings.mlr.press/v162/niu22a.html.

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. URL https://openreview.net/forum?id=g2YraF75Tj.

Teodora Popordanoska, Sebastian Gregor Gruber, Aleksei Tiulpin, Florian Buettner, and Matthew B. Blaschko. Consistent and asymptotically unbiased estimation of proper calibration errors. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3466–3474, 2024a. URL https://proceedings.mlr.press/v238/popordanoska24a.html.

Teodora Popordanoska, Gorjan Radevski, Tinne Tuytelaars, and Matthew B. Blaschko. LaSCal: Label-shift calibration without target labels. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. doi: 10.52202/079017-2088. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/783c5986e1d6112cb4688d9b2105609a-Abstract-Conference.html.

Amir Rahimi, Amirreza Shaban, Ching-An Cheng, Richard Hartley, and Byron Boots. Intra order-preserving functions for calibration of multi-class neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13456–13467, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/9bc99c590be3511b8d53741684ef574c-Abstract.html.

Rishabh Ranjan, Saurabh Garg, Mrigank Raman, Carlos Guestrin, and Zachary Lipton. Post-hoc reversal: Are we selecting models prematurely? In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2903.

Jingwu Tang, Jiayun Wu, Zhiwei Steven Wu, and Jiahao Zhang. Dimension-free decision calibration for nonlinear loss functions. *arXiv preprint arXiv:2504.15615*, 2025. doi: 10.48550/arXiv.2504.15615. URL https://arxiv.org/abs/2504.15615.

Tim van Erven, Mark D. Reid, and Robert C. Williamson. Mixability is bayes risk curvature relative to log loss. In *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pp. 233–252, 2011. URL https://proceedings.mlr.press/v19/vanerven11a.html.

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. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Jia-Qi Yang, De-Chuan Zhan, and Le Gan. Beyond probability partitions: Calibrating neural networks with semantic aware grouping. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2547.

Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42058–42080, 2023. URL https://proceedings.mlr.press/v202/zhao23d.html.

Lening Zhao, Qipeng Zhan, and Li Shen. Invertible logits transformation for accuracy-preserving post-hoc uncertainty calibration. *arXiv preprint arXiv:2608.10372*, 2026. doi: 10.48550/arXiv.2608.10372. URL https://arxiv.org/abs/2608.10372.

Baojian Zhou and Yifan Sun. Approximate Frank–Wolfe algorithms over graph-structured support sets. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27303–27337, 2022. URL https://proceedings.mlr.press/v162/zhou22i.html.

A Additional proofs

A.1 Hierarchical risk-ledger guarantees

For Theorem 3, separability gives

$$\sup_{\nu_1,\dots,\nu_n} \frac{1}{n}\sum_i \langle \nu_i, \delta_i\rangle = \frac{1}{n}\sum_i \sup_{\nu_i \in \Delta^{K-1}} \langle \nu_i, \delta_i\rangle. \tag{36}$$

Each inner objective is linear on the simplex and is maximized by a point mass on an outcome in $\arg\max_y \delta_{iy}$, proving the identity. The maximizing outcomes may differ across rows, so the statement does not rely on a common label distribution.

For Proposition 1, $a_i, c_i, s_i \geq 0$. Hence $b_i'(\gamma) = a_i + 2c_i\gamma \geq 0$ and $b_i''(\gamma) = 2c_i \geq 0$, while $l_i'(\gamma) = s_i/(1 - s_i\gamma) \geq 0$ and $l_i''(\gamma) = s_i^2/(1 - s_i\gamma)^2 \geq 0$ on the feasible interval. The objective

term $\|d_i\|_2^2(1-\gamma)^2$ is convex and strictly convex whenever $d_i \neq 0$. Box, action, local-cap, and cohort-budget constraints are convex. At $\gamma = 0$, both cohort regrets vanish; positive budgets therefore give Slater feasibility relative to the box constraints and strong duality.

