DEPLOYMENT-AWARE MODEL SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Two serving stacks can agree on every number a leaderboard publishes and still be the right and the wrong choice for the same deployment. Of 620 Open LLM models, 77 are the best available choice at some task mixture while topping none of the seven numbers published. A system's *risk profile*, its best achievable risk at each operating condition, is the minimal object an evaluation can depend on; practice puts point masses on it, and we characterize the linear and convex rules completely. What decides when that choice matters is carried not by a profile but by the *differences* between candidates'. Aggregating over a range ranks a pair exactly as scoring it at the range's average does, precisely when that difference profile is affine. Shared curvature therefore cancels, and profiles may bend arbitrarily yet reorder nothing. No regular *linear* rule reading finitely many conditions is deployment-complete, and we localize what a suite of covering radius h can miss. The boundary is also a noise-calibrated test of whether a full-profile criterion can change a selection. It is correct on all eight systems we try, and bending predicts divergence, not benefit: four bend and two pay, by 13–28% on re-steerable RLHF pipelines and 72% on moderation. A replication across 96 hospitals shows that advantage is largely a hedge against estimation error, fading as profiles improve and the rules converge (60% agreement to 92%). What pays throughout is evaluating at the deployment's own condition, 0.6–0.8 away in total variation from the headline's, using scores already published.

1 INTRODUCTION

A team building a coding assistant shortlists ten open models and takes the one ranked highest on a public leaderboard, which ranks by an equal-weight average of six benchmarks. The assistant's workload is nothing like an equal mixture. On the rankings the community consults, the top-ranked pick is not the mixture-optimal choice for plausible non-uniform mixtures in one sampled shortlist in six to one in three. That holds on HELM (Liang et al., 2023) and the archived Open LLM Leaderboard (Fourrier et al., 2024), which average fixed task sets, and on LMArena (Chiang et al., 2024), which instead fits a Bradley–Terry model to preferences pooled over its own traffic. The aggregation rule is not what separates them. A voting rule (Rofin et al., 2023), or a latent-ability model that accounts for task difficulty, still returns a single number at a single operating condition: one condition is chosen and the rest discarded. For the averaging leaderboards the scores needed to pick correctly were published all along. Clinical evaluation repeats the pattern. BRIDGE (Wu et al., 2026) was built because exam-style questions miss real clinical text, yet it ranks by an equal-weight average over 87 patient-care tasks, and its leader is not the best model for 11 of the 19 hypothetical sites its own task labels define. More data does not help: averaging at the wrong mixture with higher precision returns the same wrong answer.

The same structure appears wherever evaluation and deployment differ. An ICU mortality model is validated at prevalence $\pi \approx 0.15$ and deployed at $\pi \approx 0.04$. A moderation classifier is tuned at $\pi = 0.35$ and meets 1–5% harmful traffic. In each case the procedure evaluates at one point of a space that deployment occupies somewhere else.

When a system is evaluated under one set of conditions and deployed under another, what must the evaluation measure for the selection to be reliable?

The answer is a function, not a number. It is the *risk profile*

$$r_s : \Omega \to \mathbb{R}, \qquad r_s(\omega) = \text{best achievable risk of system } s \text{ at operating condition } \omega, \tag{1}$$

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.

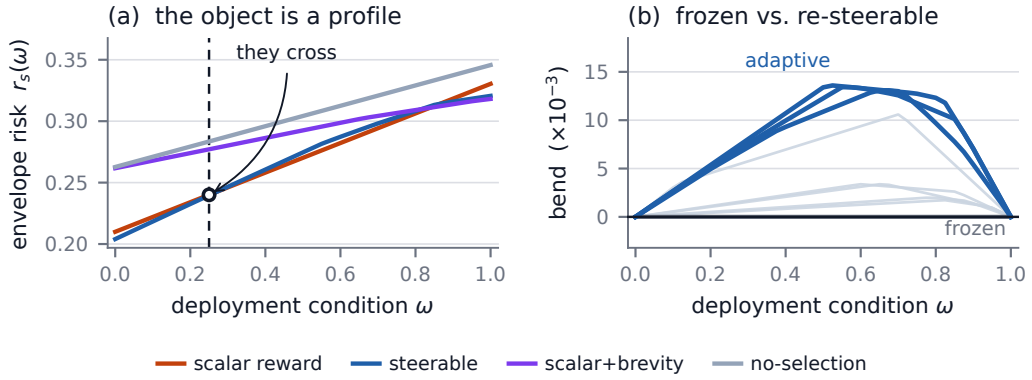


Figure 1: **The paper's argument, on four real RLHF serving pipelines. (a)** A pipeline's risk depends on the deployment it faces. Drawn along a path from a research-analyst persona to a mobile-assistant persona, that is its *risk profile*. The four curves cross, so no single condition ranks them: the best pipeline changes twice along this one path, so the differences between curves change sign, which is what decides the question. **(b)** How straight those curves are. Frozen, every curve is straight to 4.7×10^{-16}, so averaging over a range of conditions gives the same answer as scoring at its middle and the rule cannot matter (Proposition 1, which turns on the *differences* between curves). Let the deployer re-tune and the curves bend, most for the pipelines with the most to re-tune. What the bend is worth, frozen against adaptive, is 0.0% exactly against 13–28% (Table 4).

a function on a compact space Ω of *operating conditions*: prevalences, cost ratios, task mixtures, demand regimes.

Much of this object is inherited. It is the lower envelope of Drummond and Holte's cost curves (Drummond and Holte, 2006); integrating performance over operating conditions goes back to Adams and Hand (1999) and the threshold-choice frameworks of Hernández-Orallo et al. (2012) and Hernández-Orallo et al. (2013); Hand (2009) already observed that choosing a criterion is choosing a measure, and that AUC's measure is classifier-dependent; and decision-curve analysis evaluates over ranges of threshold or prevalence in exactly this spirit (Vickers and Elkin, 2006; Flores et al., 2025). What we add is the boundary: which of these rules can differ from which, and when. Shift methods correct one model toward one target (Saerens et al., 2002; Sugiyama et al., 2007; Lipton et al., 2018); robust optimization and mixture-reweighting train against a set of them (Mohajerin Esfahani and Kuhn, 2018; Duchi and Namkoong, 2021; Xie et al., 2023); benchmarks of in-the-wild shift supply conditions without a rule for aggregating over them (Koh et al., 2021). We *select* among systems already trained, across a range of targets none of them was fitted to. The systems that make this live are the ones now being deployed: cascades that escalate only when a cheap model looks unreliable (Chen et al., 2023) and routers learned from preference data (Ong et al., 2024) reconfigure at deployment, and it is exactly that reconfiguration which can bend a profile away from affine.

The thesis is that **a deployment-equivalence class is represented by its profile, and regular linear evaluation integrates that profile over deployment uncertainty**; convex rules take suprema of such integrals. We make three contributions.

1. The exact boundary (Section 2). The risk profile is the right deployment-aware object: every evaluation depending only on achievable risk factors through it (Theorem 1), linear rules are exactly measures and convex ones suprema over sets of measures (Theorem 2), which turns Hand's observation into a classification. What decides whether the choice among these rules matters is carried not by a profile but by the *differences* between candidates': aggregating against a measure ranks a pair exactly as scoring it at that measure's barycentre does, precisely when the difference profile is affine (Proposition 1). Each profile being affine, which is what a non-reconfigurable system gives, is sufficient and not necessary, since curvature shared by all candidates cancels.

2. What finite evaluation misses, and what published numbers already hide (Section 3). No regular *linear* rule consulting finitely many conditions is deployment-complete (Theorem 3); within that class completeness is equivalent to full-support integration (Corollary 1). For realizable systems we localize where a suite of covering radius h can still be fooled, at configurations optimal on no ball of radius $3h$ (Theorem 4), which reconfigurable pipelines manufacture at widths below 10^{-3}. Such

blind sets are not hypothetical: two serving stacks the Open LLM Leaderboard scores *identically* on every task differ by 190 times the gap between adjacent ranks at a mixture we specify.

3. A calibrated test, eight systems, and a prescription (Sections 4, 5). Raw curvature is not evidence, since estimation noise bends an estimated difference profile even when the true one is straight; the released `curvature_test` calibrates it against a null in which the differences are affine, at a measured false-positive rate of 3.5%. We run it on eight systems (Table 2) spanning leaderboards, re-steerable RLHF pipelines, safety moderation, logged Arena traffic, LLM routing and two clinical databases. Differences bend in four, which the test calls correctly every time, and aggregation pays in two, then only for a deployer that already knows its operating condition. A pre-registered replication across 96 hospitals (Pollard et al., 2018) shows more source data does not sharpen curvature into an advantage but makes every rule converge (60% agreement to 92%). Our prescription: publish the per-condition matrix, a sufficient statistic for non-reconfigurable systems, and evaluate at your own condition. Supporting theory: Wasserstein duality with a first-order expansion (Theorem 5) and an oracle inequality separating the three error sources a practitioner faces (Corollary 3) and related work are in the appendix (Appendix B).

2 The Object, and When the Rule Cannot Matter

Operating conditions and profiles. An *operating condition* $\omega \in \Omega$ is any deployment-time parameter that affects which system performs best and may differ between evaluation and deployment; we take Ω compact metric. Three instantiations run through the paper: a task mixture on the simplex Δ^{T-1}, a prevalence and cost ratio (π, c) on a rectangle, and a harmful-content rate π on an interval. Given a system s with deployment-time configuration τ (a threshold, a routing rule, an order quantity) and loss L continuous in ω, the deployment risk is $R_\omega(s,\tau) = \mathbb{E}_{(X,Y)\sim P_\omega}[L(a_{s,\tau}(X), Y, \omega)]$ and the *envelope risk profile*

$$r_s(\omega) \;=\; \inf_{\tau}\; R_\omega(s,\tau) \tag{2}$$

is continuous on compact Ω (maximum theorem) and records the system's best achievable risk at each condition. For thresholded binary classification, $r_s(c,\pi) = \inf_\tau[\, c(1-\pi)\mathrm{FPR}(\tau) + (1-c)\pi\,\mathrm{FNR}(\tau)\,]$; properties are in Appendix C.7.

Why this object. An evaluation functional is an *aggregation rule*: a map $\Phi : C(\Omega) \to \mathbb{R}$ turning a whole profile into one number, with selection $\arg\min_k \Phi(r_{s_k})$. Call two systems *deployment-equivalent*, $s \sim s'$, if $\inf_\tau R_\omega(s,\tau) = \inf_\tau R_\omega(s',\tau)$ for every ω. Two requirements pin the profile as *the* representation.

Axiom 1 (Risk relevance)**.** Evaluation depends on a system only through its risk surface $(\omega,\tau) \mapsto R_\omega(s,\tau)$.

Axiom 2 (Reconfiguration invariance)**.** Deployment-equivalent systems receive equal evaluations: the deployer re-optimizes τ at deployment, so evaluation may not distinguish systems whose achievable risks coincide everywhere.

Theorem 1 (The profile is the minimal sufficient representation)**.** *(a) Any evaluation E satisfying Axioms 1 and 2 factors through the profile: there is a functional Φ with $E(s) = \Phi(r_s)$, and $r_s = r_{s'} \iff s \sim s'$, so the factorization is through the exact quotient. (b) If T is any statistic through which every admissible rule (the regular class of Theorem 2) factors, then $T(s) = T(s')$ implies $r_s = r_{s'}$: no smaller representation supports deployment-aware evaluation.*

The proof (Appendix C.5) is short: (a) is the universal property of a quotient, and (b) holds because every point evaluation $r \mapsto r(\omega)$ is itself an admissible rule. Minimality rests on that choice of admissible class; nothing else does.

2.1 What the aggregation rules are

Call Φ a *linear aggregation rule* if it is monotone, positively linear and sup-norm continuous. Minimax and CVaR-style rules are monotone and continuous but not linear, and a second representation covers them, transplanted wholesale from coherent and convex risk measures (Artzner et al., 1999; Föllmer and Schied, 2002; Schmeidler, 1989; Gilboa and Schmeidler, 1989); we claim only the transplant to evaluation, where Ω holds operating conditions rather than monetary outcomes.

Theorem 2 (Classification of aggregation rules). ***(a)*** *Φ is a linear aggregation rule if and only if there is a unique finite nonnegative Borel measure μ on Ω with $\Phi(r) = \int_\Omega r\, d\mu$ (Riesz–Markov–Kakutani).* ***(b)*** *If Φ is monotone, translation-equivariant, convex, sup-norm continuous and normalised, there is a convex, weak-$*$ closed penalty $\alpha : \mathcal{P}(\Omega) \to [0, \infty]$ with $\Phi(r) = \sup_{\mu \in \mathcal{P}(\Omega)} \{\int_\Omega r\, d\mu - \alpha(\mu)\}$; if Φ is also positively homogeneous then $\Phi(r) = \sup_{\mu \in \mathcal{M}} \int r\, d\mu$ for a closed convex $\mathcal{M} \subseteq \mathcal{P}(\Omega)$.*

So a linear rule is a weighted average of the profile and two such rules differ only in the weighting μ, which we call the *criterion measure*; a convex rule is a worst case over a set of weightings. Part (a) has the shape of Schervish's characterization of proper scoring rules as mixtures of cost-weighted losses (Schervish, 1989; Buja et al., 2005), with the integral running over deployment conditions rather than forecast thresholds. Designing a criterion is choosing where to put the weight; designing a robust criterion is choosing what to hedge against. Canonical sets and their penalties are tabulated in Appendix D.3; the proofs apply the classical theorems directly (Appendix C.1).

Table 3 (Appendix F.1) places every procedure in this classification. The procedures that dominate practice, cross-validation at development conditions, benchmark averages and fixed-threshold clinical metrics, are all *the same aggregation rule*: a Dirac measure at one operating condition, differing only in which condition. AUC escapes the framework rather than sitting degenerately in it: scorers sharing an ROC hull, and so sharing a profile, can differ in AUC on dominated parts of their curves, while the area under the hull *is* an aggregation rule, improper in Hand's sense (Hand, 2009; Flores et al., 2025).

Example 1 (OPERA: a default full-support rule). Take μ non-degenerate on Ω and select $s^* = \arg\min_k \Phi_\mu(r_{s_k})$, $\Phi_\mu(r) = \int r\, d\mu$: rank candidates by their average risk over the conditions you might actually face. Use the deployment distribution when it can be elicited. Otherwise, for prevalence intervals, we default to the *log-odds uniform* measure $d\mu_{\mathrm{LOU}} \propto d\pi/[\pi(1-\pi)]$, which is logit-translation invariant on the untruncated scale and is the fixed point of the problem's symmetry group there (Appendices E.2, E.4); truncated to a deployment interval it is a default whose normaliser depends on the endpoints (Appendix I.5). OPERA is Bayes-optimal for conditions drawn from μ, and against a point mass its μ-expected regret is never larger (Appendix C.1). Its robust variant R-OPERA replaces μ by a Wasserstein ball (Example 2).

2.2 The collapse boundary

Call a system *non-reconfigurable* if its rule cannot be changed at deployment, so R_ω is evaluated at a fixed τ. The following is the paper's pivot: it says exactly when everything above is unnecessary.

Proposition 1 (Collapse boundary). *Let $\Omega \subset \mathbb{R}^d$ be convex and compact, let s_j, s_k be candidates with continuous profiles, and write $g := r_{s_j} - r_{s_k}$ for their* difference profile. *For a probability measure μ on Ω let $\bar{\omega}_\mu := \int \omega\, d\mu$ be its barycentre.*

(i) (sufficient, and exact) *If g is affine, then for every μ*

$$\int r_{s_j}\, d\mu - \int r_{s_k}\, d\mu \;=\; r_{s_j}(\bar{\omega}_\mu) - r_{s_k}(\bar{\omega}_\mu),$$

so every linear aggregation rule ranks s_j against s_k exactly as the point mass at its own barycentre does.

(ii) (necessary) *Conversely, if $\int g\, d\mu = g(\bar{\omega}_\mu)$ for every μ, then g is affine.*

(iii) (the boundary is about differences) *If every candidate's profile is affine then every difference profile is affine and (i) applies to the whole candidate set, which is the classical statement. The converse fails: a curved component* common *to all candidates cancels in every difference and cannot change any selection, however large it is.*

Proof. (i) An affine function integrates to its value at the barycentre, and the integral is linear, so $\int g\, d\mu = g(\bar{\omega}_\mu)$; a common ranking follows because the sign of the difference is preserved. (ii) Take $\mu = \frac{1}{2}(\delta_x + \delta_y)$, whose barycentre is $\frac{1}{2}(x+y)$; the hypothesis makes g midpoint-affine, and a continuous midpoint-affine function on a convex set is affine. (iii) The first claim is immediate. For the second, let $r_{s_k} = a_k + h$ with a_k affine and h common to all candidates: each difference $a_j - a_k$ is affine regardless of h. □

Remark 1 (What to measure). The quantity deciding whether the aggregation rule matters is the curvature of the *difference* profiles, not of the profiles themselves. The distinction is not academic: in the clinical experiment of Section 4 the candidates have the largest per-profile curvature anywhere in this paper and the smallest difference curvature, because they bow together, and there the integral does not beat a barycentre-matched point mass. We report both statistics throughout.

Deployment-time adaptation threshold re-optimization, routing, re-pricing is what *can* bend the envelope away from affine, but it does not force it to: one configuration may stay optimal throughout Ω. What takes a system out of the collapsed regime is not adaptation but non-affine *difference* profiles; adaptation only makes those possible (Remark 12).

3 What a Finite Suite Misses

Why use the entire profile? Because three requirements, each innocuous on its own, cannot hold together. A criterion should be a sensible summary of risk (*regularity*: monotone, positively linear, sup-norm continuous, Axiom 3); it should consult finitely many operating conditions (*locality*, Axiom 4); and a system strictly better across a whole region of deployment conditions should score strictly better (*deployment completeness*, Axiom 5). Every criterion in the first block of Table 3 satisfies the first two, and so does any benchmark suite, however many benchmarks it contains. The third is what a deployment stakeholder means by "the evaluation sees where my system will run." Formal statements are in Appendix C.6. Throughout, Ω is compact metric with no isolated points.