With normalized cohort multipliers $\lambda_B, \lambda_L \geq 0$, the Lagrangian is a constant plus a sum of row functions

$$\phi_i(\gamma) = \|d_i\|_2^2(1-\gamma)^2 + \lambda_B b_i(\gamma)/\epsilon_B + \lambda_L l_i(\gamma)/\epsilon_L. \tag{37}$$

Its derivative

$$\phi_i'(\gamma) = 2\|d_i\|_2^2(\gamma - 1) + \frac{\lambda_B}{\epsilon_B}(a_i + 2c_i\gamma) + \frac{\lambda_L}{\epsilon_L}\frac{s_i}{1 - s_i\gamma} \tag{38}$$

is monotone nondecreasing, so the clipped scalar root is the unique row minimizer whenever $d_i \neq 0$. Finally, uniform pointwise clipping at ϵ_s obeys every local cap when $\tau_s \geq \epsilon_s$ and its mean regrets are at most ϵ_s; it is therefore feasible for Program 9. Global optimality proves the stated candidate-distance dominance.

A.2 Closed-form ray endpoints

For Brier score,

$$\mathcal{B}(p + \gamma d, y) - \mathcal{B}(p, y) = K^{-1}\{2\gamma\langle d, p - e_y\rangle + \gamma^2\|d\|^2\}. \tag{39}$$

Only $-d_y$ depends on y, so the worst label minimizes d_y. Moreover $A \geq 0$ because $\langle d, p\rangle$ is a convex combination of coordinates of d and is at least $\min_j d_j$. For log loss, coordinates with $d_y \geq 0$ cannot worsen; exponentiation and rearrangement on $d_y < 0$ gives the main-text endpoint. For decisions, every anchor-class margin $(p_c - p_j) + \gamma(d_c - d_j)$ is affine.

A.3 Closed-form Adaptive PSH oracle and certificate

For the ray from p to simplex vertex e_j, let $m_j = \max_{k \neq j} p_k$ and $c = \arg\max_k p_k$. Substitution into the ray formulas gives

$$A_j = \frac{2}{K}\big(p_j - \|p\|_2^2 + m_j\big), \tag{40}$$

$$B_j = \frac{1}{K}\big(1 - 2p_j + \|p\|_2^2\big). \tag{41}$$

so $\gamma_{B,j}$ is the same stable quadratic root as in Section 3.2 with $(A, B) = (A_j, B_j)$. All non-j coordinates contract by $1 - \gamma$, hence $\gamma_{L,j} = 1 - \exp(-\epsilon_L)$. The decision endpoint is

$$\gamma_{D,j} = \begin{cases} 1, & j = c, \\ (p_c - p_j)/(1 + p_c - p_j), & j \neq c, \end{cases} \tag{42}$$

with the largest floating-point strict predecessor when $j < c$ and equality would invoke the lower-index tie rule. Thus $\gamma_j = \min\{\gamma_{B,j}, \gamma_{L,j}, \gamma_{D,j}\}$ and $v_j^\star = (1 - \gamma_j)p + \gamma_j e_j$. For any gradient g, its oracle score is $\langle g, v_j^\star\rangle = (1-\gamma_j)\langle g, p\rangle + \gamma_j g_j$. Computing the largest and second-largest coordinates supplies every m_j in $O(K)$; scanning these scores and M stored model endpoints gives the stated $O(K + M)$ oracle without a $K \times K$ candidate tensor.

Proof of Proposition 3. Every atom lies in $\mathcal{C}(p)$ by the ray endpoint theorem, and Theorem 5 makes every convex update feasible. Exact line search includes $\eta = 0$, proving monotonicity, and introduces at most one atom per iteration. Let x^* minimize f over $\operatorname{conv}\mathcal{A}(p)$ and $g_t = \nabla f(x_t)$. Convexity and the linear-oracle definition give

$$f(x_t) - f(x^*) \leq \langle g_t, x_t - x^*\rangle \leq \langle g_t, x_t - s_t\rangle = G_t. \tag{43}$$

For $f(x) = \frac{1}{2}\|x - q\|_2^2$, the Frank–Wolfe curvature over a set of diameter D is at most D^2. The standard conditional-gradient rate is therefore $2D^2/(t+2)$ (Jaggi, 2013); exact line search cannot be worse than its prescribed step. □

A.4 GLOBAL SAFE BODY AND PROJECTION GAP

Proof of Proposition 2. For each hypothetical label y,

$$\mathcal{B}(r, y) - \mathcal{B}(p, y) = K^{-1}\big(\|r - e_y\|_2^2 - \|p - e_y\|_2^2\big), \tag{44}$$

so the original representation is an intersection of K closed balls. Expanding and maximizing over y instead gives

$$\max_y\{\mathcal{B}(r, y) - \mathcal{B}(p, y)\} = K^{-1}\left(\|r\|_2^2 - \|p\|_2^2 - 2\min_j(r_j - p_j)\right). \tag{45}$$

Introducing $t \le r_j - p_j$ for every j is exact because the largest feasible t equals the minimum displacement; the remaining quadratic inequality is convex and has a rotated-SOC representation.