Theorem 3 (Blind-spot impossibility)**.** *No evaluation functional satisfies Axioms 3–5. Moreover, any Φ satisfying Axioms 3 and 4 with witness set Ω_0 has a* blind set*: a nonempty open $U \subseteq \Omega$ such that for every sufficiently small $h > 0$ there exist L-Lipschitz profiles r_1, r_2 with $\Phi(r_1) = \Phi(r_2)$ and $r_2 - r_1 \geq h/2$ on an open subset of U, so deployment there incurs regret the criterion registers as exactly zero. On convex Ω the witnesses may be taken* realizable*, as envelopes of finitely many affine functions with independently specifiable gradients (Lemma 2); the blind set's size is governed by the covering radius of Ω_0 (Proposition 3).*

Any evaluation that consults finitely many conditions therefore leaves a whole region of deployments in which one system is worse than another by a fixed margin while the score says they are equal, and the witnesses are not pathological: they are profiles realizable by finite configuration sets. Within the linear class the escape is exactly one thing.

Corollary 1 (Completeness in the linear class)**.** *An evaluation functional satisfies Axioms 3 and 5 if and only if it is integration against a measure with* full support *on Ω.*

This says nothing about nonlinear rules: a set of measures that all have full support is sufficient for completeness of a convex rule, but Wasserstein balls do not meet that condition, so the argument does not establish completeness for R-OPERA (Appendix D.2). The quantifier also matters. Theorem 3 ranges over *all* Lipschitz profiles, and restricting the class overturns it: if profiles are known to lie in a fixed d-dimensional linear subspace, d chosen conditions determine them (Proposition 9), and Proposition 1 is the affine instance. Realizable envelopes form a cone rather than a linear space, so the right question is *which* realizable systems a finite suite can see. The answer is a property of the system, not of the criterion. Call a configuration *active* if it attains the minimum somewhere, and an envelope $r_s = \min_\tau a_\tau$ *w-fat* if every active configuration is optimal on some ball of radius w contained in Ω.

Theorem 4 (Fatness threshold)**.** *Let Ω be convex and compact and let Ω_0 have covering radius h. If r_1, r_2 are w-fat envelopes with $h < w/3$ and $r_1 = r_2$ on Ω_0, then $r_1 = r_2$ on Ω. More generally, if any two envelopes agree on Ω_0, each lies weakly below every configuration of the other that is optimal on a ball of radius greater than $3h$; they can disagree only on the union of the regions of configurations, of either, optimal on no such ball.*

A finite suite is blind only to configurations that win on thin slivers of the condition space. Steerability is what manufactures the slivers, and the effect is stark. Exact Chebyshev-ball linear programs on the pipelines of Section 4.2 give $w \in [0.025, 0.071]$ for the six non-steerable pipelines, while the steerable one has 90 active configurations with $w = 2.9 \times 10^{-4}$: seeing it exactly would need a covering radius below 10^{-4}, against 0.56 for a six-condition suite. One optimality region makes a profile affine; many thin ones make it curved *and* invisible. A fixed benchmark suite reports finitely

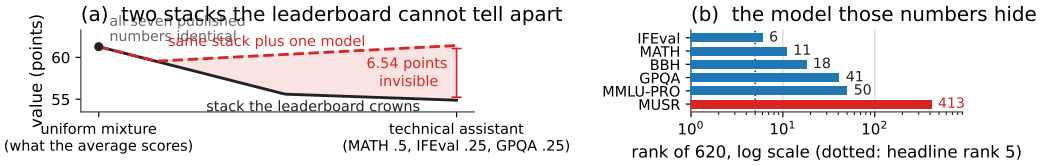


Figure 2: **A blind set, exhibited on a public leaderboard.** **(a)** Two routing stacks the Open LLM Leaderboard scores *identically* on all six tasks and on the average, differing only in whether `Qwen2.5-72B-Instruct-abliterated` is a member. At a technical-assistant mixture we specify, they differ by 6.54 points, about 190 times the median gap between adjacent ranks. **(b)** That model is unremarkable on every task and 413th of 620 on MUSR, which holds its average down, yet it is the best choice at that mixture. Theorem 3 on published scores, with stacks we construct.

Table 1: Open LLM Leaderboard, 620 models: the headline #1 (`calme-3.2-instruct-78b`) against realistic deployment mixtures. The point mass is free where the deployment sits near the uniform mixture and costs several points where it does not, the blind-spot geometry of Theorem 3 on a public artefact. Shaded: the deployments where taking the headline #1 costs points.

Deployment persona	Best model at that mixture	Cost of the headline #1
math-tutor	`Qwen2.5-32B-Instruct`	**5.53 pts**
coding-agent	`Qwen2.5-72B-Instruct`	**3.82 pts**
chat-assistant	`calme-3.2-instruct-78b`	0.00 pts
research-analyst	`calme-3.2-instruct-78b`	0.00 pts

many conditions, so Corollary 4 (Appendix F.1) applies verbatim: within the linear class no such suite is deployment-complete, whatever its size. Proofs are in Appendices C.6 and F.2.

The blind set is not hypothetical. The Open LLM Leaderboard publishes seven numbers per model for a pool of 620: six per-task scores and their equal-weight average. Each individual model's profile is affine and those seven numbers determine it, so a leaderboard used to choose among *static* models has no blind set at all. The blind set appears exactly when a deployer may *route*: a serving stack that picks among several models at deployment has as its profile the lower envelope of its members', which is concave and no longer pinned by the vertices. Routing is routine. Call a model *invisible* when it attains the maximum at none of the seven published numbers, so adding it to the crowned stack changes no reported score. Among the 620 models, 615 are invisible and 77 of those lower deployment risk somewhere on the mixture simplex, computed exactly by linear programming. Two routing stacks scored identically on every task and on the average differ by $\mathbf{6.54}$ average-points at a realistic mixture, against a median difference of 0.034 points between adjacently ranked models in the top 200: the invisible gap is $\mathbf{190}\times$ the difference the community already treats as a meaningful ranking signal (Figure 2). Taken each at its own best mixture, the 77 hide a median of 62 times the pool-wide adjacent-rank gap and $316\times$ at the most extreme. These are existence statements over mixtures, not claims about typical deployments; Table 1 gives the cost at four deployments we name in advance. The same construction on our own serving pipelines needs roughly 150 reported conditions to reach adjacent-rank resolution, some twenty-five times what a six-task leaderboard publishes; the invisible gap decays about two and a half times faster than the worst case, so real achievable sets are low-complexity, but its exponent still tracks $1/d$ at $r = 0.98$ (Appendix G.1).

4 Experiments

Section 4.1 varies *where* systems are evaluated, on four public leaderboards; Section 4.2 varies whether reconfiguration makes the choice of *aggregation rule* matter; Section 4.3 turns the boundary into a test and audits it on all eight systems. Six rules compete throughout: a point mass at the evaluation condition, the uniform measure, minimax, CVaR(0.9), OPERA and R-OPERA, each judged by realized regret against the deployment's oracle-best candidate, and each granted deployment-time re-optimization equally (Appendix G.7 adds stress tests and legacy benchmarks).

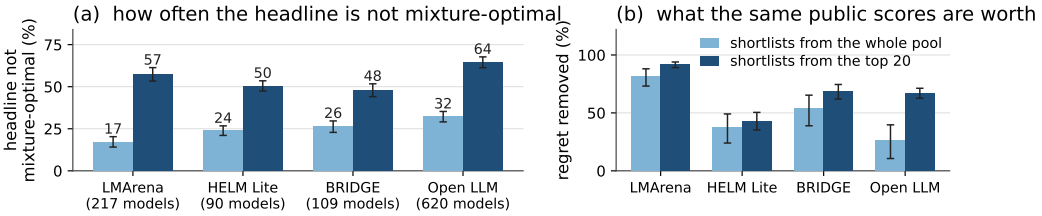


Figure 3: **Four public leaderboards, one degeneracy. (a)** How often the headline-ranked model is not the mixture-optimal candidate under sampled mixtures, for shortlists of ten drawn from the whole pool and from the headline's own top 20. **(b)** Mean-regret reduction of the best profile rule against the headline, on each leaderboard's own published scores. Intervals are 95% cluster bootstraps. Open LLM was archived in March 2025, and the finding concerns the rule, which every live leaderboard still uses (Appendices G.3, G.4).

4.1 Where the rule cannot matter: four leaderboards

Every ranking the community consults is a point-mass rule applied to a profile it already possesses, though they reach that point differently: LMArena (Chiang et al., 2024) fits a Bradley–Terry model to preferences pooled over the arena's own traffic, while HELM (Liang et al., 2023), the archived Open LLM Leaderboard (Fourrier et al., 2024) and the clinical BRIDGE (Wu et al., 2026) take equal-weight averages over fixed task sets. Holding their published per-condition scores fixed, evaluating at a *hypothesized* deployment mixture rather than the headline's displaces the top-ranked model in 17–32% of sampled shortlists, and 48–65% among each headline's own top 20 (Figure 3). These deployments are ours, not observed: at the leaderboard's own uniform mixture the headline is near-optimal (6.3%) and integrating is far worse (Appendix G.2). No cleverer fixed rule recovers the difference either, since Borda and Copeland aggregation (Rofin et al., 2023) are worse than the mean on three of the four leaderboards (Appendix G.5).

Every candidate here is non-reconfigurable, so each profile is affine in the mixture and every difference profile is affine exactly: no profile bends by more than 2.6×10^{-15} across all 620 Open LLM and 109 BRIDGE models. By Proposition 1 no linear rule can rank these candidates differently from its own barycentre-matched point mass, and none does. What this section measures is therefore the cost of evaluating at the *wrong* condition, and nothing about the rule.

How far deployments actually sit. That cost scales with the distance from the headline's condition to the deployment's, so we measure that distance wherever deployments are observable; this has not been measured before. BRIDGE's 19 metadata-defined sites sit at median total variation 0.71 from the condition its headline uses and RouterBench's 67 task sources at 0.62; separately, every MIMIC-IV care unit (Johnson et al., 2023) runs at prevalence 0.038–0.152 against the 0.35 its models are developed at. These are different geometries and we do not combine them. The exception is instructive: eleven *language markets* on logged Arena traffic (LMArena, 2025) sit at median 0.085, and there the headline is mixture-optimal in 98% of shortlists, CI $[97.4, 99.5]$ (Appendix G.6); a language market is an aggregate, not a deployment. For systems used as they are no new benchmark is needed: weighting published per-task scores by the site's case mix is exact under the stated linear-utility and commensurability assumptions (Appendix G.4).

4.2 The boundary, crossed: preference weightings in RLHF

This experiment tests the boundary itself, by turning deployment-time reconfiguration off and on with everything else held fixed.

Setup. $\Omega = \Delta^4$ is the simplex of deployment *preference weightings* over five axes a served response is judged on, instantiated on UltraFeedback (Cui et al., 2024) (88,843 completions) and on HelpSteer2 (Wang et al., 2024b), where all five axes are human-annotated. A candidate is a *serving pipeline*: a reward-model family $\{R_\theta : \theta \in \Theta_s\}$ with a best-of-n budget, both reconfigurable at deployment. On a shared pool of real responses per prompt,

$$R_\omega(s, \tau) = \mathbb{E}_x\big[1 - \omega \cdot a\big(\arg\max_{y \in S_n(x)} R_\theta(y)\big)\big] + \lambda(n-1), \qquad \tau = (\theta, n), \tag{3}$$

Table 2: **The diagnostic, and what it predicted.** Curvature of each candidate's own profile, of the pairwise *difference* profiles, and the latter relative to the between-candidate spread, the quantity Proposition 1 makes decisive. "Rules differ" is how often the integral and a barycentre-matched point mass select differently; "gain" is the integral's regret reduction against that point mass. Own $= \max_\pi |r_k - \mathrm{aff}(r_k)|$ averaged over candidates; Diff. the same for every pairwise difference $r_j - r_k$; Diff./spread divides by the mean between-candidate spread. [†]Four leaderboards and the frozen arm of every reconfiguration study. Deployments are observed ([o]), defined by benchmark metadata ([m]), hypothesized persona mixtures ([h]), a specified operating range ([s]), or constructed offline from real responses ([c]). **Highlighted**: the four systems whose difference profiles curve.

System	Own $\times 10^3$	Diff. $\times 10^3$	Diff./spread	Rules differ	Gain
Non-reconfigurable[†,h]	0	0	0	0%	0.0% exactly
Arena vendors[o]	0.006	0.013	5×10^{-5}	0%	0.0%
BRIDGE, strategy choice[m]	1.53	2.84	0.010	0.2%	nil
RLHF, re-steerable[c]	**3.08**	**3.58**	**0.082**	**12%**	**+13 to +28%**
Moderation, thresholds[s]	**7.63**	**2.70**	**0.383**	**30%**	**+72%**
MIMIC-IV care units[o]	**12.06**	**1.76**	**0.396**	**34%**	**−4.0% n.s.**
LLM routing[o,s]	**84.4**	**96.6**	**0.412**	**20%**	**−2.2% unstable**

so the profile is a lower envelope of affine functions, curved exactly to the extent that the deployer can re-steer: multi-objective reward models can be re-weighted at deployment and specialist ones cannot, so profiles cross (Appendix H.2). This is an offline construction with a deployer that knows the weighting; it buys control, and the real-system replication below relaxes each simplification.

Results. Frozen, every profile bends by at most 4.7×10^{-16} (Figure 1(b)), and the integral and a *barycentre-matched* point mass select the same system in every shortlist, with a reduction of 0.0%, CI $[0.0, 0.0]$, in all five variants (Table 4, 240 replications, $K = 6$). Let the deployer choose θ and n and the profile curves, from 28–30 $\times\ 10^{-3}$ for scalar-reward pipelines to 113–116 $\times\ 10^{-3}$ for steerable ones. The two rules now diverge, and the gain concentrates there: they agree on 82–93% of shortlists, but where they differ the barycentre-matched point mass gives up 36–78% of its regret, an overall reduction of 13–28% with every interval excluding zero. Curvature predicts the boundary, not the magnitude: it separates the frozen regime, where the rules always agree and the reduction is exactly zero, from the adaptive one, but does not track the size of the gain ($r = -0.004$). A scalar-reward point mass agrees only 9% of the time, but loses 32–57% even frozen, so only the barycentre-matched comparison isolates the rule. Under 18 rescalings of the preference axes the frozen reduction stays 0.0% and the adaptive gain vanishes where the kink leaves the measure's support, exactly as Proposition 1 requires (Appendix H.4).

A real adaptive system, and the precondition it exposes. We rebuild the chart from fresh best-of-8 generations of Qwen2.5-7B-Instruct (Team, 2024) on 480 held-out prompts, selected by ArmoRM's heads (Wang et al., 2024a) and graded by an independent OLMo-2-13B judge (OLMo, 2024). Frozen, the reduction is again 0.0%. An oracle deployer gains 68% $[30, 82]$, on a small base, or 47% $[9, 61]$ when only the configuration means are estimated. Both require the deployment's preference weighting to be *known*. Estimating it yields no gain distinguishable from zero at any budget tested: -3.9% $[-8.0, -0.2]$ at $m = 10$ pairwise labels, significantly *worse* than the point mass, then -0.6%, $+9.7\%$ and $+22.8\%$ $[-1.3, 34.8]$ at $m = 30, 100, 300$, every interval covering zero (Figure 8(b), Appendix H.3). Curvature is necessary for the rule to matter; knowing your own operating condition is what makes it pay.

4.3 The diagnostic, and what it predicted

Raw curvature is not evidence on its own, because estimation noise bends an estimated difference profile even when the true one is straight. The released `curvature_test` calibrates it. The statistic is the mean over pairs of the maximum deviation of $r_j - r_k$ from affinity; the null takes the affine part of each estimated difference as truth; the noise comes from resampling whole replications, so candidates scored on the same examples stay paired. It rejects at the null's 95th percentile, with a measured false-positive rate of 3.5%.

Table 2 is the diagnostic and its audit at once, over all eight systems. The statistic separates systems where the rule cannot help from those where it can. The top three rows cannot, and two of those

can reconfigure and still do not curve. Among those that curve it predicts divergence, not benefit. Moderation and care units make the point together: their difference curvature is the same to within 3% and their outcomes are opposite, but their calibrated ratios differ fivefold, 1.58 against 7.61 ($p < 0.001$ each, 20 replications), because the clinical bend clears its own noise floor only narrowly. The gains follow that calibrated ordering rather than the raw one: -4.0% and $+72\%$. Both columns are computable before adopting anything. Hence the decision rule. If the bend is indistinguishable from noise, evaluate at your own condition and stop. If it clears noise only narrowly, expect the rule to change selections without reliably improving them. In moderation a plain uniform measure beats our own default ($+87\%$ against 72%), which is the default used in every experiment here and not chosen after seeing this. The same ordering appears wherever profiles are noisy, because averaging over a range hedges an estimation error that a point mass at any single condition inherits.

4.4 Replications: ninety-six hospitals, one site, moderation, routing

Our largest natural test is eICU-CRD (Pollard et al., 2018): 128,169 ICU stays across 96 hospitals, each with its own case mix and observed mortality, spanning prevalence 0.024–0.178. We fixed a forecast in a timestamped, hashed protocol before touching the data: pooling an order of magnitude more source data than the single-site study would turn its null into a gain. It did not. At full source the integral beats a barycentre-matched point mass by 2.4%, but the rules agree on every hospital in six of eight replications, so that average sits on a bootstrap distribution with an atom at zero. A point mass at the source prevalence does better (3.2%): our default's barycentre is 0.054 against a median hospital at 0.085, so this is relocation again, not the rule. What moves monotonically is agreement, not gain. Capping the source at 9k, 20k, 40k and 53k rows, the rules agree on 60%, 67%, 83% and 92% of hospitals. More data does not sharpen curvature into an advantage; it makes every rule converge (Appendix I.2).

MIMIC-IV (Johnson et al., 2023) is the same picture at one site: the rules select differently on 34% of cells, yet the pooled reduction is -4.0% $[-16.3, 6.6]$ and every care unit's interval covers zero. Baselines using information OPERA does not, including label-shift estimation at the deployment's own prevalence (Lipton et al., 2018), do worse still, because the estimated profile no longer describes the unit it is read at (Appendix I.1). Moderation is the paper's clearest aggregation gain, $+72\%$ $[38, 91]$ over 30 replications on two orders of magnitude more data (Appendix I.3). Routing on RouterBench (Hu et al., 2024) bends without paying: the rules differ on 20% of shortlists, but at the observed task composition the integral loses 2.2% and across willingness-to-pay ranges the gain runs from 0.3% to 41.7% (Appendix I.4).

5 Discussion

Evaluation under shift needs a function, not a number. The risk profile is the minimal object a deployment-aware rule can depend on, every linear rule is a measure on it, and current practice is a point mass. The *differences* between candidates' profiles decide whether the measure matters: when they are affine, as for any system that cannot reconfigure, every linear rule ranks as the point mass at its own barycentre (Proposition 1), and the only question left is *where* you evaluate. There published numbers fail: deployments sit 0.6–0.8 in total variation from the headline's condition, and no finite suite sees configurations winning on thin regions, which reconfiguration manufactures (Theorems 3, 4). The calibrated diagnostic called all eight systems correctly, and curvature predicts divergence: aggregation paid on RLHF pipelines and moderation, not on care units or routing. When the system reconfigures and the diagnostic clears its noise floor, integrate against the deployment distribution, or a full-support default (Examples 1, 2) when it cannot be elicited. Every measurable gain required a deployer already knowing its condition; estimating preferences from a hundred labels erased it. Naming the operating condition is the hard part, not integrating over it (Appendix F.7).

AI USE STATEMENT

We used generative AI tools (Claude) to edit manuscript, write and review the code, to draft and check several appendix derivations, and to assist with running and monitoring experiments. The research questions, the method, and the experimental design are the authors'. All derivations were verified by the authors, all numbers are generated from logged results by scripts the authors reviewed, and the authors take full responsibility for the content.

REFERENCES

N. M. Adams and D. J. Hand. Comparing classifiers when the misallocation costs are uncertain. *Pattern Recognition*, 32(7):1139–1147, 1999.

P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

J. Blanchet and K. Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. *WWW Companion*, 2019.

A. Buja, W. Stuetzle, and Y. Shen. Loss functions for binary class probability estimation and classification: Structure and applications. *Technical report, University of Pennsylvania*, 2005.

L. Chen, M. Zaharia, and J. Zou. Frugalgpt: How to use large language models while reducing cost and improving performance, 2023.

W.-L. Chiang, L. Zheng, Y. Sheng, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *ICML*, 2024.

G. Cui, L. Yuan, N. Ding, et al. Ultrafeedback: Boosting language models with scaled ai feedback. In *ICML*, 2024.

C. Drummond and R. C. Holte. Cost curves: An improved method for visualizing classifier performance. *Machine Learning*, 65(1):95–130, 2006.

J. C. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *Annals of Statistics*, 49(3):1378–1406, 2021.

W. Ehm, T. Gneiting, A. Jordan, and F. Krüger. Of quantiles and expectiles: consistent scoring functions, choquet representations and forecast rankings. *JRSS-B*, 78(3):505–562, 2016.

A. N. Elmachtoub and P. Grigas. Smart "predict, then optimize". *Management Science*, 68(1):9–26, 2022.

G. A. Flores, A. H. Smith, J. A. Fukuyama, and A. C. Wilson. Aligning evaluation with clinical priorities: Calibration, label shift, and error costs. *NeurIPS 2025*, 2025.

H. Föllmer and A. Schied. Convex measures of risk and trading constraints. *Finance and Stochastics*, 6(4):429–447, 2002.

C. Fourrier, N. Habib, A. Lozovskaya, K. Szafer, and T. Wolf. Open llm leaderboard v2. Hugging Face, 2024. URL https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard.

R. Gao and A. Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

I. Gilboa and D. Schmeidler. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2):141–153, 1989.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *JASA*, 102 (477):359–378, 2007.

D. J. Hand. Measuring classifier performance: a coherent alternative to the area under the roc curve. *Machine Learning*, 77(1):103–123, 2009.

J. Hernández-Orallo, P. Flach, and C. Ferri. A unified view of performance metrics: translating threshold choice into expected classification loss. *JMLR*, 13:2813–2869, 2012.

J. Hernández-Orallo, P. Flach, and C. Ferri. Roc curves in cost space. *Machine Learning*, 93(1): 71–91, 2013.

Q. J. Hu, J. Bieker, X. Li, N. Jiang, B. Keigwin, G. Ranganath, K. Keutzer, and S. K. Upadhyay. Routerbench: A benchmark for multi-llm routing system, 2024.

A. E. W. Johnson, L. Bulgarelli, L. Shen, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. A. Earnshaw, I. S. Haque, S. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn, and P. Liang. Wilds: A benchmark of in-the-wild distribution shifts. In *ICML*, 2021.

P. Liang, R. Bommasani, T. Lee, et al. Holistic evaluation of language models. *TMLR*, 2023.

Z. Lipton, Y.-X. Wang, and A. Smola. Detecting and correcting for label shift with black box predictors. in. In *ICML*, 2018.

LMArena. arena-human-preference-140k. Hugging Face dataset, CC-BY-4.0, 2025. URL https://huggingface.co/datasets/lmarena-ai/arena-human-preference-140k.

P. Mohajerin Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1–2):115–166, 2018.

Team OLMo. 2 OLMo 2 furious, 2024.

I. Ong, A. Almahairi, V. Wu, W.-L. Chiang, T. Wu, J. E. Gonzalez, M. W. Kadous, and I. Stoica. Routellm: Learning to route llms with preference data, 2024.

T. J. Pollard, A. E. W. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5:180178, 2018.

M. Rofin, V. Mikhailov, M. Florinsky, A. Kravchenko, T. Shavrina, E. Tutubalina, D. Karabekyan, and E. Artemova. Vote'n'rank: Revision of benchmarking with social choice theory. in. *Proceedings of EACL*, 2023.

M. Saerens, P. Latinne, and C. Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural Computation*, 14(1):21–41, 2002.

M. J. Schervish. A general method for comparing probability assessors. *Annals of Statistics*, 17(4): 1856–1879, 1989.

D. Schmeidler. Subjective probability and expected utility without additivity. *Econometrica*, 57(3): 571–587, 1989.

M. Sugiyama, M. Krauledat, and K.-R Müller. Covariate shift adaptation by importance weighted cross validation. *JMLR*, 8:985–1005, 2007.

Qwen Team. Qwen2.5 technical report, 2024.

A. J. Vickers and E. B. Elkin. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574, 2006.

H. Wang, W. Xiong, T. Xie, H. Zhao, and T. Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Findings of EMNLP, pages 10582–10592*, 2024a.

Z. Wang, Y. Dong, O. Delalleau, et al. Helpsteer2: Open-source dataset for training top-performing reward models. *NeurIPS Datasets and Benchmarks Track*, 2024b.

J. Wu, B. Gu, R. Zhou, et al. Bridge: benchmarking large language models for understanding real-world clinical practice texts. *Nature Biomedical Engineering*, 2026.

S. M. Xie, H. Pham, X. Dong, N. Du, H. Liu, Y. Lu, P. Liang, Q. V. Le, T. Ma, and A. W. Yu. Doremi: Optimizing data mixtures speeds up language model pretraining, 2023.

Appendix Contents

A Notation . . . 13
B Related Work . . . 13
C Proofs of the Main Results . . . 15
D Robust Aggregation: Duality, Hierarchy, and Alternative Geometries . . . 19
E Symmetry, Universality, and Finite-Sample Theory . . . 23
F Relocated Statements, Scope, and Practitioner Guidance . . . 27
G Leaderboard Experiments: Blind Sets, Personas, and Logged Traffic . . . 31
H RLHF Serving Pipelines . . . 37
I Clinical and Safety Replications . . . 40

A Notation

Deployment space and systems	
Ω	compact metric space of *operating conditions*; $\omega \in \Omega$ a single condition
$\pi,\ c,\ w$	prevalence, cost ratio, task mixture, the instantiations of ω used here
Δ^{T-1}	the T-task mixture simplex
$s,\ s_1, \ldots, s_K$	decision systems; a shortlist of K candidates
$\tau \in \Theta_s$	deployment-time configuration (threshold, routing rule, best-of-n budget) and the set s can reach
$R_\omega(s, \tau)$	deployment risk of s in configuration τ at condition ω
$r_s(\omega)$	*envelope risk profile*, $\inf_\tau R_\omega(s, \tau)$; the object of the paper (2)
FPR, FNR	false-positive / false-negative rates; n_0, n_1 their class sample sizes
Aggregation rules	
$C(\Omega),\ \mathcal{P}(\Omega)$	continuous functions on Ω; Borel probability measures on Ω
Φ	an evaluation functional, or *aggregation rule*, $C(\Omega) \to \mathbb{R}$; selection is $\arg\min_k \Phi(r_{s_k})$
$\mu,\ \mu_0$	criterion measure (Theorem 2); μ_0 the reference measure
δ_ω	Dirac measure at ω, the geometry of every point-mass criterion
μ_{LOU}	log-odds-uniform measure, $d\mu \propto d\pi/[\pi(1-\pi)]$; the invariant default
$\Phi_\mu(r)$	the integral rule $\int_\Omega r\, d\mu$ (OPERA)
$\mathcal{M},\ \mathcal{M}_\varepsilon$	uncertainty geometry; the W_1 ball $\{\mu : W_1(\mu, \mu_0) \le \varepsilon\}$ of radius ε
$[r]^\lambda$	Lipschitz regularization $\sup_{\omega'}\{r(\omega') - \lambda\, d(\omega', \omega)\}$, from the W_1 dual (4)
$\widehat{\Phi}_\varepsilon$	first-order robust criterion $\Phi_0 + \varepsilon\Lambda$ (5) (R-OPERA as implemented)
ε_{sat}	radius at which the exact robust value saturates at minimax
Summaries and guarantees	
$\Lambda(r_s; \mu_0)$	*deployment sensitivity*: derivative of the robust value at $\varepsilon = 0$
$\Gamma(s)$	*risk–sensitivity pair* $(\Phi_0(s),\ \Lambda(r_s; \mu_0))$, integrated risk and sensitivity (9)
$\text{Lip}(r_s),\ L_{\max}$	Lipschitz constant of a profile; the largest over a shortlist
$\Omega_0,\ h(\Omega_0)$	the finitely many conditions a local criterion consults, and their covering radius
δ_{grid}	discretization error of an implemented rule (its own blind set)
$\gamma,\ \bar{\sigma}$	selection margin; mean estimation error of the profile
$G,\ K$	size of the evaluation grid; size of the candidate shortlist

B Related Work

This appendix extends the related-work discussion of Section 1: first the neighbouring literatures, then exactly what is inherited from the cost-curve literature and what is added.

Scoring rules. Schervish (1989) proves that a regular scoring rule for probability forecasts is proper if and only if it is a mixture of cost-weighted zero–one losses against a nonnegative measure on the cost–loss ratio; Buja et al. (2005) restate this as a Choquet representation and Gneiting and Raftery (2007) as their Theorem 3, and Ehm et al. (2016) extend mixture representations to

quantiles and expectiles and plot the resulting family (Murphy diagrams) to obtain dominance statements. These have the shape of Theorem 2(a), an integral of elementary losses against a measure, but the integration runs over thresholds of a *forecast*; here it runs over the conditions a *deployment* may meet, and the object integrated is achievable risk after reconfiguration. Decision curve analysis (Vickers and Elkin, 2006) is the same integral at a single threshold probability with an affine rescaling.

Risk measures and robust decisions. Theorem 2(b) transplants the dual representations of coherent and convex risk measures (Artzner et al., 1999; Föllmer and Schied, 2002) and of multiple-prior decision theory (Schmeidler, 1989; Gilboa and Schmeidler, 1989); R-OPERA's dual is Wasserstein distributionally robust optimization (Mohajerin Esfahani and Kuhn, 2018; Blanchet and Murthy, 2019; Gao and Kleywegt, 2023), and ϕ-divergence geometries relate to Duchi and Namkoong (2021). We import these results rather than extend them; what is new is their use to classify *evaluation* procedures and the completeness analysis of Appendix D.2.

Shift and decisions. Label-shift estimation (Saerens et al., 2002; Lipton et al., 2018) and importance-weighted validation (Sugiyama et al., 2007) target one known deployment distribution; the collapse boundary says when that suffices (non-reconfigurable systems, where evaluating at the right point is exact) and when it does not. Decision-focused learning (Elmachtoub and Grigas, 2022) trains a predictor for a downstream decision; our envelope instead lets the deployer re-optimize the decision at deployment and asks how to evaluate systems that will be so re-optimized.

Benchmark aggregation. Social-choice aggregation (Rofin et al., 2023) replaces the mean over tasks with voting rules more robust to outlying tasks; Bradley–Terry and Elo ratings (Chiang et al., 2024) and item-response models estimate a latent capability while accounting for task difficulty. All remain fixed rules over a fixed benchmark, and none changes the operating condition at which the rule evaluates: an Elo rating fitted to pooled arena traffic is still a point evaluation, at that traffic's mixture rather than a deployment's.

Adaptive serving. Cascades that escalate only when a cheap model looks unreliable (Chen et al., 2023) and routers learned from preference data (Ong et al., 2024) are exactly the reconfigurable systems whose profiles this paper evaluates; the cost-quality curves they report are objects our boundary says to compare rather than to read at one point. Reweighting *training* mixtures (Xie et al., 2023) and group DRO optimise a worst case or a trade-off during fitting, whereas we select among already-trained systems. Benchmarks of in-the-wild shift (Koh et al., 2021) supply deployment conditions but not a rule for aggregating over them, and are a natural next testbed for the diagnostic.

Relation to the cost-curve literature. The envelope risk profile (2) is the lower envelope of the cost curves of Drummond and Holte (2006), whose Section 5 already selects over configurations more general than a threshold. Hand (2009) and Hernández-Orallo et al. (2012) write it as an integral, Adams and Hand (1999) elicit a distribution over operating points, and Hernández-Orallo et al. (2012) integrate over an "operating condition" tuple that is our Ω. All of this lives on a one-dimensional binary cost axis: Hernández-Orallo et al. (2012) pose a multivariate $w(\theta)$ but reduce it to a one-dimensional skew, and Hernández-Orallo et al. (2013) treat thresholded classifiers. We carry the same object to an arbitrary compact Ω and general configuration sets, which is where task mixtures, preference weightings, routing and best-of-n selection live. That criterion design is measure design is Hand's point, since one "must choose a function $w(c)$ which does not depend on the score distributions"; Buja et al. (2005) make it constructive, and on the cost axis Schervish's converse characterises the proper criteria. Theorem 2 gives the corresponding characterisation over any Ω. That is Riesz–Markov–Kakutani, and we treat it as bookkeeping. The AUC-improperness argument of Section 2.1 is also Hand's. What this literature lacks is any account of when the choice of measure is irrelevant, or of what a finitely supported criterion misses. Proposition 1 supplies the first, as a scope statement on difference profiles, and it is the claim we test most directly (Section 4.2). Theorems 3 and 4, with Proposition 3, supply the second.

C Proofs of the Main Results

C.1 Geometry and representation

This appendix proves the classification of aggregation rules (Theorem 2) and the optimality of integration.

C.2 Representation Theorem (Theorem 2)

Proof. ($\Leftarrow$) Given μ, define $\Phi(r) = \int r\, d\mu$. Monotonicity: if $r_1 \le r_2$ pointwise, then $\int r_1\, d\mu \le \int r_2\, d\mu$ since $\mu \ge 0$. Linearity: linearity of integration. Continuity: $|\Phi(r_1) - \Phi(r_2)| \le \|r_1 - r_2\|_\infty \cdot \mu(\Omega)$.