The log constraints are $r \ge \alpha p$. If they hold, $s = (r - \alpha p)/(1 - \alpha)$ is nonnegative and sums to one, so $r = \alpha p + (1 - \alpha)s$ with $s \in \Delta^{K-1}$; the converse is immediate. Euclidean projection through this affine map yields the stated log-only formula. Finally, action preservation gives D_c for top-1 classification, or Equation 2 for a general finite action set. The intersection contains p and is convex and compact; strict convexity of squared Euclidean distance gives a unique global projection. □

Proof of Proposition 5. The lower bound follows from $\mathcal{P}_U \subseteq \mathcal{C}(p)$. Let $c^\star$ minimize distance from q to $\mathcal{C}(p)$ and set $\delta = d_H(\mathcal{P}_U, \mathcal{C}(p))$. There exists $z \in \mathcal{P}_U$ with $\|z - c^\star\|_2 \le \delta$, hence $\operatorname{dist}(q, \mathcal{P}_U) \le \|q - z\|_2 \le \operatorname{dist}(q, \mathcal{C}(p)) + \delta$. Nested hulls decrease distance by set inclusion, and Theorem 6 sends δ to zero for dense directions. □

A.5 PROOF OF THE RADIAL APPROXIMATION RESULT

Proof of Theorem 6. Work in the affine hull $L = \operatorname{aff}(\mathcal{C}(p)) - p$. Compactness makes each radial maximum $\rho(u)$ attainable. For any $x \in \mathcal{C}(p) \setminus \{p\}$, set $u = (x - p)/\|x - p\|$. Convexity and the definition of ρ give $x \in [p, e(u)]$; the reverse inclusion follows because both endpoints lie in $\mathcal{C}(p)$. Convexity also implies $\mathcal{P}_U \subseteq \mathcal{C}(p)$, and adding vertices proves monotonicity.

Because p is in the relative interior, $\mathcal{C}(p) - p$ contains a ball of radius $r > 0$ in L and is contained in a ball of radius $R < \infty$. The radial function of such a convex body is continuous on the unit sphere of L and hence uniformly continuous. If the largest angular gap of U_n tends to zero, every $e(u)$ has some $e(v_n)$, $v_n \in U_n$, converging to it uniformly. Writing any $x \in \mathcal{C}(p)$ as $x = (1 - \lambda)p + \lambda e(u)$ and replacing $e(u)$ by $e(v_n)$ produces a point in $\mathcal{P}_{U_n}$ converging uniformly to x. Since $\mathcal{P}_{U_n} \subseteq \mathcal{C}(p)$, this is exactly $d_H(\mathcal{P}_{U_n}, \mathcal{C}(p)) \to 0$. □

B SUPPORTING EMPIRICAL RESULTS

Table 1: Global-safe projection diagnostic on 10,000 balanced VLCS predictions. Capture is candidate-distance reduction relative to the exact global projection; labels do not construct the edits.

Safe set	Directions	$\|r - q\|_2 \downarrow$	Capture↑	Brier↓	NLL↓
Anchor	0	.25815	0%	.085885	.9898
Single model ray	1	.22534	79.3%	.084605	.9330
Two-model projected PSH	2	.22425	81.9%	.084406	**.9275**
Adaptive PSH (32 steps)	1.89 avg.	.21853	95.8%	.084401	.9306
+ simplex-vertex rays	7	.21843	96.0%	**.084401**	.9307
Global safe projection	SOCP	**.21677**	100%	.084405	.9300

Table 2: Guarantee interfaces. Each family addresses a distinct unit of guarantee.

Family	Labels to fit	Unit of guarantee	Decision kept	All-label harm cap
Accuracy-preserving calibration	Yes	map/confidence	Yes	No
Shift calibration	Source labels	target population	Yes	No
Decision/multicalibration	Yes	groups/agents	In agreement	No
Risk control/selection	Yes	population	No	No
Proper-score pooling	Yes	population/sequence	No	Loss-specific
PSPL/PSH	No at edit time	prediction vector	Yes	Yes

Table 3: ImageNetV2 matched-frequency ($K = 1000$, $n = 10{,}000$), with $\epsilon_L = .1$ and $\epsilon_B = .05/K$. Candidate order is locked by source metadata.