($\Rightarrow$) Φ is a positive (by monotonicity + linearity: $r \ge 0 \implies \Phi(r) \ge \Phi(0) = 0$), linear, continuous functional on $C(\Omega)$ where Ω is compact Hausdorff. By the Riesz–Markov–Kakutani representation theorem, there exists a unique finite signed regular Borel measure ν such that $\Phi(r) = \int r\, d\nu$. Positivity of Φ forces $\nu \ge 0$. Set $\mu = \nu$. □

C.3 Convex Geometries (Theorem 2(b))

Proof sketch. This is the convex/coherent risk-measure duality of Föllmer and Schied (2002) and Artzner et al. (1999) (see also Gilboa and Schmeidler, 1989; Schmeidler, 1989), stated on $C(\Omega)$ rather than L^∞. Monotonicity plus translation-equivariance make Φ 1-Lipschitz in the sup-norm; convexity and continuity allow the Fenchel–Moreau biconjugation $\Phi(r) = \sup_\nu\{\langle r, \nu\rangle - \Phi^*(\nu)\}$ over the dual $(C(\Omega))^* =$ finite signed Radon measures (Riesz). Monotonicity forces the effective domain of Φ^* into nonnegative measures; translation-equivariance forces total mass one, hence $\nu \in \mathcal{P}(\Omega)$, and $\alpha := \Phi^*|_{\mathcal{P}(\Omega)}$. Positive homogeneity makes Φ^* an indicator, i.e. $\alpha \in \{0, \infty\}$, giving the support-function form over $\mathcal{M} = \{\mu : \alpha(\mu) = 0\}$, closed and convex. □

C.4 Optimality of integration

Proposition 2 (Optimality of Φ_μ)**.** *Let μ be a probability measure on Ω.* ***(a) Bayes optimality.*** *If the deployment condition is drawn from μ, then any Φ_μ-selection minimises μ-expected risk, and hence μ-expected regret, among* all *rules that select without observing the realised condition; in particular among all linear aggregation rules. When the* $\arg\min$ *is not unique every tie-breaking choice attains the same value, so the statement is about the value, not about a distinguished candidate.* ***(b) Crossing improvement.*** *If μ has full support (the proof uses only that μ charges every open set; no reference measure, hence no notion of absolute continuity, is needed on a general Ω) and two candidates have crossing profiles ($r_{s_j}(\omega_1) < r_{s_k}(\omega_1)$ and $r_{s_j}(\omega_2) > r_{s_k}(\omega_2)$), then integration against μ has μ-expected regret no larger than that of any point-mass rule at a single ω_0, and strictly smaller exactly when the candidate the point mass selects has strictly larger μ-expected risk than the Φ_μ-optimal value. Selecting a different candidate is not enough: two candidates can have equal μ-expected risk, and then the rules differ in their pick and not in their regret. Crossing alone does not give strict improvement: a point mass on the side of the crossing that μ predominantly weights selects the same candidate and incurs identical regret.*

Proof. (a) A deployment-blind rule commits to one candidate k before ω is drawn, so its μ-expected risk is $\int r_{s_k} d\mu$ and its μ-expected regret is $\int r_{s_k} d\mu - \mathbb{E}_{\omega\sim\mu}[\min_j r_{s_j}(\omega)]$, which differ by a constant. Minimising over k gives $\arg\min_k \int r_{s_k} d\mu$, which is exactly the Φ_μ-selection, and any tie attains the same integral. Restricting to linear rules is then a special case: by Theorem 2 each is Φ_ν for some ν, and it too commits before ω is drawn. (b) Part (a) gives the weak inequality against every other selection rule. Regret depends on a rule only through the candidate it selects, so equality holds precisely when the point mass returns the Φ_μ-optimal candidate. Strictness therefore requires a hypothesis locating ω_0 on the side of the crossing opposite the bulk of the μ-mass, and does not follow from crossing alone. □

C.5 MINIMALITY THEOREM

Proof of Theorem 1. (a) By Axiom 1, $E(s) = F(R_s)$ for some functional F of the risk surface. By definition of the envelope, $r_s = r_{s'}$ if and only if $s \sim s'$, so the map $s \mapsto r_s$ is a complete invariant of deployment equivalence: it induces a bijection between equivalence classes and realized profiles. Axiom 2 says E is constant on equivalence classes; a map constant on the fibers of a quotient factors uniquely through the quotient, i.e. $E(s) = \Phi(r_s)$ with Φ defined on realized profiles by $\Phi(r) := E(\text{any } s \text{ with } r_s = r)$, well-defined by invariance.

(b) For each $\omega \in \Omega$, the point evaluation $\Phi_\omega(r) = r(\omega)$ is monotone, positively linear, and sup-norm continuous, an admissible aggregation rule (Theorem 2 with $\mu = \delta_\omega$). If every admissible aggregation rule factors through T, then $T(s) = T(s')$ implies $\Phi_\omega(r_s) = \Phi_\omega(r_{s'})$ for every ω, i.e. $r_s(\omega) = r_{s'}(\omega)$ pointwise. Hence T determines the profile. □

C.6 BLIND-SPOT THEOREM, IN FULL

Throughout, Ω is compact metric with no isolated points, metric d_Ω, and $B(\omega, \rho)$ denotes the open ball. Section 3 states the three requirements of Theorem 3 in prose; here they are, formally.

Axiom 3 (Regularity)**.** Φ is monotone, positively linear, and sup-norm continuous.

Axiom 4 (Locality)**.** There is a finite $\Omega_0 \subset \Omega$ such that $r = r'$ on Ω_0 implies $\Phi(r) = \Phi(r')$: evaluation consults the profile at finitely many operating conditions.

Axiom 5 (Deployment completeness)**.** If $r_1 \leq r_2$ with strict inequality on some nonempty open $U \subseteq \Omega$, then $\Phi(r_1) < \Phi(r_2)$: a system that is strictly better on a non-negligible set of deployment conditions receives a strictly better score.

Remark 2 (Reading the axioms). Every criterion in the first block of Table 3 satisfies the first two, and so does any benchmark suite, however many benchmarks it contains. The third is what a deployment stakeholder means by "the evaluation sees where my system will run." The three axioms are sufficient for Theorem 3 without being minimal.

Lemma 1 (Locality pins the support)**.** *If Φ satisfies Axioms 3 and 4 with witness set Ω_0, then its representing measure (from Theorem 2) satisfies* $\operatorname{supp}\mu \subseteq \Omega_0$.

Proof. Suppose $\omega^* \in \operatorname{supp}\mu \setminus \Omega_0$. Since Ω_0 is finite (hence closed) and $\omega^* \notin \Omega_0$, there is $\rho > 0$ with $B(\omega^*, \rho) \cap \Omega_0 = \emptyset$. Take the continuous bump $\varphi(\omega) = \max\{0,\, 1 - d_\Omega(\omega, \omega^*)/\rho\}$, which vanishes on Ω_0 and is positive on $B(\omega^*, \rho)$. Then $\varphi = 0$ on Ω_0, so locality forces $\Phi(\varphi) = \Phi(0) = 0$; but $\omega^* \in \operatorname{supp}\mu$ gives $\mu(B(\omega^*, \rho/2)) > 0$ and hence $\Phi(\varphi) = \int \varphi\, d\mu \geq \frac{1}{2}\mu(B(\omega^*, \rho/2)) > 0$, a contradiction. □

Proof of Theorem 3. Let Φ satisfy Axioms 3 and 4. By Theorem 2 and Lemma 1, $\Phi = \int \cdot\, d\mu$ with $\operatorname{supp}\mu \subseteq \Omega_0$ finite.

Blind set. Since Ω has no isolated points and Ω_0 is finite, $U := \Omega \setminus \Omega_0$ is open, nonempty, and $\mu(U) = 0$. Pick $\omega_c \in U$ and $\rho > 0$ with $\overline{B(\omega_c, \rho)} \subseteq U$; for $0 < h \leq \rho$ define the 1-Lipschitz bump $\beta_h(\omega) = \max\{0,\, h - d_\Omega(\omega, \omega_c)\}$, supported in $B(\omega_c, h) \subseteq U$.

Failure of completeness. For any profile $r_1 \geq 0$ set $r_2 := r_1 + \beta_h$. Then $r_1 \leq r_2$ with strict inequality on the open set $B(\omega_c, h)$, yet $\Phi(r_2) - \Phi(r_1) = \int \beta_h\, d\mu = 0$ since $\mu(B(\omega_c, h)) = 0$. Axiom 5 fails, and $r_2 - r_1 \geq h/2$ on $B(\omega_c, h/2)$: deployment there incurs regret $\geq h/2$ registered as zero. Both profiles are 1-Lipschitz whenever r_1 is (take $r_1 \equiv$ const).

Realizability of the witnesses. The bump β_h is convex at the boundary of its support, so $r_1 + \beta_h$ is not concave and is not the envelope of any system (Lemma 4(iii)). The construction is repaired inside the realizable cone by Lemma 2 below, which supplies envelope profiles with the same conclusion; the bump version is retained only because it gives the shortest proof that Axioms 4 and 5 are incompatible. □

Lemma 2 (Realizable witnesses; the usage cell is convex)**.** *Let Ω be compact convex, let $r_1 = \min_{i \leq m} a_i$ be an envelope of affine functions (the profile of a system with m configurations), and let*

a be affine. Then $r_2 := \min(r_1, a)$ is the profile of the system that adds one configuration, and the set on which the two differ,

$$\Delta \;:=\; \{\omega : r_2(\omega) < r_1(\omega)\} \;=\; \bigcap_{i \le m} \{\omega : a(\omega) < a_i(\omega)\},$$

is convex*, an intersection of open half-spaces. Consequently:*

(i) if $\Delta \cap \Omega_0 = \emptyset$ then $r_1 = r_2$ on Ω_0 and no local criterion with witness set Ω_0 can distinguish the two systems;

(ii) for $\dim \Omega = d$ and $m \ge d+1$ pieces arranged as a peak at a point ω_c, Δ can be made a bounded polytope contained in any prescribed ball $B(\omega_c, \rho)$;

(iii) choosing ω_c at maximal distance h from Ω_0 and $\rho = h$ gives realizable profiles, both L-Lipschitz, that agree on Ω_0 and differ by $\kappa_m L h$ at ω_c, where κ_m is the inradius of the polytope spanned by the peak's unit gradient directions: $\kappa_{d+1} = 1/d$ for a regular simplex and $\kappa_{2d} = 1/\sqrt{d}$ for a cross-polytope, increasing to 1 as m grows.

Proof. $a - r_1$ is convex (affine minus concave), so $\Delta = \{a - r_1 < 0\}$ is a sublevel set of a convex function, hence convex; the displayed intersection form is immediate from $r_1 = \min_i a_i$.

For (ii) and (iii) we give the construction explicitly. Work in the affine hull of Ω, of dimension d, fix $\omega_c \in \Omega$ and unit vectors $v_1, \dots, v_m$ whose positive hull is all of $\mathbb{R}^d$, and set

$$a_i(\omega) \;=\; b + L\,\langle v_i, \omega - \omega_c\rangle, \qquad a(\omega) \;=\; b - \eta, \qquad \kappa_m \;:=\; \min_{\|u\|=1} \max_{i \le m} \langle v_i, u\rangle > 0.$$

Each a_i has gradient of norm L, so $r_1 = \min_i a_i$ is L-Lipschitz, and so is $r_2 = \min(r_1, a)$ because a is constant; this is the claim that adding the configuration preserves the bound. Writing $y = \omega - \omega_c$,

$$r_1(\omega) \;=\; b + L \min_i \langle v_i, y\rangle \;=\; b - L \max_i \langle -v_i, y\rangle \;\le\; b - L\,\kappa_m \|y\|$$

for every $y \neq 0$: evaluating the definition of κ_m at the unit vector $u = -y/\|y\|$ gives $\max_i \langle v_i, -y/\|y\|\rangle \ge \kappa_m$, i.e. $\max_i \langle -v_i, y\rangle \ge \kappa_m \|y\|$. No symmetry of the $\{v_i\}$ is needed, and indeed the regular simplex directions are not symmetric under negation. Hence

$$\Delta \;=\; \{r_1 > b - \eta\} \;\subseteq\; \{\,\kappa_m L \|y\| < \eta\,\} \;=\; B\Big(\omega_c, \tfrac{\eta}{\kappa_m L}\Big),$$

so choosing $\eta = \kappa_m L h$ confines Δ to $B(\omega_c, h)$. Only the containment is used: for the cross-polytope directions the cell is a box, not a Euclidean ball, and norms throughout are Euclidean, which is what fixes the constants below. And the two profiles differ at ω_c by exactly $r_1(\omega_c) - r_2(\omega_c) = \eta = \kappa_m L h$. Taking ω_c at maximal distance $h = h(\Omega_0)$ from Ω_0 puts no point of Ω_0 inside $B(\omega_c, h)$, so $r_1 = r_2$ on Ω_0, which is (i) and (iii). If ω_c lies on the relative boundary of Ω the same argument applies with $\Delta \cap \Omega$ in place of Δ; only the containment is used.

Two families fix the constants. The cross-polytope directions $\{\pm e_j\}_{j \le d}$ give $m = 2d$ pieces and $\kappa_{2d} = \min_{\|u\|=1} \max_j |u_j| = 1/\sqrt{d}$, attained at $u = d^{-1/2}(1, \dots, 1)$; r_2 then has $2d+1$ configurations, and the gap is $Lh/\sqrt{d}$, which is the bound quoted in Proposition 3(ii). The $d+1$ vertex directions of a regular simplex give $\kappa_{d+1} = 1/d$. As m grows with the v_i equidistributed on the unit sphere, $\kappa_m \to 1$, since $\max_i \langle v_i, u\rangle \to \max_{\|v\|=1} \langle v, u\rangle = 1$ uniformly in u; the gap the construction achieves therefore approaches Lh, which is the trivial upper bound for an L-Lipschitz pair agreeing at distance h. □

Proposition 3 (Covering-radius blind-spot bound)**.** *Let profiles be L-Lipschitz and let Ω_0 have covering radius $h(\Omega_0) := \max_{\omega \in \Omega} \min_{\omega_0 \in \Omega_0} d_\Omega(\omega, \omega_0)$. Then*

(i) (upper bound) *if $r_1 = r_2$ on Ω_0 then $\|r_1 - r_2\|_\infty \le 2L\,h(\Omega_0)$: a suite reporting Ω_0 is $2Lh$-complete;*

(ii) (lower bound) *on a convex Ω of dimension d there exist realizable profiles, of $2d$ and $2d+1$ configurations respectively, agreeing on Ω_0 and differing by $L\,h(\Omega_0)/\sqrt{d}$ (Lemma 2); the upper and lower bounds therefore match up to a factor $2\sqrt{d}$ on the realizable class;*

(iii) (sufficiency) *on the d-dimensional mixture simplex, for every $\eta > 0$ there is a set Ω_0 with $h(\Omega_0) \le \eta/(2L)$ and $|\Omega_0| \le C_d(2L/\eta)^d$, and by (i) any such set is η-complete;*

(iv) (necessity) *conversely, if Ω_0 is η-complete for the realizable class of (ii), then $h(\Omega_0) \le \sqrt{d}\,\eta/L$ and hence $|\Omega_0| \ge c_d\,(L/(\sqrt{d}\,\eta))^d$.*

The two directions differ by the factor $2\sqrt{d}$ of (ii), so the count G needed for η-completeness is exponential in d either way. Here c_d, C_d are the packing and covering constants of the simplex.

Proof. (i) For any ω pick the nearest $\omega_0 \in \Omega_0$; then $|r_1(\omega) - r_2(\omega)| \le |r_1(\omega) - r_1(\omega_0)| + |r_2(\omega_0) - r_2(\omega)| \le 2L\,d(\omega, \omega_0)$, and taking the maximum over ω gives $2Lh$. (ii) is Lemma 2(iii). (iii) is the standard covering estimate for a d-dimensional convex body: a maximal ρ-separated set is a ρ-net and has cardinality at most $C_d\rho^{-d}$; take $\rho = \eta/(2L)$. (iv) is its converse. If $h(\Omega_0) > \sqrt{d}\,\eta/L$, apply (ii) at the point realising the covering radius to obtain two realizable profiles that agree on Ω_0 yet differ by $Lh(\Omega_0)/\sqrt{d} > \eta$, contradicting η-completeness; the cardinality bound is then the packing lower bound $|\Omega_0| \ge c_d h(\Omega_0)^{-d}$, valid for a full-dimensional compact convex body with nonempty relative interior, with c_d depending on its geometry and normalisation. □

Remark 3 (The honest reading of the rate). Part (iii) is a *worst-case* statement over L-Lipschitz profiles. Real systems need not attain it, and in Section G.1 we measure that they do not: on real serving pipelines the largest invisible gap decays with exponent -0.75 rather than the worst-case $-1/d = -0.25$, because achievable sets of real pipelines are low-complexity. What survives measurement is the *dimensional mechanism*, the measured exponent tracks $1/d$ across $d = 1, \dots, 4$ with correlation 0.98, and the practical conclusion that the suite sizes actually in use are one to two orders of magnitude too small.

Proof of Corollary 1. ($\Leftarrow$) If $\operatorname{supp}\mu = \Omega$ and $r_1 \le r_2$ with strict inequality on open $U \ne \emptyset$, then by continuity there is an open $V \subseteq U$ and $\eta > 0$ with $r_2 - r_1 \ge \eta$ on V, whence $\Phi(r_2) - \Phi(r_1) \ge \eta\,\mu(V) > 0$ by full support. Axioms 3 and 5 hold. ($\Rightarrow$) If $\operatorname{supp}\mu \ne \Omega$, the open set $U = \Omega \setminus \operatorname{supp}\mu$ is nonempty with $\mu(U) = 0$, and the bump construction above violates Axiom 5. □

Lemma 3 (Single-atom instability)**.** *For any non-isolated $\omega_0 \in \Omega$ and every $\varepsilon > 0$, there exist 1-Lipschitz risk profiles r_1, r_2 and a measure ν_ε with $W_1(\delta_{\omega_0}, \nu_\varepsilon) < \varepsilon$ such that: (i) point-mass at ω_0 selects model 1; (ii) under ν_ε, model 2 is uniquely optimal; (iii) the regret is at least $\frac{1}{2}W_1(\delta_{\omega_0}, \nu_\varepsilon)$.*

Proof. Since ω_0 is non-isolated, choose $\omega_\varepsilon \in \Omega \setminus \{\omega_0\}$ with $\Delta_\varepsilon := d_\Omega(\omega_\varepsilon, \omega_0) < \varepsilon$. Set $\nu_\varepsilon := \delta_{\omega_\varepsilon}$. Define $r_1(\omega) := \Delta_\varepsilon/2$ (constant) and $r_2(\omega) := d_\Omega(\omega, \omega_\varepsilon)$ (1-Lipschitz). At ω_0: $r_1 = \Delta_\varepsilon/2 < \Delta_\varepsilon = r_2$, so point-mass selects model 1. Under ν_ε: $\int r_1 = \Delta_\varepsilon/2$ and $\int r_2 = 0$, so model 2 is optimal. Regret: $\Delta_\varepsilon/2 = \frac{1}{2}W_1(\delta_{\omega_0}, \nu_\varepsilon)$. □

C.7 Envelope risk profiles: properties

Definition 1 (Score envelope risk profile)**.** For scoring function $s : \mathcal{X} \to [0, 1]$ and decision rule $\delta_{s,\tau}(x) = \mathbf{1}(s(x) \ge \tau)$: $r_s(c, \pi) = \inf_\tau [c(1-\pi)\,\mathrm{FPR}(\tau) + (1-c)\pi\,\mathrm{FNR}(\tau)]$.