Method	Top-1	Top-5	Brier↓	NLL↓	ΔDec
ConvNeXt anchor	.7318	.9089	.0004114	1.3620	0
Mean-3	.7319	.9124	.0004208	1.3718	.1013
SafeRay-ViT	.7318	.9122	**.0004033**	**1.3172**	**0**
SafeRay-ResNet	.7318	.9104	.0004222	1.3895	**0**
Uniform PSH	.7318	.9115	.0004119	1.3481	**0**
SafeRay-Mean3	.7318	.9116	.0004104	1.3430	**0**
Adaptive PSH, model atoms (16)	.7318	.9118	.0004143	1.3545	**0**
Adaptive PSH, +1000 vertices (4)	.7318	.9081	.0004125	1.3329	**0**
Adaptive PSH, +1000 vertices (16)	.7318	.9106	.0004135	1.3351	**0**

Table 4: ImageNetV2 contract diagnostics from frozen predictions. Panel A uses all 10,000 examples; update retained is candidate-distance reduction from the anchor. Panel B uses one example per class ($n = K = 1000$); capture is relative to the compact exact global projection. No target label enters any projection.

A. Decision-only tail harm

Method	Update retained	Log q95	Log q99	Log q99.9	Max log	ΔDec
Unconstrained candidate	100.0%	1.021	1.065	1.087	1.096	.1013
Decision-only projection	94.1%	1.015	1.061	1.085	1.096	0
Joint safe ray	19.9%	.100	.100	.100	.100	0
Adaptive PSH (16)	37.4%	.100	.100	.100	.100	0

B. Adaptive capture of the compact exact projection

Steps	Mean capture	Median capture	5th percentile	Mean excess distance	Violations
16	86.9%	92.1%	59.2%	.00505	0
32	89.8%	93.8%	68.1%	.00375	0
64	92.0%	95.0%	75.8%	.00286	0
128	93.7%	96.0%	79.8%	.00226	0
256	94.9%	97.0%	82.2%	.00186	0

Table 5: Vision macro results. Accuracy is unchanged. Conf. denotes frozen confirmation.

Dataset/model	Status	Accuracy	Brier: anchor→PSPL	NLL: anchor→PSPL
C10-C Standard	Post-hoc	.8163	.03079→.03045	.9753→.9445
C10-C AugMix WRN	Conf.	.8807	.01818→.01803	.4087→.4013
C10-C AugMix ResNeXt	Conf.	.9032	.01461→.01452	.3246→.3193
C100-C AugMix WRN	Conf.	.6439	.004825→.004806	1.3896→1.3800

Table 6: OfficeHome fresh-seed confirmation (four domains, seeds 2711–2715). Brier is normalized by 65. ΔDec is the fraction of anchor decisions changed.

Method	Accuracy	Brier↓	NLL↓	ECE↓	ΔDec
Anchor	.7662	.005014	.9399	.0472	0
Candidate	.7060	.006192	1.1964	.0536	.2405
Fixed mix .25	**.7708**	.004945	.9211	.0597	.0456
Validation mix	.7708	.004946	**.9210**	.0592	.0413
Temperature scaling	.7662	.005071	.9644	.0543	0
Argmax isotonic	.7662	.005084	.9856	.0505	0
Single safe ray (PSPL)	.7662	.004964	.9245	.0534	**0**
Uniform safe hull (PSH)	.7662	.004955	.9196	.0531	**0**
TS + fixed mix .25	.7705	**.004910**	.9213	**.0390**	.0392
TS + validation mix	.7704	.004913	.9222	.0394	.0446
TS + PSPL	.7662	.004971	.9293	.0443	**0**

Table 7: Official-algorithm SimplexTS comparison on the locked CIFAR-100 test split ($n = 5{,}000$). "Fixed acc." uses the original logits prediction, as in the official ECE code; "Prob. acc." takes argmax of the returned probability vector.

Method	Fixed acc.	Prob. acc.	Brier↓	NLL↓	ECE↓	ΔDec
Uncalibrated	.7786	.7786	.003139	.8908	.0433	0
Scalar TS	.7786	.7786	.003139	.8907	.0436	0
SimplexTS ($\beta = 1.0$)	.7786	.7766	.003198	.9811	.0492	.0408

The authors' implementation is run on its official ResNet-18 split with 200 classifier epochs, 19 Multi-Mixup values, and 500 calibration epochs per value. Only the loop generating 15 million Gumbel variables is tensorized; 30,000-sample checks give maximum mean/variance discrepancies .00439/.00110, and an unmodified scalar run gives ECE $.0473$, within the vectorized 20-seed range $[.0437, .0502]$.