Lemma 4 (Properties)**.** *For any scoring function s: (i) $r_s \in C(\Omega)$; (ii) $r_s \ge 0$; (iii) r_s is concave in π (infimum of affine functions), hence **not** generally linear, the optimal threshold $\tau^*(c, \pi)$ moves with π; (iv) assume* label shift*, that the class-conditional score distributions $P(s \mid Y)$ are the same at π_0 and at π; the odds update below uses exactly this. Then the optimal threshold depends on the distribution at which s is calibrated. If s is calibrated at prevalence π_0, that is $s(x) = P_{\pi_0}(Y = 1 \mid x)$, then at deployment prevalence π*

$$\tau^*(c, \pi) \;=\; \frac{K}{1+K}, \qquad K \;=\; \frac{c}{1-c} \cdot \frac{1-\pi}{\pi} \cdot \frac{\pi_0}{1-\pi_0}.$$

Two cases are worth naming. For a score calibrated at the balanced distribution $\pi_0 = 1/2$ this is $\tau^* = c(1-\pi)/[(1-c)\pi + c(1-\pi)]$. For a score recalibrated at the deployment prevalence itself, $\pi_0 = \pi$, it is the constant $\tau^* = c$: the threshold on the posterior *scale stops moving, while the induced threshold on any fixed score scale still moves with π, which is what (iii) requires.*

Proof. For fixed τ the map $\pi \mapsto f_\pi(\tau) = c(1-\pi)\mathrm{FPR}(\tau) + (1-c)\pi\mathrm{FNR}(\tau)$ is affine with slope $(1-c)\mathrm{FNR}(\tau) - c\,\mathrm{FPR}(\tau) \in [-c, 1-c]$, so the family is uniformly Lipschitz in π with constant $\max(c, 1-c)$. An infimum of a uniformly Lipschitz family is Lipschitz with the same constant, which gives (i); it is nonnegative because rates lie in $[0,1]$ and the coefficients are nonnegative, which gives (ii); and an infimum of affine functions is concave, which gives the first half of (iii). It is not generally affine because the minimising τ moves with π, and the envelope is affine on exactly the region where one τ stays optimal. □

Proof of (iv). The risk is $R = c\,P(\hat{Y}=1, Y=0) + (1-c)\,P(\hat{Y}=0, Y=1) = \mathbb{E}_\pi[\,c(1-\eta_\pi(X))\mathbf{1}\{\hat{Y}=1\} + (1-c)\eta_\pi(X)\mathbf{1}\{\hat{Y}=0\}\,]$ with $\eta_\pi(x) = P_\pi(Y=1 \mid x)$, so the Bayes rule predicts 1 iff $\eta_\pi(x) > c$. Bayes' rule relates the posteriors at the two prevalences through the odds: $\frac{\eta_\pi}{1-\eta_\pi} = \frac{\eta_{\pi_0}}{1-\eta_{\pi_0}} \cdot \frac{\pi}{1-\pi} \cdot \frac{1-\pi_0}{\pi_0}$. Setting $\eta_\pi = c$ and solving for $\eta_{\pi_0} = s$ gives odds K and hence $\tau^* = K/(1+K)$. □

Remark 4 (Contrast with fixed rules). For a fixed decision rule (fixed threshold), $R_{c,\pi}(\delta) = \alpha\pi + \beta$ is affine in π and Proposition 1 applies: integration collapses to the barycenter. The envelope escapes the collapse because the threshold adapts. This is the boundary of the theory made concrete.

D Robust Aggregation: Duality, Hierarchy, and Alternative Geometries

D.1 R-OPERA duality and the Bayesian-to-minimax hierarchy

Example 2 (R-OPERA: a robust variant). Take $\mathcal{M}_\varepsilon = \{\mu : W_1(\mu, \mu_0) \le \varepsilon\}$ around $\mu_0 = \mu_{\mathrm{LOU}}$. Wasserstein duality gives the *exact* robust value (Theorem 5, Appendix D.1)

$$\Phi_{\mathcal{M}_\varepsilon}(r_s) \;=\; \inf_{\lambda \ge 0} \Big\{ \lambda\varepsilon + \int [r_s]^\lambda \, d\mu_0 \Big\}, \qquad [r]^\lambda(\pi) = \sup_{\pi'} \{ r(\pi') - \lambda|\pi' - \pi| \}, \tag{4}$$

which is concave and non-decreasing in ε, with derivative $\Lambda(r_s; \mu_0)$ at $\varepsilon = 0$, so

$$\widehat{\Phi}_\varepsilon(s) \;:=\; \underbrace{\int r_s \, d\mu_0}_{\text{integrated risk}} \;+ \varepsilon \cdot \underbrace{\Lambda(r_s; \mu_0)}_{\text{sensitivity}} \tag{5}$$

is the *first-order* robust criterion: an upper bound on (4), tangent at $\varepsilon = 0$, and the object implementations compute. In words: rank by average risk plus a charge for sensitivity, with ε setting how much caution you buy. It is not a closed form, and Appendix D.5 reports where the two diverge.

Theorem 5 (R-OPERA dual)**.** *Let Ω be a compact metric space and $\mathcal{M}_\varepsilon = \{\mu : W_1(\mu, \mu_0) \le \varepsilon\}$. The duality and the first-order expansion below hold at this generality; it is the finite-sample slope estimator of Corollary 3, not this theorem, that we develop only for a prevalence interval. Write d for the metric, so that $d(\pi', \pi) = |\pi' - \pi|$ on such an interval.*

$$\Phi_{\mathcal{M}_\varepsilon}(r_s) \;=\; \inf_{\lambda \ge 0} \Big\{ \lambda\varepsilon + \int [r_s]^\lambda \, d\mu_0 \Big\}, \qquad [r]^\lambda(\pi) := \sup_{\pi'} \{ r(\pi') - \lambda\, d(\pi', \pi) \}. \tag{6}$$

Define the deployment sensitivity of r relative to μ_0

$$\Lambda(r; \mu_0) \;:=\; \operatorname*{ess\,sup}_{\pi \in \operatorname{supp} \mu_0} \; \sup_{\pi' \ne \pi} \frac{r(\pi') - r(\pi)}{d(\pi', \pi)} \;\le\; \mathrm{Lip}(r), \tag{7}$$

with equality $\Lambda(r; \mu_0) = \mathrm{Lip}(r)$ whenever $\operatorname{supp} \mu_0 = \Omega$, the full-support case that Corollary 1 forces on any deployment-complete criterion. Then:

(i) (upper bound, all ε) $\Phi_{\mathcal{M}_\varepsilon}(r_s) \le \int r_s \, d\mu_0 + \varepsilon\Lambda(r_s; \mu_0)$*, and the dual minimiser satisfies $\lambda^\star \in [0, \Lambda(r_s; \mu_0)]$;*

(ii) (first-order exactness)

$$\Phi_{\mathcal{M}_\varepsilon}(r_s) \;=\; \int r_s \, d\mu_0 + \varepsilon \cdot \Lambda(r_s; \mu_0) \;+\; o(\varepsilon) \qquad (\varepsilon \downarrow 0); \tag{8}$$

(iii) (saturation) $\Phi_{\mathcal{M}_\varepsilon}(r_s) = \sup_\pi r_s$ *for every* $\varepsilon \geq \varepsilon_{\mathrm{sat}}(r_s) := \min_{\pi^\star \in \arg\max r_s} W_1(\delta_{\pi^\star}, \mu_0)$. *This radius is sufficient, not necessarily the smallest at which saturation occurs: when* r_s *has several maximisers, a mixture supported on them is also a maximiser of* $\int r_s \, d\mu$ *and may lie strictly closer to* μ_0 *than any single Dirac mass.*

The linear form (8) *is an upper bound and a first-order expansion,* not *an identity:* $\lambda^\star$ *may lie strictly inside* $(0, \mathrm{Lip}(r))$ *(Remark 5, where* $\lambda^\star = \Lambda = 1 < \mathrm{Lip}(r) = 2$*).*

Proof. Equation (6) is strong duality for Wasserstein distributionally robust optimisation over a compact ground space (Mohajerin Esfahani and Kuhn, 2018; Blanchet and Murthy, 2019; Gao and Kleywegt, 2023). The objective $g(\lambda) = \lambda\varepsilon + \int [r]^\lambda d\mu_0$ is convex in λ, being an infimum-free supremum of affine functions of λ integrated against μ_0. For $\lambda \geq \Lambda(r; \mu_0)$ we have $[r]^\lambda = r$ μ_0-a.e., so $g(\lambda) = \lambda\varepsilon + \int r \, d\mu_0$ is increasing; hence $\lambda^\star \leq \Lambda$, and evaluating g at $\lambda = \Lambda$ gives (i). For (ii), the upper bound is (i), so only a matching lower bound is needed. Fix $\delta > 0$. By definition of $\Lambda := \Lambda(r; \mu_0)$ as an essential supremum, the set $\{\pi \in \mathrm{supp}\, \mu_0 : \sup_{\pi'} (r(\pi') - r(\pi))/|\pi' - \pi| > \Lambda - \delta\}$ has positive μ_0-measure. Since r is continuous, the inner supremum may be restricted to a countable dense set of π', so that set is a countable union over π' of the sets $A_{\pi'} := \{\pi : r(\pi') - r(\pi) > (\Lambda - \delta)|\pi' - \pi|\}$, and therefore $\mu_0(A_{\pi'}) > 0$ for at least one π'. Fix such a π', write $A := A_{\pi'} \setminus \{\pi'\}$ and $C := \int_A |\pi' - \pi| \, d\mu_0(\pi) \in (0, \infty)$.

For $\varepsilon \leq C$ set $\theta := \varepsilon / C \in (0, 1]$ and let μ move a θ-fraction of the mass on A to the point π', leaving the rest of μ_0 in place. This plan has cost $\theta \int_A |\pi' - \pi| d\mu_0 = \varepsilon$, so $W_1(\mu, \mu_0) \leq \varepsilon$ and $\mu \in \mathcal{M}_\varepsilon$. Its value is

$$\int r \, d\mu - \int r \, d\mu_0 \;=\; \theta \int_A \big(r(\pi') - r(\pi)\big) d\mu_0(\pi) \;>\; \theta(\Lambda - \delta) \int_A |\pi' - \pi| d\mu_0 \;=\; (\Lambda - \delta)\, \varepsilon,$$

using the defining inequality of A pointwise. Hence $\Phi_{\mathcal{M}_\varepsilon}(r) \geq \int r \, d\mu_0 + (\Lambda - \delta)\varepsilon$ for every $\varepsilon \leq C(\delta)$. Combining with the upper bound of (i),

$$\Lambda - \delta \;\leq\; \liminf_{\varepsilon \downarrow 0} \frac{\Phi_{\mathcal{M}_\varepsilon}(r) - \int r \, d\mu_0}{\varepsilon} \;\leq\; \limsup_{\varepsilon \downarrow 0} \frac{\Phi_{\mathcal{M}_\varepsilon}(r) - \int r \, d\mu_0}{\varepsilon} \;\leq\; \Lambda,$$

and letting $\delta \downarrow 0$ gives the directional derivative Λ, which is (8). Note the argument needs only that the transport rate is realised on a set of positive measure, not on one of measure bounded away from zero.

For (iii), if $\varepsilon \geq W_1(\delta_{\pi^\star}, \mu_0)$ then $\delta_{\pi^\star} \in \mathcal{M}_\varepsilon$, so $\Phi_{\mathcal{M}_\varepsilon}(r) \geq r(\pi^\star) = \sup r$; the reverse inequality is immediate. □

Remark 5 (The dual minimiser can be strictly interior). Take $\Omega = [0, 1]$, $\mu_0 = \delta_0$ and $r(\pi) = \pi^2$, so $\mathrm{Lip}(r) = 2$. Since $W_1(\mu, \delta_0) = \mathbb{E}_\mu[\pi]$, the robust value is $\sup\{\mathbb{E}_\mu[\pi^2] : \mathbb{E}_\mu[\pi] \leq \varepsilon\} = \varepsilon$ for $\varepsilon \in [0, 1]$, attained by moving mass ε to $\pi = 1$. The dual minimiser is $\lambda^\star = 1$, strictly between 0 and $\mathrm{Lip}(r)$, and the naïve linear form $\int r \, d\mu_0 + \varepsilon \mathrm{Lip}(r) = 2\varepsilon$ overstates the value by a factor of two. The corrected coefficient is exact here: $\Lambda(r; \delta_0) = \sup_{\pi'} \pi'^2/\pi' = 1$. Note that $\mu_0 = \delta_0$ is itself a degenerate aggregation rule; under the full-support measures Corollary 1 forces we have $\Lambda = \mathrm{Lip}$, and (8) is then the correct first-order expansion, an upper bound with $O(\varepsilon^{3/2})$ gap in our numerical checks, not an identity.

Remark 6 (Where the linear form is used, and what it licenses). Equation (8) underlies the risk–sensitivity pair (9) and the frontier (Proposition 8). Because it is a first-order expansion rather than an identity, those objects are statements about the *first-order robust criterion* $\widehat{\Phi}_\varepsilon := \Phi_0 + \varepsilon\Lambda$, which is also what any implementation computes. Selections of $\widehat{\Phi}_\varepsilon$ agree with exact R-OPERA for ε small enough that no two candidates' robust values cross between the two. The two criteria have *different* large-radius limits: exact R-OPERA saturates at the minimax selection $\arg\min_k \sup r_{s_k}$ once $\varepsilon \geq \max_k \varepsilon_{\mathrm{sat}}(r_{s_k})$, whereas the surrogate is unbounded in ε and drifts to $\arg\min_k \Lambda(r_{s_k}; \mu_0)$, which need not be the minimax candidate and in our data is not (Section D.5). We report the first-order object throughout, do not claim exactness for it, and measure the gap.

Proposition 4 (Bayesian-to-minimax hierarchy)**.** *The map* $\varepsilon \mapsto \Phi_{\mathcal{M}_\varepsilon}$ *is non-decreasing,* concave, *and continuous, with* $\varepsilon = 0$ *giving OPERA (Bayesian) and every* $\varepsilon \geq \max_k \varepsilon_{\mathrm{sat}}(r_{s_k})$ *giving minimax selection. The* value *interpolates continuously; the selected candidate does not, changing*

discontinuously wherever two candidates cross. The hierarchy is therefore attained at a finite radius rather than only in the limit.

Proof. Monotonicity holds because $\mathcal{M}_\varepsilon$ expands with ε. By Theorem 5, $\Phi_{\mathcal{M}_\varepsilon}(r)$ is an infimum over $\lambda \geq 0$ of functions *affine* in ε, hence concave on $[0, \infty)$ and therefore continuous on the open interval $(0, \infty)$. Concavity alone does not give continuity at the endpoint $\varepsilon = 0$; there it follows from monotonicity together with the upper bound of Theorem 5(i), which sandwiches the value between $\int r\, d\mu_0$ and $\int r\, d\mu_0 + \varepsilon\Lambda(r; \mu_0)$. Once $\varepsilon \geq \varepsilon_{\mathrm{sat}}(r)$ the value equals $\sup_\pi r(c, \pi)$, which is the minimax rule. (Earlier drafts asserted convexity; the dual is an infimum, not a supremum, of affine functions of the radius, so the value function is concave.) □

Lemma 5 (Regret bound)**.** *For any deployment condition ω with $W_1(\delta_\omega, \mu_0) \leq \varepsilon$: $r_{s^*_{\mathrm{DR}}}(\omega) - \min_k r_{s_k}(\omega) \leq 2\varepsilon \cdot \max_k \mathrm{Lip}(r_{s_k})$.*

Proof. Let $s^* = \arg\min_k r_{s_k}(\omega)$. Since $\delta_\omega \in \mathcal{M}_\varepsilon$: $\Phi_{\mathcal{M}_\varepsilon}(r_{s^*_{\mathrm{DR}}}) \leq \Phi_{\mathcal{M}_\varepsilon}(r_{s^*}) \leq \int r_{s^*}\, d\mu_0 + \varepsilon\mathrm{Lip}(r_{s^*})$ by Theorem 5(i), and $r_{s^*_{\mathrm{DR}}}(\omega) \leq \Phi_{\mathcal{M}_\varepsilon}(r_{s^*_{\mathrm{DR}}})$. Bounding $\int r_{s^*}\, d\mu_0 \leq r_{s^*}(\omega) + \varepsilon\mathrm{Lip}(r_{s^*})$ (reverse Lipschitz) and combining gives the claim. □

Lemma 6 (Dominance over point masses)**.** *R-OPERA dominates every point-mass criterion* in the worst case over the uncertainty set*: for any π^*,*

$$\sup_{\mu \in \mathcal{M}_\varepsilon} \int r_{s^*_{\mathrm{DR}}}\, d\mu \;\leq\; \sup_{\mu \in \mathcal{M}_\varepsilon} \int r_{s^*(\pi^*)}\, d\mu,$$

where $s^*(\pi^*)$ is the candidate the point mass at π^* selects. Dominance in μ-expected regret at a fixed *μ does not hold and is not claimed: by Proposition 2 the μ-optimal linear criterion is Φ_μ, so whenever a point mass happens to reproduce the Φ_μ selection while R-OPERA hedges away from it, R-OPERA is strictly worse at that μ, the hedging cost that Section 4 measures directly.*

Proof. Immediate from the definition of s^*_{DR} as the minimiser of $\sup_{\mu \in \mathcal{M}_\varepsilon} \int r_{s_k}\, d\mu$ over k: evaluate that inequality at k indexing $s^*(\pi^*)$. □

Lemma 7 (Selection stability)**.** *If the Φ_μ minimizer is unique with margin $\gamma_\mu := \min_{s \neq s_\mu}\{\Phi_\mu(s) - \Phi_\mu(s_\mu)\} > 0$, then for any ν with $W_1(\mu, \nu) < \gamma_\mu/(2L_{\max})$ (where $L_{\max} = \max_s L_s$), the minimizer is unchanged.*

Proof. By Kantorovich–Rubinstein, $|\Phi_\mu(s) - \Phi_\nu(s)| \leq L_{\max} W_1(\mu, \nu)$ for each s. For any competitor $s \neq s_\mu$: $\Phi_\nu(s) - \Phi_\nu(s_\mu) \geq \gamma_\mu - 2L_{\max} W_1(\mu, \nu) > 0$. □

D.2 Deployment completeness beyond the linear class

Corollary 1 characterizes completeness for linear rules only. For the convex rules of Theorem 2(b) we have a sufficient condition, and it is not satisfied by Wasserstein balls.