Table 8: VLCS locked external confirmation (four domains, five seeds).

Method	Acc.	Bal. acc.	Brier↓	NLL↓	ECE↓	ΔDec
Anchor	.7469	.6773	.07743	.8422	.1248	0
Validation mix	.7561	.6801	.07453	.7868	.1102	.0661
Temperature scaling	.7469	.6773	.07684	.8115	.1171	0
Single safe ray (PSPL)	.7469	.6773	.07603	.7879	.1209	**0**
Uniform safe hull (PSH)	.7469	.6773	.07571	.7781	.1191	**0**
TS + validation mix	**.7564**	**.6812**	**.07439**	.7665	**.1094**	.0661
TS + PSPL	.7469	.6773	.07565	**.7655**	.1150	**0**

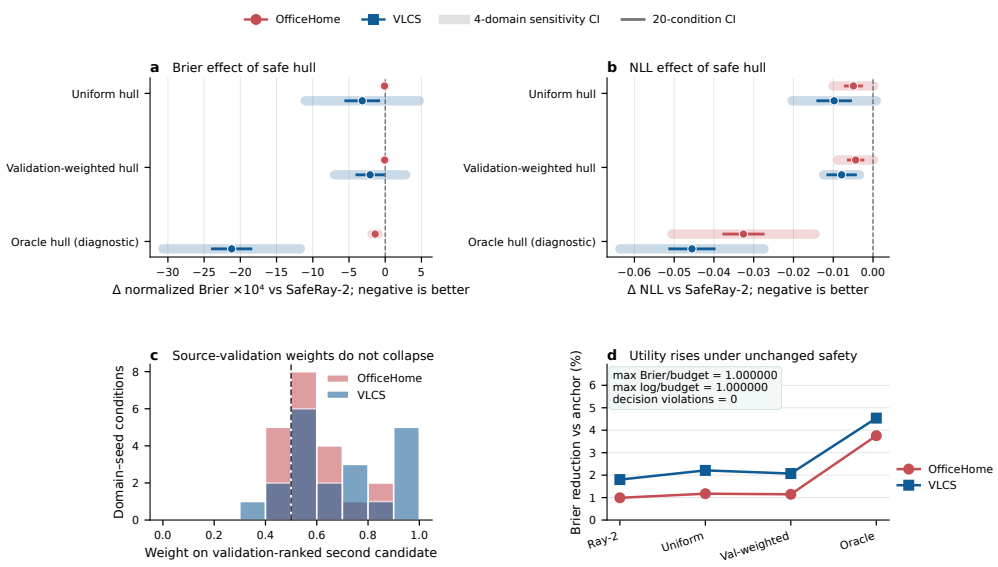

Figure 3: Complementary hull analysis. (a,b) Paired Brier and NLL differences relative to the validation-ranked safe ray; thin lines are 95% intervals over 20 conditions and wide lines are four-domain sensitivity intervals. (c) Source-validation weights do not collapse to a single endpoint. (d) Mean Brier reduction from the anchor; the target-label oracle is diagnostic only. The annotation reports maximum budget use and total decision violations.

C Experimental status and protocols

The following ledger marks confirmation, development, and stress-test evidence and states the permitted inference for each block.

Block	Status	Permitted inference
C10 AugMix WRN EATA	Frozen confirmatory	Data/model confirmation
C100 AugMix WRN EATA	Frozen confirmatory	Class-count confirmation
C10 AugMix ResNeXt EATA	Frozen confirmatory	Architecture confirmation
C10 Standard EATA	Post-hoc	Supporting evidence
C10 Tent–SAR	Development	Mechanism exploration
ImageNetV2-1K	Frozen confirmatory	Real 1000-class utility
Official SimplexTS	Frozen direct baseline	Accuracy-semantics comparison
Medical Raman replay	Exploratory	Cross-domain stress test
ComFilE nested veto	Development	End-to-end failure analysis
OfficeHome stable heads	Fresh-seed confirmation	Same-target seed robustness
VLCS stable heads	Locked external confirmation	Cross-dataset robustness
VLCS global projection	Locked diagnostic sample	Safe-set approximation