Proposition 5 (A sufficient condition for convex rules)**.** *Let $\mathcal{M} \subseteq \mathcal{P}(\Omega)$ be weak-$*$ closed and convex, and $\Phi_{\mathcal{M}}(r) = \sup_{\mu \in \mathcal{M}} \int r\, d\mu$. If every $\mu \in \mathcal{M}$ has full support, then $\Phi_{\mathcal{M}}$ satisfies Axiom 5.*

Proof. $\mathcal{P}(\Omega)$ is weak-$*$ compact because Ω is compact metric, so $\mathcal{M}$ is compact, and $\mu \mapsto \int r_1\, d\mu$ is weak-$*$ continuous for continuous r_1. The supremum defining $\Phi_{\mathcal{M}}(r_1)$ is therefore attained at some $\mu_1 \in \mathcal{M}$. Let $r_1 \leq r_2$ with $r_2 > r_1$ on a nonempty open U. By continuity there are $\delta > 0$ and a nonempty open $V \subseteq U$ with $r_2 - r_1 \geq \delta$ on V, and $\mu_1(V) > 0$ by full support. Then $\Phi_{\mathcal{M}}(r_2) \geq \int r_2\, d\mu_1 = \Phi_{\mathcal{M}}(r_1) + \int (r_2 - r_1)\, d\mu_1 \geq \Phi_{\mathcal{M}}(r_1) + \delta\, \mu_1(V) > \Phi_{\mathcal{M}}(r_1)$. □

The condition is sufficient, not necessary, and the Wasserstein balls behind R-OPERA do not meet it. For $\varepsilon > 0$ and an absolutely continuous μ_0, moving the mass of a small ball B to a point outside it costs at most $\mu_0(B)\,\mathrm{diam}(\Omega)$, so $\mathcal{M}_\varepsilon$ contains measures that vanish on B. Completeness can in fact fail. Once $\varepsilon \geq \varepsilon_{\mathrm{sat}}(r_1)$ the robust value equals $\sup r_1$ (Proposition 4). If $r_2 \geq r_1$ is raised only on an open set away from the maximizers of r_1 and stays below $\sup r_1$, then $\sup r_2 = \sup r_1$, so monotonicity and $\Phi_{\mathcal{M}_\varepsilon}(r_2) \leq \sup r_2$ give $\Phi_{\mathcal{M}_\varepsilon}(r_2) = \Phi_{\mathcal{M}_\varepsilon}(r_1)$: R-OPERA in its minimax regime does not register the improvement. Whether it is complete at smaller radii is not settled here. Robustness and completeness are thus separate properties, and we claim only the first for R-OPERA.

D.3 ALTERNATIVE UNCERTAINTY GEOMETRIES

Throughout this appendix $r \in C(\Omega)$ is non-constant, μ_0 is a probability measure, and each ball is a set of probability measures $\mu \ll \mu_0$.

Lemma 8 (χ^2-divergence geometry)**.** *Let $\chi^2(\mu\|\mu_0) = \int (d\mu/d\mu_0 - 1)^2 d\mu_0$ and $\mathcal{M}_\varepsilon^{\chi^2} = \{\mu : \chi^2(\mu\|\mu_0) \le \varepsilon\}$ and write $V = \mathrm{Var}_{\mu_0}(r)$, $m = \int r\, d\mu_0$. Then*

$$\sup_{\mu \in \mathcal{M}_\varepsilon^{\chi^2}} \int r\, d\mu \;\le\; \min\Big\{ m + \sqrt{\varepsilon V},\; \operatorname*{ess\,sup}_{\mu_0} r \Big\},$$

with equality in the first term if and only if the tilted density stays nonnegative, that is

$$\varepsilon \;\le\; \varepsilon^{\chi^2}(r) \;:=\; \frac{V}{\big(m - \operatorname{ess\,inf}_{\mu_0} r\big)^2}.$$

For $\varepsilon > \varepsilon^{\chi^2}(r)$ the supremum is strictly smaller than $m + \sqrt{\varepsilon V}$, and it tends to $\operatorname{ess\,sup} r$ as $\varepsilon \to \infty$.

Proof. Write $d\mu = (1+g)\, d\mu_0$ with $\int g\, d\mu_0 = 0$, $1 + g \ge 0$ μ_0-a.s. and $\chi^2(\mu\|\mu_0) = \int g^2 d\mu_0 \le \varepsilon$. Then $\int r\, d\mu = m + \int g\,(r - m)\, d\mu_0$, and Cauchy–Schwarz gives $\int g(r-m) d\mu_0 \le \sqrt{\varepsilon V}$, with equality for $g^\star = \sqrt{\varepsilon / V}\,(r - m)$. The relaxation that drops $1 + g \ge 0$ therefore has value $m + \sqrt{\varepsilon V}$, which bounds the constrained supremum; the bound by $\operatorname{ess\,sup} r$ is immediate since μ is a probability measure. The maximiser $g^\star$ is feasible exactly when $1 + \sqrt{\varepsilon / V}\,(r - m) \ge 0$ μ_0-a.s., i.e. when $\sqrt{\varepsilon / V}\,(m - \operatorname{ess\,inf} r) \le 1$, which is the stated condition; as the relaxed problem has a unique maximiser, the constrained value is strictly smaller when $g^\star$ is infeasible. Monotone convergence of the feasible set gives the limit $\operatorname{ess\,sup} r$. □

Lemma 9 (KL-divergence geometry)**.** *Let $\mathcal{M}_\varepsilon^{\mathrm{KL}} = \{\mu : \mathrm{KL}(\mu\|\mu_0) \le \varepsilon\}$ with r bounded. Then*

$$\sup_{\mu \in \mathcal{M}_\varepsilon^{\mathrm{KL}}} \int r\, d\mu \;=\; \inf_{\eta > 0} \Big\{ \eta\varepsilon + \eta \log \int e^{r/\eta}\, d\mu_0 \Big\},$$

and as $\varepsilon \to 0$ the value is $\int r\, d\mu_0 + \sqrt{2\varepsilon \mathrm{Var}_{\mu_0}(r)} + O(\varepsilon)$. If $\mathrm{Var}_{\mu_0}(r) = 0$ then r is μ_0-a.s. constant and the supremum is that constant for every ε.

Proof. The Donsker–Varadhan variational formula gives, for every $\eta > 0$, $\int r\, d\mu \le \eta \log \int e^{r/\eta} d\mu_0 + \eta\, \mathrm{KL}(\mu\|\mu_0) \le \eta\varepsilon + \eta \log \int e^{r/\eta} d\mu_0$ on the ball, so the left side is at most the infimum. The reverse inequality is Lagrangian duality for the concave problem $\sup_\mu \{\int r\, d\mu : \mathrm{KL}(\mu\|\mu_0) \le \varepsilon\}$, whose Slater condition holds because μ_0 is feasible with $\mathrm{KL} = 0 < \varepsilon$; the supremum is attained at the exponential tilt $d\mu_\eta \propto e^{r/\eta} d\mu_0$ for the optimal η. For the expansion, $\eta \log \int e^{r/\eta} d\mu_0 = \int r d\mu_0 + \mathrm{Var}_{\mu_0}(r)/(2\eta) + O(\eta^{-2})$, so the objective is $\int r d\mu_0 + \eta\varepsilon + \mathrm{Var}_{\mu_0}(r)/(2\eta) + O(\eta^{-2})$, minimised at $\eta = \sqrt{\mathrm{Var}_{\mu_0}(r)/(2\varepsilon)}$ with value $\int r d\mu_0 + \sqrt{2\varepsilon \mathrm{Var}_{\mu_0}(r)} + O(\varepsilon)$. □

Geometry	Sensitivity penalty	Interpretation
Wasserstein-1	$\varepsilon \cdot \Lambda(r_s; \mu_0)$ (1st order)	Lipschitz constant of risk profile
χ^2	$\sqrt{\varepsilon \cdot \mathrm{Var}[r_s]}$ (small ε)	Std. deviation of risk profile
KL	$\sqrt{2\varepsilon \cdot \mathrm{Var}[r_s]}$ (small ε)	entropic risk; exact dual in Lemma 9
Singleton $\{\mu_0\}$	0	no robustness (OPERA)
$\mathcal{P}(\Omega)$	$\sup r_s - \int r_s\, d\mu_0$	full minimax

D.4 CHOOSING ε

Data-driven: $\hat{\varepsilon} = W_1(\hat{P}_{\text{val}}, \mu_0)$ given deployment-distribution validation data. Conservative default: $\varepsilon = (b - a)/4$, a heuristic rather than a theoretically calibrated choice. Diagnostic: compare selections at $\varepsilon = 0$ and near-minimax $\varepsilon = (b - a)/2$; agreement means the choice is inconsequential, disagreement flags genuine deployment-uncertainty sensitivity.

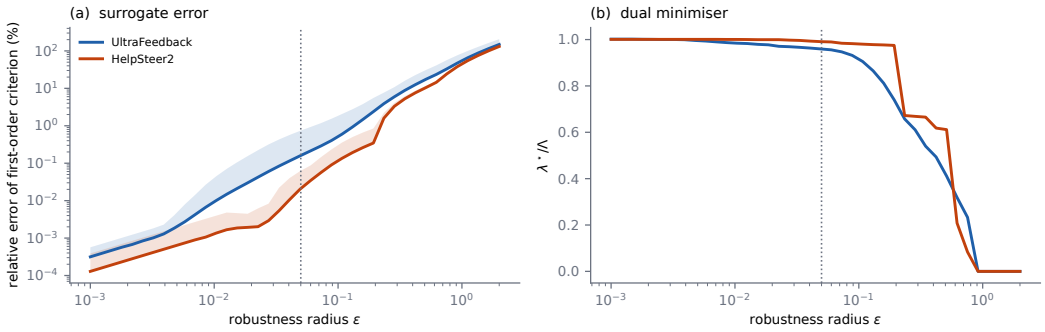


Figure 4: The first-order criterion against the exact Wasserstein dual. **(a)** Relative error of $\widehat{\Phi}_\varepsilon$ (mean over candidates; shaded to the worst case). Below the reported radius (dotted) the error is under 0.2%, and the two criteria select the same candidate in every one of the 495 lattice points, at a numerical tolerance of 10^{-9}. **(b)** The dual minimiser $\lambda^\star$ equals $\Lambda(r;\mu_0)$ throughout the operating range and departs from it only at radii where the uncertainty set has swallowed most of the simplex.

D.5 Exact R-OPERA versus the first-order criterion

Everything the paper plots uses the first-order criterion $\widehat{\Phi}_\varepsilon = \Phi_0 + \varepsilon\Lambda$ of (5). Since that is an upper bound rather than an identity, the honest question is how much it costs. We answer it by computing the exact dual (4) on the RLHF chart.

Because $g(\lambda) := \int [r]^\lambda d\mu_0$ does not depend on ε, the exact value at every radius follows from one pass over a λ-grid per candidate; we use the 495-point mixture lattice with the total-variation metric and μ_0 uniform.

Inside the operating range the two agree completely. For $\varepsilon \le 0.10$, a range that contains the default $\varepsilon = (b-a)/4$ used in every experiment here, exact R-OPERA and $\widehat{\Phi}_\varepsilon$ make the *same selection at every radius on both datasets*, and the surrogate's value error is at most 0.40% on average (1.54% worst-case) on UltraFeedback and 0.09% (0.23%) on HelpSteer2. The dual minimiser sits at $\lambda^\star/\Lambda = 0.984$ at $\varepsilon = 0.01$, confirming first-order exactness numerically (Figure 4).

Outside it they diverge, and in the direction theory predicts. By $\varepsilon = 0.5$ the surrogate overstates the robust value by 17% on average and $\lambda^\star$ has fallen to $0.41\,\Lambda$; by $\varepsilon = 2$ the overstatement is 149% and $\lambda^\star = 0$. Selection agreement over the full sweep, including radii far outside any plausible deployment uncertainty, is 90% (UltraFeedback) and 80% (HelpSteer2). The limits differ as they must: exact R-OPERA saturates at the minimax candidate, whereas the surrogate drifts toward $\arg\min_k \Lambda(r_{s_k};\mu_0)$, here the degenerate no-selection pipeline, which has the flattest profile and the worst risk. Anyone using the frontier at large ε should use the exact dual; at the radii robustness is actually specified at, the linear picture is faithful.

E Symmetry, Universality, and Finite-Sample Theory

E.1 The canonical measure

This appendix states when evaluation is independent of how operating conditions are parameterized, and what that implies for the default measure on prevalence intervals.

E.2 Equivariance and canonical measures

Nothing above should depend on how operating conditions are parameterized, prevalence versus log-odds, task weights versus logits. It does not:

Proposition 6 (Reparameterization equivariance)**.** *Let $\varphi : \Omega \to \Omega'$ be a homeomorphism identifying two parameterizations of the same deployment problem, so that each candidate's profiles*

correspond: $r'_s = r_s \circ \varphi^{-1}$. *Then for every finite Borel measure* μ *on* Ω,

$$\Phi_{\varphi_\# \mu}(r'_s) \;=\; \Phi_\mu(r_s) \qquad \textit{for all candidates } s,$$

where $\varphi_\# \mu$ *is the pushforward. Consequently rankings and selections are invariant, full support is preserved, and the entire classification of Table 3 is parameterization-independent. The same holds for convex aggregation rules with* $\mathcal{M}' = \varphi_\# \mathcal{M}$.

Corollary 2 (Invariant measures are symmetry fixed points)**.** *Let* G *be a group of admissible reparameterizations* $\varphi : \Omega \to \Omega$ *(the problem's symmetry group). If the candidate profiles range over a class that separates measures, for instance all of* $C(\Omega)$*, then* μ *yields* G*-invariant evaluation for every candidate set if and only if* $\varphi_\# \mu = \mu$ *for all* $\varphi \in G$*. (Without a separating class the condition is sufficient but not necessary: a* μ *that is not invariant can still rank a particular finite candidate set the same way.) On the untruncated prevalence scale* $(0, 1)$*, with* G *the group of logit translations acting simply transitively, the unique* σ*-finite invariant Radon measure up to a positive scale factor is its Haar measure, the log-odds-uniform measure; it is infinite, and on a truncated deployment range the group does not act, so what we use is its normalised restriction (Proposition 7, Appendix E.4).*

Proofs are a change of variables (Appendix E.3). The methodological force is weaker than a fixed-point argument and we state it as such: on the untruncated scale the log-odds-uniform measure is the Haar measure of the symmetry the prevalence coordinate already carries, so it is canonical in the sense that an improper invariant prior is canonical. Truncation to a deployment range breaks exact invariance, because the group no longer acts, so on $[a, b]$ the measure is a principled default, not a forced choice, and a practitioner with genuine information about the deployment range should use it (Section 4 does exactly this with persona-centred measures). What the equivariance theorem does deliver unconditionally is that rankings do not depend on how the operating condition is parameterised.

E.3 Equivariance: proofs

Proof of Proposition 6. By the change-of-variables formula for pushforward measures, $\int_{\Omega'} r'_s \, d(\varphi_\# \mu) = \int_\Omega (r'_s \circ \varphi) \, d\mu = \int_\Omega r_s \, d\mu$, using $r'_s \circ \varphi = r_s$. Rankings coincide because every candidate's score is preserved. Full support is preserved because φ is a homeomorphism: $\operatorname{supp}(\varphi_\# \mu) = \varphi(\operatorname{supp} \mu)$. For convex aggregation rules, $\sup_{\mu \in \mathcal{M}} \int r_s \, d\mu = \sup_{\nu \in \varphi_\# \mathcal{M}} \int r'_s \, d\nu$ term by term. □

Proof of Corollary 2. ($\Leftarrow$) If $\varphi_\# \mu = \mu$ for all $\varphi \in G$, then by Proposition 6 evaluation under μ is unchanged by every $\varphi \in G$. ($\Rightarrow$) If $\varphi_\# \mu \neq \mu$ for some φ, the two measures differ on some continuous test function $r \geq 0$ (Riesz uniqueness, Theorem 2), so $\int r \, d\mu \neq \int r \, d(\varphi_\# \mu)$; assume without loss of generality that the first is smaller and pick a constant c with $\int r \, d\mu < c < \int r \, d(\varphi_\# \mu)$. A constant integrates to itself under every probability measure, so the two-candidate set $\{r, c\}$ is ranked $r \prec c$ under μ and $c \prec r$ under $\varphi_\# \mu$: a *strict* reversal, not a tie. (Taking the constant $\int r \, d\mu$ instead would only produce a tie under μ and would not establish the claim.) This makes the selection differ across the two parameterizations. The prevalence-interval case is Proposition 7. □

E.4 Log-odds uniform measure: invariance

Definition 2 (Logit-scale translation)**.** For $t \in \mathbb{R}$, define $T_t : (0, 1) \to (0, 1)$ by $T_t(\pi) = \sigma(\sigma^{-1}(\pi) + t)$ where $\sigma(x) = (1 + e^{-x})^{-1}$.

Proposition 7 (Uniqueness)**.** *On the* untruncated *scale* $\Omega = (0, 1)$*, the logit-translation group* $\{T_t\}_{t \in \mathbb{R}}$ *acts freely and transitively, and* $d\mu_{\mathrm{LOU}}(\pi) \propto d\pi / [\pi(1 - \pi)]$ *is the unique translation-invariant (Haar) measure up to scale. It is* σ*-finite and infinite, so it admits no normalisation: there is no invariant* probability *measure for this action. On a truncated range* $[a, b]$ *the group does not act (*$T_t[a, b] \not\subseteq [a, b]$ *for* $t \neq 0$*), so no measure on* $[a, b]$ *is a fixed point of it; the normalised restriction* $\frac{1}{Z} \frac{d\pi}{\pi(1-\pi)}$, $Z = \log \frac{b(1-a)}{a(1-b)}$*, is the restriction of the Haar measure, which we adopt as a default rather than derive as an invariance fixed point; through* Z*, it depends on the endpoints* a *and* b*.*

Proof. Under $\eta = \sigma^{-1}(\pi)$ the action becomes translation on the whole line $\mathbb{R}$, whose unique invariant measure up to scale is Lebesgue (Haar for $(\mathbb{R}, +)$); pulling back via $d\eta = d\pi / [\pi(1 - \pi)]$

gives the stated density on $(0, 1)$. Lebesgue measure on $\mathbb{R}$ is infinite, hence not normalisable, and translations do not map a bounded sub-interval to itself; both facts are why the truncated version is a restriction rather than a fixed point. □

E.5 Universality of the construction

Any decision system whose realized loss depends on a deployment-time parameter induces a profile; no theorem is needed, only the construction (2) of Section 2. A reinforcement-learning policy's return over a set of environment parameters (transition or reward drift) is a profile, with deployment-time adaptation given by re-tuning or fine-tuning. An auction mechanism's revenue over a family of bidder-type distributions is a profile, with the reserve price as the adapted configuration. A controller's cost over a class of disturbance spectra is a profile, with gain re-tuning as adaptation. A newsvendor's cost over (demand level, critical fractile) is a profile, with the order quantity adapted. We develop the clinical, task-mixture, and moderation instantiations fully (Section 4) and leave these as constructions; each inherits every result of the paper verbatim.

E.6 Finite-sample implementation and theory

Given K-fold cross-validation, a grid $\{\omega_1, \dots, \omega_G\} \subset \Omega$, and radius ε: $\hat{\Phi}_{\mathrm{DR}}(s_k) = \sum_{j=1}^{G} w_j \hat{r}^{\mathrm{CV}}_{s_k}(\omega_j) + \varepsilon \cdot \widehat{\mathrm{Lip}}(r_{s_k})$, with w_j quadrature weights for the chosen measure and $\widehat{\mathrm{Lip}}$ the maximal empirical difference quotient. Under label shift, class-conditionals are stable, so FPR/FNR are estimated once and recombined with deployment prevalence as a known coefficient, no importance weighting. Overhead beyond standard CV: $O(nG)$ multiplications plus $O(G^2)$ comparisons.

Lemma 10 (Uniform concentration of the empirical ROC)**.** *With probability $\geq 1 - \delta'$,* simultaneously for all thresholds $\tau \in \mathbb{R}$, $|\widehat{\mathrm{FPR}}(\tau) - \mathrm{FPR}(\tau)| \leq \sqrt{\log(2/\delta')/(2n_0)}$*; the same holds for FNR with n_1.*

Proof. The bound is stated for an independent hold-out: DKW requires i.i.d. samples, while out-of-fold scores from K-fold cross-validation are dependent because the folds share training data, so applying it there is a standard approximation rather than a theorem. With that said, $\mathrm{FPR}(\tau) = 1 - F_0(\tau)$ for the score c.d.f. F_0 of the negative class, and $\widehat{\mathrm{FPR}}$ is the corresponding empirical c.d.f., so the claim is the Dvoretzky–Kiefer–Wolfowitz inequality with Massart's constant: $P(\sup_\tau |\hat{F}_0 - F_0| > t) \leq 2e^{-2n_0 t^2}$. Uniformity in τ is what the next lemma needs, because the thresholds it ranges over are chosen from the same sample. □

Lemma 11 (Concentration of envelope risk)**.** *Fix s. With probability $\geq 1 - \delta'$,* simultaneously for every $\pi \in [0, 1]$,

$$|\hat{r}_s(\pi) - r_s(\pi)| \;\leq\; c(1-\pi)\sqrt{\tfrac{\log(4/\delta')}{2n_0}} + (1-c)\pi\sqrt{\tfrac{\log(4/\delta')}{2n_1}} \;\leq\; \sqrt{\tfrac{\log(4/\delta')}{2\min(n_0,n_1)}}.$$

Proof. Apply Lemma 10 to each class at level $\delta'/2$ and intersect the two events. On that event, for every π and every τ simultaneously, the pointwise objective $f_\pi(\tau) = c(1-\pi)\mathrm{FPR}(\tau) + (1-c)\pi\mathrm{FNR}(\tau)$ satisfies $|\hat{f}_\pi(\tau) - f_\pi(\tau)| \leq c(1-\pi)t_0 + (1-c)\pi t_1 =: \Delta(\pi)$ with t_i the two DKW radii. Since $|\inf_\tau \hat{f}_\pi - \inf_\tau f_\pi| \leq \sup_\tau |\hat{f}_\pi - f_\pi|$, and $r_s(\pi) = \inf_\tau f_\pi(\tau)$, $\hat{r}_s(\pi) = \inf_\tau \hat{f}_\pi(\tau)$, the claim follows. ($\hat{f}_\pi$ is piecewise constant with jumps only at observed scores, so its infimum over $\mathbb{R}$ is attained on the sample and the empirical minimisation is the one the estimator performs.) The final inequality uses $c(1-\pi) + (1-c)\pi \leq 1$.

The uniformity matters: the earlier route, a union bound over the $\leq n$ empirical thresholds, is not valid, because those thresholds are functions of the same sample and the corresponding events are not fixed in advance. DKW removes the difficulty and, as a by-product, the $\log n$ and the grid size G drop out of the rate. □

Theorem 6 (Finite-sample selection guarantee)**.** *Let Φ_G be a grid aggregation rule with nonnegative weights summing to one and* no *sensitivity penalty, and suppose its population minimiser k^* has margin $\gamma > 0$. Then with probability $\geq 1 - \delta$ empirical selection returns k^* provided*

$$\min(n_0, n_1) \;\geq\; 8\log(4K/\delta)/\gamma^2.$$

Proof. Union bound over the K systems with Lemma 11 at level δ/K gives, with probability $\geq 1-\delta$,

$$\sup_{\pi} \max_{k} |\hat{r}_{s_k}(\pi) - r_{s_k}(\pi)| \;\leq\; \bar{\sigma} := \sqrt{\log(4K/\delta)/(2\min(n_0, n_1))}.$$

Because the weights are nonnegative and sum to one, the same bound transfers to Φ_G: $\max_k |\hat{\Phi}_G(s_k) - \Phi_G(s_k)| \leq \bar{\sigma}$. Hence $\hat{\Phi}_G(s_k) - \hat{\Phi}_G(s_{k^*}) \geq \gamma - 2\bar{\sigma}$ for every $k \neq k^*$, which is strictly positive once $\bar{\sigma} < \gamma/2$; the stated sample size gives $\bar{\sigma} \leq \gamma/4$ and margin $\gamma/2$. The statement excludes the sensitivity penalty on purpose: the penalty is not an average of profile values, and its estimation is handled separately in Corollary 3. □

Remark 7 (No importance weights). Sample complexity depends on $\min(n_0, n_1)$, not on the ratio of deployment to evaluation prevalence. Importance weighting would require $n \cdot (\pi_{\text{deploy}}/\pi_{\text{eval}})^2$ effective samples.

Corollary 3 (Oracle inequality)**.** *Let $\hat{k}$ minimise the empirical first-order criterion*

$$\widehat{\Phi}^G_\varepsilon(k) \;=\; \sum_j w_j\, \hat{r}_{s_k}(\pi_j) \;+\; \varepsilon \widehat{\Lambda}_k, \qquad \widehat{\Lambda}_k \;=\; \max_{i \neq j} \frac{|\hat{r}_{s_k}(\pi_i) - \hat{r}_{s_k}(\pi_j)|}{|\pi_i - \pi_j|},$$

on the grid $\{\pi_1, \ldots, \pi_G\} \subset [a, b]$ with nonnegative weights summing to one that represent μ_0, meaning $w_j = \mu_0(V_j)$ for the Voronoi cells V_j of the grid. Some such condition is necessary: for arbitrary weights the discrete measure need not approximate μ_0 at all, and the quadrature term below fails. Put $L_{\max} := \max_k \mathrm{Lip}(r_{s_k})$, $\Delta_{\min} := \min_{i \neq j} |\pi_i - \pi_j|$,

$$\delta_{\text{grid}} := L_{\max} \max_{\pi \in [a,b]} \min_j |\pi - \pi_j|, \qquad \delta_{\text{slope}} := \max_k \left|\Lambda(r_{s_k}; \mu_0) - \Lambda^G_k\right|,$$

where Λ^G_k is the population difference quotient maximised over the grid. The bound is one-dimensional: a multivariate deployment space needs a mesh geometry and a multivariate slope estimator, which we do not develop. Then for any *deployment prevalence $\pi_d \in [a, b]$, with probability $\geq 1 - \delta$,*

$$r_{s_{\hat{k}}}(\pi_d) - \min_k r_{s_k}(\pi_d) \;\leq\; 2L_{\max}\, W_1(\delta_{\pi_d}, \mu_0) \;+\; \varepsilon L_{\max} \;+\; 2\bar{\sigma}\Big(1 + \frac{2\varepsilon}{\Delta_{\min}}\Big) \;+\; 2\delta_{\text{grid}} + 2\varepsilon\, \delta_{\text{slope}},$$

with $\bar{\sigma}$ as in Theorem 6. The first term is the residual distance from the deployment condition to the criterion measure; it is not absorbed into ε.

Remark 8 (Reading the two new terms). The term $4\varepsilon\bar{\sigma}/\Delta_{\min}$ is the price of estimating a slope from noisy profile values: a difference quotient divides estimation error by the spacing. It pulls against δ_{grid}, which wants a fine grid. With uniform spacing h the two contribute $\approx L_{\max} h + 4\varepsilon\bar{\sigma}/h$, minimised at $h \asymp \sqrt{\varepsilon\bar{\sigma}/L_{\max}}$ for a total of order $\sqrt{\varepsilon\bar{\sigma}L_{\max}}$. When the profiles are envelopes of finitely many affine pieces, as in Section 2, a better route is available: in one dimension $\mathrm{Lip}(r_s)$ is the largest slope among the active configurations, and in general the largest dual norm of their gradients, computable in closed form rather than estimated. Two conditions are needed before that removes the slope terms. The implemented criterion must itself use $\mathrm{Lip}(r_s)$, and μ_0 must have full support on $[a, b]$, since $\Lambda(r; \mu_0) \leq \mathrm{Lip}(r)$ with equality exactly in that case. Without full support, substituting Lip for Λ does not estimate the same criterion more accurately, it *changes* the criterion, and the comparison is no longer governed by this bound. Note also that δ_{slope} is defined as an absolute difference precisely because Λ^G_k can exceed $\Lambda(r_{s_k}; \mu_0)$ when the grid visits points outside $\operatorname{supp} \mu_0$; assuming full support removes that possibility and makes the grid quotient a lower bound on Λ.

Proof. Write $\Phi_\varepsilon(k) = \int r_{s_k} d\mu_0 + \varepsilon\Lambda(r_{s_k}; \mu_0)$ for the population criterion and $\Phi^G_\varepsilon(k) = \sum_j w_j r_{s_k}(\pi_j) + \varepsilon\Lambda^G_k$ for its grid version. Four approximations are chained.

(i) Quadrature. Each r_{s_k} is $L_{\max}$-Lipschitz, so its grid quadrature moves the integral by at most δ_{grid}, and the grid difference quotient moves the sensitivity by at most δ_{slope} by definition. Hence $|\Phi^G_\varepsilon(k) - \Phi_\varepsilon(k)| \leq \delta_{\text{grid}} + \varepsilon\delta_{\text{slope}}$ for every k.

(ii) Estimation of the average. By Theorem 6, $\sup_\pi \max_k |\hat{r}_{s_k}(\pi) - r_{s_k}(\pi)| \leq \bar{\sigma}$ with probability $\geq 1 - \delta$; on that event the weighted average moves by at most $\bar{\sigma}$.

Table 3: Evaluation procedures as aggregation rules on the risk profile. Shaded: the only non-degenerate rule. Current practice sits entirely in the first block, and differs only in *which* condition it evaluates at.

Procedure	**Geometry**	**Degeneracy**
CV at development conditions	$\mu = \delta_{(c_0,\pi_0)}$	point mass
DCA at threshold t (Vickers and Elkin, 2006)	$\mu = \delta_{(t,\pi_0)}$	point-mass family
Leaderboard task-average	$\mu = \delta_{w_{\text{unif}}}$	point mass on the simplex
AUCH (area under the hull)	candidate-dependent μ_s	improper (violates fixed-μ)
AUC-ROC	n/a	not a functional of r_s
Minimax	$\mathcal{M} = \mathcal{P}(\Omega)$	maximal geometry
CVaR_α	distortion geometry	tail reweighting
OPERA / R-OPERA	**full-support μ, or a W_1 ball**	**non-degenerate**

(iii) Estimation of the slope. On the same event, for any grid pair $i \neq j$ the numerator of the difference quotient moves by at most $2\bar{\sigma}$ and the denominator is at least $\Delta_{\min}$, so each quotient moves by at most $2\bar{\sigma}/\Delta_{\min}$; a maximum of uniformly perturbed quantities is perturbed by at most the same amount, giving $|\widehat{\Lambda}_k - \Lambda_k^G| \leq 2\bar{\sigma}/\Delta_{\min}$. Combining with (ii), $|\widehat{\Phi}_\varepsilon^G(k) - \Phi_\varepsilon^G(k)| \leq \bar{\sigma}(1 + 2\varepsilon/\Delta_{\min})$. This is the step that a bound of $\bar{\sigma}(1+\varepsilon)$ would get wrong: $\widehat{\Lambda}_k$ is not an average of profile values.

(iv) Transport to the deployment condition. Kantorovich–Rubinstein gives $|r_{s_k}(\pi_d) - \int r_{s_k} d\mu_0| \leq \text{Lip}(r_{s_k}) W_1(\delta_{\pi_d}, \mu_0) \leq L_{\max} W$, writing $W := W_1(\delta_{\pi_d}, \mu_0)$. This is where the deployment condition enters, and it is *not* controlled by ε: no step of the argument requires $\delta_{\pi_d} \in \mathcal{M}_\varepsilon$.

Now let $k^\dagger := \arg\min_k r_{s_k}(\pi_d)$. Since $\hat{k}$ minimises $\widehat{\Phi}_\varepsilon^G$, steps (i)–(iii) give $\Phi_\varepsilon(\hat{k}) \leq \Phi_\varepsilon(k^\dagger) + 2\bar{\sigma}(1 + 2\varepsilon/\Delta_{\min}) + 2\delta_{\text{grid}} + 2\varepsilon\delta_{\text{slope}}$. Because $0 \leq \Lambda(r_{s_k}; \mu_0) \leq L_{\max}$, dropping the nonnegative penalty on the left and bounding it by $\varepsilon L_{\max}$ on the right turns this into a statement about $\int r_{s_k} d\mu_0$; applying (iv) once on each side to convert both integrals to values at π_d adds $2L_{\max} W$ and yields the claim. □

Remark 9 (The criterion's own blind set, and its error bar). Any implementation evaluates the profile on a finite grid and is therefore *local* in the sense of Axiom 4; by Theorem 3 it too carries blind sets. This is not a contradiction but a quantitative statement: δ_{grid} above is exactly the size of that blind set, it is reportable, and it shrinks as the grid is refined. The distinction we draw is not finite versus infinite computation (all computation is finite) but between a criterion that *targets* a full-support functional and reports its discretisation error, and one that targets a point mass and reports nothing. Section G.1 measures both sides of this comparison on real systems.

Remark 10 (Sensitivity penalty as regularizer). Among candidates with equal integrated risk, R-OPERA prefers the smaller Lipschitz constant: lower worst-case risk (robustness) and lower estimation variance (regularization). The pair's second coordinate plays both roles.

F Relocated Statements, Scope, and Practitioner Guidance

F.1 Statements relocated from the main text

This appendix collects statements referenced from the body.

Remark 11 (Constrained deployers). Axiom 2 assumes the deployer optimizes τ freely. A deployer constrained to a class $\tilde{\mathcal{T}} \subseteq \mathcal{T}$ (regulatory thresholds, latency budgets) induces the constrained envelope $\inf_{\tau \in \tilde{\mathcal{T}}} R_\omega(s, \tau)$; every result applies verbatim with that profile. The extreme case $|\tilde{\mathcal{T}}| = 1$ (no adaptation) yields affine profiles and lands in the collapse regime of Proposition 1, so the rational-deployer assumption is not a hidden strength: weakening it moves the theory continuously toward the regime where it is not needed at all.

Remark 12 (Adaptation creates curvature). The envelope (2) is an infimum of functions affine in ω at fixed τ, hence *concave* whenever L is affine in ω, as on the mixture simplex, or on a prevalence interval at fixed cost ratio. On the two-dimensional clinical domain $\omega = (c, \pi)$ the loss is *bilinear* rather than affine and the envelope is concave in each coordinate separately but not jointly; every

statement below that uses concavity is applied on a fixed-c slice or on the simplex. Deployment-time adaptation is the mechanism that can bend risk profiles away from affine; distinct geometries select differently exactly where the resulting *difference* profiles are non-affine, which adaptation permits without guaranteeing. Empirically (Section 4.1), a leaderboard's task-average and any Dirichlet-integrated criterion pick identical models, until the deployment pipeline is allowed to adapt (threshold or routing choice), at which point they diverge. The theory's domain of relevance is adaptive deployment, which is the modern norm.