D Medical transfer and candidate-veto stress test

Saved-prediction replay on head–neck and prostate Raman tasks preserves every anchor decision and improves Brier and NLL. For the public ComFilE stroke and Alzheimer-related data, we rerun patient-level nested cross-validation from raw data: 20 seeds, five outer folds, and three inner folds. Within each outer fold, the model with lower inner Brier is the anchor. Action is selected in 100/100 Alzheimer-related folds, increasing balanced accuracy from the old fixed-Source value of .129 to .422. In stroke, noisy fold-wise selection chooses Source in 92/100 folds and Action in eight, yielding .823 balanced accuracy versus .845 for a fixed Source.

The medical stress test exposes when borrowing from the rejected model improves one proper score while worsening another. Unrestricted PSPL improves stroke NLL from .5487 to .4472 but changes

Brier from .13453 to .13499; on the Alzheimer-related data it worsens Brier from .21519 to .21972 and NLL from 1.0803 to 1.0977. A frozen *unanimous candidate veto* was developed for this failure mode: it enables editing only if PSPL improves both Brier and NLL in each of the three inner folds, without using outer-test labels. It rejects 100/100 Alzheimer-related edits and permits one of 100 stroke folds. We therefore report the veto as a development-stage mechanism that requires independent confirmation.

Table 9: Medical stress tests. Decisions are unchanged. Head–neck and prostate are saved-prediction replay; stroke and AD use end-to-end nested CV with the unanimous veto.

Setting	Acc.	Brier: anchor→PSPL	NLL	ECE
Head–neck	.555	.2864→.2810	.841→.790	.247→.227
Prostate	.791	.1165→.1156	.666→.643	.183→.173
Stroke veto	.823	.13453→.13425	.5487→.5480	.0611→.0611
AD veto	.459	.21519→.21519	1.0803→1.0803	.0686→.0686

E OfficeHome and VLCS uncertainty and audits

Table 10 gives the predeclared paired contrasts; each panel has 20 paired domain–seed rows, not 20 independent datasets.

Table 10: Paired mean differences and 95% t intervals. Negative values favor the edited method.

OfficeHome

Edit	Reference / metric	Mean	95% interval
PSPL	Anchor / Brier	-0.000050	$[-0.000064, -0.000035]$
PSPL	Anchor / NLL	-0.01532	$[-0.02142, -0.00922]$
PSPL	Anchor / ECE	$+0.00624$	$[+0.00086, +0.01163]$
TS+PSPL	TS / Brier	-0.000100	$[-0.000125, -0.000075]$
TS+PSPL	TS / NLL	-0.03507	$[-0.04620, -0.02395]$
TS+PSPL	TS / ECE	-0.00993	$[-0.01532, -0.00454]$

VLCS

Edit	Reference / metric	Mean	95% interval
PSPL	Anchor / Brier	-0.001395	$[-0.002230, -0.000559]$
PSPL	Anchor / NLL	-0.05426	$[-0.09362, -0.01490]$
PSPL	Anchor / ECE	-0.00388	$[-0.01501, +0.00725]$
TS+PSPL	TS / Brier	-0.001182	$[-0.002087, -0.000276]$
TS+PSPL	TS / NLL	-0.04602	$[-0.08656, -0.00548]$
TS+PSPL	TS / ECE	-0.00215	$[-0.01467, +0.01037]$

All 60 logistic heads per dataset converge, both PSPL variants have zero decision violations, and all Brier/log regrets obey the declared .005/.1 budgets up to float64 tolerance. At the four-domain level, PSPL Brier/NLL intervals are $[-0.0001013, +0.0000020]$/$[-0.03749, +0.00685]$ on OfficeHome and $[-0.00441, +0.00162]$/$[-0.1940, +0.0855]$ on VLCS. The first locked OfficeHome run with overconfident SGD heads is retained only as a diagnostic; the stronger VLCS TS+validation mix changes 6.61% of decisions and is not safety-equivalent to PSPL.

F Global safe-projection solver audit

The diagnostic draws 500 class-balanced predictions from each of 20 locked VLCS domain–seed conditions. For Brier and candidate-fidelity objectives, we solve the exact global body and four nested hulls with second-order cone programs, then pass every numerical solution through the analytic ray checker to realize strict tie semantics. Of 80,000 solves, 79,883 terminate as optimal and 117 as close to optimal; all repaired outputs have zero decision violations, worst Brier regret .005, worst log regret .1, and no nesting violations above 10^{-5}. The projection objective uses the validation-ranked candidate but no target labels; labels enter only the reported Brier/NLL metrics.

Table 11: Nested safe-set diagnostics on 10,000 VLCS predictions. Candidate capture is label-free; Brier capture uses a target-label oracle only as a representational upper bound.

Direction set	Total directions	Candidate capture	Brier-oracle capture
One model ray	1	79.3%	40.7%
Two model rays	2	81.9%	40.7%
+ simplex vertices	7	96.0%	95.4%
+ ordered transfers	27	96.2%	95.4%
Global safe body	SOCP	100%	100%

Table 12: ImageNetV2 paired differences for 16-step Adaptive PSH with 1,000 vertex atoms (method minus reference), with 95% t intervals over 1,000 class clusters of ten images. Negative is better except for top-5.

Reference	Candidate distance	Brier ($\times 10^{-6}$)	NLL	Top-5
SafeRay-Mean3	$-.02202$ $[-.02235, -.02170]$	$+3.03$ $[2.39, 3.67]$	$-.00789$ $[-.01204, -.00375]$	$-.0010$ $[-.0031, .0011]$
Model-atoms-16	$-.00578$ $[-.00610, -.00546]$	$-.86$ $[-1.22, -.49]$	$-.01937$ $[-.02340, -.01534]$	$-.0012$ $[-.0033, .0009]$

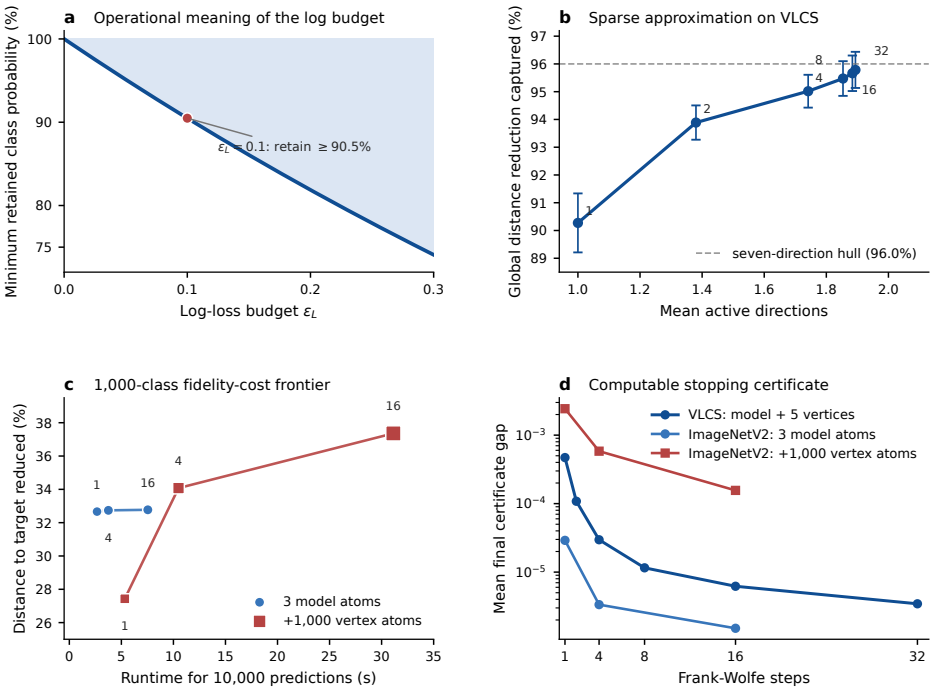

Figure 4: Budget semantics and adaptive projection trade-offs. (a) The all-label log budget imposes $r_y/p_y \geq e^{-\epsilon_L}$; the main $\epsilon_L = .1$ setting retains at least 90.48% of every anchor probability. (b) On 10,000 balanced VLCS predictions, numbers label Frank–Wolfe steps, error bars are 95% t intervals over 20 locked domain–seed conditions, and the dashed line is the full seven-direction hull. (c) On 10,000 ImageNetV2 predictions, marker area encodes mean active directions and labels give the step budget; lines connect runs with the same atom set. (d) Mean per-example Frank–Wolfe certificate versus steps. All empirical points use labels only for reporting, and all have zero Brier, log, and decision violations.