Corollary 4 (No fixed benchmark suite is deployment-complete)**.** *Let Ω be the simplex of deployment task mixtures and let an evaluation report any finite number of aggregate scores, each a fixed weighted combination of per-task results. Then there is a nonempty open set of deployment mixtures U and, for every sufficiently small $h > 0$, a pair of models whose deployment risks differ by at least h throughout an open subset of U while every reported score is identical. Adding benchmarks shrinks the blind set but cannot eliminate it; within the linear class, only full-support integration over mixtures can (Corollary 1); nonlinear rules require the separate analysis of Appendix D.2.*

Proof. Each aggregate score is the profile evaluated at one mixture, so the suite is local with Ω_0 the finite set of reported mixtures. Apply Lemma 2 at a mixture ω_c at maximal distance h from Ω_0: it produces two envelopes of $2d$ and $2d+1$ affine pieces that agree on Ω_0 and differ by $Lh/\sqrt{d}$ on an open neighbourhood of ω_c, with L the Lipschitz constant of Lemma 2. Both are profiles of systems with finitely many configurations. Counting configurations is not by itself enough: the pipeline must be able to realize $2d+1$ *independently specifiable* affine task-risk vectors, so that the gradients $\{\pm e_j\}$ of Lemma 2 are available. On the task-mixture simplex that holds whenever per-task risks can be traded against one another, as routing, thresholding or per-deployment tuning allow; a pipeline with many configurations whose risk vectors span a lower-dimensional set does not admit the construction. For a non-reconfigurable pipeline the profile is affine (Proposition 1) and no such pair exists, which is exactly the boundary of Remark 13. □

Remark 13 (Consistency with the collapse). For *non-reconfigurable* systems, profiles on the mixture simplex are affine (Proposition 1), and $d+1$ affinely independent mixtures determine an affine profile exactly, so a finite suite *is* complete for that class. The blind-spot phenomenon and the collapse are two sides of one boundary: a finite suite is a sufficient statistic whenever profiles are affine, and can fail as soon as reconfiguration makes them strictly curved. The converse does not hold configuration by configuration: a reconfigurable system whose profile happens to be affine, for instance one configuration that is optimal throughout Ω, is also seen exactly. Reconfiguration is what makes curvature *possible*, not what makes it inevitable.

Remark 14 (Reading the theorem). The proof is elementary; the force of the result is in its packaging of a trade-off that practice resolves silently and badly. Every point-mass rule in Table 3 chooses locality and gives up completeness, usually without anyone deciding to. The theorem says this is not fixable by choosing the point well, adding points, or adding data: *any* finite set of operating conditions leaves open regions of the deployment space where regret accumulates invisibly. The only exits are full-support integration (OPERA-type rules) or leaving the linear class for a convex geometry (Theorem 2(b)), which hedges blind spots at the price of conservatism.

Proposition 8 (The deployment frontier)**.** *Plot each candidate at $\Gamma(s_k) \in \mathbb{R}^2$ and let $\widehat{\Phi}_\varepsilon(s) := \Phi_0(s) + \varepsilon\,\Lambda(r_s;\mu_0)$ be the first-order robust criterion of Theorem 5(ii), the object an implementation computes. Write H for the convex hull of $\{\Gamma(s_1),\dots,\Gamma(s_K)\}$. Then:*

- *(i) for each $\varepsilon \geq 0$ the set of $\widehat{\Phi}_\varepsilon$-minimisers is exactly the set of candidates whose signature lies on the face of H exposed by $-(1,\varepsilon)$, the sign convention for minimising $\langle(1,\varepsilon),\cdot\rangle$; this face is a single vertex unless two or more signatures lie on an edge whose minimising normal (the direction in which the linear functional is minimised over H) is proportional to $(1,\varepsilon)$, in which case every candidate on that edge ties;*
- *(ii) a candidate is the* unique *selection at some $\varepsilon \geq 0$ if and only if its signature is a vertex of H whose normal cone contains $(1,\varepsilon)$ in its interior; as ε increases these vertices are traversed in order of decreasing sensitivity and increasing Φ_0;*
- *(iii) a candidate whose signature is not on the lower-left boundary of H, in particular any candidate strictly inside H, is never selected at any robustness level.*

Only directions with a strictly positive first coordinate arise, so a vertex exposed solely by $(0, 1)$*, minimum sensitivity but not minimum* Φ_0 *is reached only in the limit* $\varepsilon \to \infty$ *of the first-order criterion. Exact R-OPERA instead reaches minimax at the finite radius* $\max_k \varepsilon_{\mathrm{sat}}(r_{s_k})$.

Proof. $\widehat{\Phi}_\varepsilon(s_k) = \langle (1, \varepsilon), \Gamma(s_k) \rangle$ is a linear functional of the signature, so minimising it over the finite set is the same as minimising it over H, and the minimising set of a linear functional over a polytope is the exposed face in that direction; a candidate attains the minimum exactly when its signature lies on that face, which gives (i). Since $(1, \varepsilon)$ has a nonzero first coordinate, the exposed face has dimension at most one, so it is a vertex or a single edge, and it is an edge precisely when that edge's minimising normal is parallel to $(1, \varepsilon)$ – the tie case. For (ii), a vertex is the unique minimiser for a direction exactly when that direction lies in the interior of its normal cone, and $\varepsilon \mapsto (1, \varepsilon)$ sweeps the directions with positive first coordinate monotonically; monotone traversal follows because the map from ε to the exposed vertex is the subdifferential of the convex support function $\varepsilon \mapsto \min_k \widehat{\Phi}_\varepsilon(s_k)$, whose slope is the selected sensitivity and is nonincreasing. For (iii), a point not on the exposed face for any admissible direction is never a minimiser, and a point in the interior of H is exposed by no direction at all. □

F.2 Scope and standing of the claims

This appendix collects what the main text does not claim: why AUC and the information criteria fall outside the classification, the remaining proofs of Section 3, and the standing of the leaderboard experiment.

F.3 AUC and the information criteria

AUC is not a functional of r_s *at all.* The envelope (2) is the lower envelope over configurations, which for a thresholded scorer is the ROC *convex hull*; two scorers with the same hull can differ on dominated portions of their ROC curves and therefore differ in AUC while sharing a profile. What *is* an aggregation rule on r_s is the area under the hull (AUCH), improper in Hand's sense because its implicit measure depends on the candidate's own score distribution, so different candidates are scored by different geometries (Hand, 2009; Flores et al., 2025). *Likelihood-based information criteria are also outside.* AIC and BIC estimate predictive divergence for a specified likelihood model, with a complexity penalty that is not a functional of the achievable-risk profile; identifying AIC with a point evaluation at symmetric loss would require the further assumptions that the induced decision rule is scored by 0–1 loss, at equal costs, and at the development prevalence. We do not make them.

F.4 Finite sufficiency: proof

Proposition 9 (Finite sufficiency on a linear profile class)**.** *Suppose every candidate profile is known to lie in a fixed linear subspace* $V \subseteq C(\Omega)$ *of finite dimension* d*. Then there are conditions* $\omega_1, \ldots, \omega_d$ *at which evaluation determines* r_s *exactly, so a criterion consulting those* d *conditions is deployment-complete on* V*. An affine family over a* d*-dimensional condition space has dimension* $d + 1$ *in general; the count here refers to the linear part fixed by* V*.*

The impossibility of Theorem 3 is therefore about not knowing the shape of profiles in advance, and Proposition 1 is the affine instance.

Proof of Proposition 9. Point evaluations separate $C(\Omega)$, so on the finite-dimensional V one may choose $\omega_1, \ldots, \omega_d$ making $r \mapsto (r(\omega_i))_{i \le d}$ injective; r_s is then recoverable from those values, and any functional of r_s is computable from them. □

F.5 Proof of the fatness threshold

Throughout this section balls and distances are relative to the affine hull of Ω, the right convention when Ω is a simplex and so has empty interior in the ambient space, and Ω is assumed to have nonempty relative interior so that balls of small radius exist.

Proof of Theorem 4. Let a be a configuration of r_1 optimal on a region R (convex, as an intersection of half-spaces with Ω) containing a ball $B(z,\rho)$ with $\rho > 3h$, and let $N = \Omega_0 \cap B(z,\rho)$. *Step 1 ($N$ spans).* Every $y \in B(z,\rho-h)$ has a point of Ω_0 within h, which lies in $B(z,\rho)$; so N is an h-net of $B(z,\rho-h)$. If some x with $\|x-z\| < \rho-2h$ lay outside $\mathrm{conv}(N)$, a unit u would separate it, $\langle u,n\rangle < \langle u,x\rangle$ for all $n \in N$; but $y = x + hu \in B(z,\rho-h)$ has some $n \in N$ with $\|n-y\| \le h$, giving $\langle u,n\rangle \ge \langle u,x\rangle$. Hence $\mathrm{conv}(N) \supseteq B^\circ(z,\rho-2h)$, and the point $n^\star \in N$ nearest z, within h of it, is interior to $\mathrm{conv}(N)$ because $h < \rho-2h$. *Step 2 ($r_2 = a$ on $\mathrm{conv}(N)$).* $\mathrm{conv}(N) \subseteq R$, so $r_1 = a$ there, and $r_2 = r_1 = a$ on N. As r_2 is concave it lies above the affine interpolant of its values on N, so $\varphi := r_2 - a \ge 0$ on $\mathrm{conv}(N)$, with $\varphi(n^\star) = 0$. For any x in $\mathrm{conv}(N)$, extend the segment from x through $n^\star$ to a point x' still in $\mathrm{conv}(N)$ (possible as $n^\star$ is interior); concavity gives $0 = \varphi(n^\star) \ge \lambda\varphi(x) + (1-\lambda)\varphi(x') \ge 0$ with $\lambda > 0$, so $\varphi(x) = 0$. *Step 3 ($r_2 \le a$ everywhere).* $r_2 = a$ on the open set $B^\circ(z,\rho-2h)$, so at any of its points r_2 is differentiable with gradient ∇a, and concavity gives $r_2(\omega) \le a(\omega)$ for all ω. Applying Step 3 to every configuration of r_1 optimal on a ball of radius $> 3h$ gives $r_2 \le \min$ of those configurations, which equals r_1 off the union of the remaining regions; exchanging roles gives the same for r_1. If both envelopes are w-fat with $w > 3h$, every configuration qualifies, so $r_2 \le r_1$ and $r_1 \le r_2$ on Ω. □

The fatness values quoted in Section 3 are exact: for each configuration a of a pipeline, its optimality region is the polytope $\{\omega \in \Delta : \omega\cdot(m_b - m_a) \le \lambda(c_b - c_a)\ \forall b\}$, and its inradius within the simplex is a linear program (a Chebyshev ball, with constraint normals projected onto the simplex's tangent space). The script is released with the paper.

F.6 Standing of the leaderboard experiment

What is measured there is regret against *hypothesised* deployment mixtures over the leaderboards' own published per-condition scores. It is not observed regret in a real deployment, and it inherits four assumptions we cannot test: that mixtures over published categories model deployment, that those categories span deployment utility, that per-category scores are commensurate and linear in that utility, and that latency, cost, prompt distribution and deployment-time adaptation can be set aside. The persona sweep addresses the first and none of the other three.

F.7 Practitioner guidance and limitations

Practitioner guidance. If deployment matches evaluation and systems cannot adapt, point evaluation is optimal (Proposition 1) and this paper is not needed. If the deployment distribution over conditions can be elicited, from logs or from the people deploying the system, integrate against it. If it cannot, and systems adapt, use a full-support default such as OPERA, report how selections move with the interval's endpoints, read robustness off the frontier, and let ε express how far deployment may drift; the $\varepsilon = 0$ versus near-minimax agreement diagnostic (Appendix D.4) says when that choice is inconsequential.

Scope of the claims. The object and the integral criterion are inherited from the cost-curve literature (Appendix B). What is new is the collapse boundary, the account of what finite suites miss on realizable systems, exhaustiveness over an abstract Ω, and the transport to multi-dimensional condition and configuration spaces. The convex-geometry duality is a transplant of known risk-measure theory.

Evidence. The leaderboard deployments of Section 4.1 are hypothesized mixtures, and regret is measured against the mixture-optimal candidate rather than observed in deployment. On logged Arena traffic (Appendix G.6) broad language markets sit near the pooled mixture and rankings rarely change, so the leaderboard effect applies to concentrated deployments; the observed segments are language markets, not individual products, because sessions are too short to recover individual workloads. The adaptive evidence comes from offline pipelines, its gain vanishes when preferences are estimated from 100 labels or fewer, and the one adaptive system we could build from logged traffic shows no curvature where its deployments sit. No production routing system with logged choices was available to us. The clinical replication is mixed and, before we pinned the software environment, was not reproducible at the level of individual care units.

The default measure. On a truncated deployment range the logit-translation group does not act, so the log-odds-uniform measure is a default rather than an invariance fixed point. Its endpoints change OPERA's selection in up to 22% of replications on MIMIC-IV without changing which care units it wins (Appendix I.5), which is a reason to prefer an elicited deployment distribution whenever one exists.

Robustness and completeness. The robust value is only *first-order* linear in the radius: the dual minimiser can lie strictly inside $(0, \Lambda)$ (Remark 5), so the risk–sensitivity pair and the frontier characterise $\widehat{\Phi}_\varepsilon$, the object implementations compute, rather than exact R-OPERA. Completeness is characterized only for linear rules; for convex rules we give a sufficient condition that Wasserstein balls do not meet, and R-OPERA in its minimax regime is not complete (Appendix D.2). Deployment completeness is stated for open sets of conditions; measure-zero deployment sets remain invisible to any integral criterion.

Two uncertainties. Profiles are estimated, and the deployment measure is uncertain. We treat these separately: the finite-sample guarantee (Theorem 6) controls selection error from estimating the profiles, and the radius ε is meant to express only uncertainty about the measure. Inflating ε to absorb estimation error as well would count robustness twice, since the sensitivity term then penalizes deviations that the finite-sample margin already covers. A principled combination, for instance adding the finite-sample deviation to the robust value through a union bound, is not developed here.

Rates and implementation. Proposition 3(iii) is a worst-case rate that real systems do not attain, since measured decay is about two and a half times steeper (Section G.1), so we do not claim the exponential reading. Any implementation is itself local and carries the blind set quantified by δ_{grid} (Remark 9); the distinction we draw is between targeting a full-support functional with a reportable discretisation error and targeting a point mass with none. The representations characterize monotone aggregation, so non-monotone desiderata, such as preferring variability for exploration, are outside the theory, and the impossibility theorem's axioms are sufficient for its conclusion without being minimal.

G Leaderboard Experiments: Blind Sets, Personas, and Logged Traffic

G.1 An exhibited blind set, and how fast it shrinks

Theorem 3 asserts that a suite reporting finitely many conditions cannot see everything. Here we exhibit the blind set on real systems rather than construct it.

A benchmark suite publishes per-task scores (the risk at each vertex of the mixture simplex) plus a headline aggregate, the equal-weight average: $G = D + 1 = 6$ operating conditions, the shape of a six-task leaderboard. Any nonempty subset of a configuration pool is itself a realizable pipeline, so we search over subsets for a pair the suite cannot distinguish. Let B be the pipeline optimal at every reported condition (the system the suite would crown) and $A = B \cup \{\tau\}$ for a configuration τ that is never better at a reported condition. By Lemma 2 the region where τ is actually used is convex, and when that cell avoids Ω_0 the two pipelines report *identical* numbers.

Such pairs are abundant: 778 of the 1,729 available configurations are invisible to the suite. The largest invisible deployment gap is 28.2×10^{-3} of risk, at a mixture weighting truthfulness and conciseness, contributed by a best-of-n budget of the scalar-plus-brevity pipeline that is 37.5×10^{-3} *worse* than B at every reported condition and only earns its keep in the interior. For calibration, the headline scores of all ten real candidates span 39.2×10^{-3} and adjacent ranks differ by a median of 2.43×10^{-3}: **the gap the suite cannot see is $11.6\times$ the difference it treats as a meaningful ranking signal, and 72% of its entire visible spread**.

The same measurement on a public leaderboard. The construction above uses our own serving pipelines, so we repeat it on an artefact we did not build. The Open LLM Leaderboard publishes seven numbers per model (six per-task scores and their equal-weight average) for a pool of 620 models. As just noted, each individual model's profile is affine and those seven numbers determine

it, so a leaderboard used to choose among *static* models has no blind set at all. The blind set appears exactly when a deployer may *route*: a serving stack that picks among several models at deployment has as its profile the lower envelope of its members', which is concave and no longer pinned by the vertices. Routing is routine, model routers, cascades, per-workload deployment.

Call a model *invisible* to the published numbers when it attains the maximum at none of them, so that adding it to the crowned stack changes no reported score; ties are resolved by the linear program at a tolerance of 10^{-9}, and the count below is unchanged at every tolerance from 10^{-12} to 10^{-4}. Among the 620 models, 615 are invisible to all seven published numbers, and 77 of those lower deployment risk somewhere on the mixture simplex (computed exactly, by linear programming over the simplex). Two routing stacks that the leaderboard scores *identically* on every task and on the average differ by **6.54 leaderboard average-points** at a realistic mixture. For calibration, the median difference between adjacently ranked models in the top 200 is 0.034 points: the invisible gap is **190**$\times$ the difference the community already treats as a meaningful ranking signal, and 44% of the spread across those 200 models.

The census behind the figure quoted in Section 4 is the same computation run for every model rather than one pair, and reports each model at *its own* best mixture against the pool-wide median adjacent-rank gap (0.0225 points), rather than one specified mixture against the top-200 gap (0.034). On that basis the 77 models hide a median of 62$\times$ the adjacent-rank gap, 230$\times$ at the ninetieth percentile and 316$\times$ at the most extreme, the last being the same model as above evaluated at 7.11 points rather than 6.54. Both statements are existence claims over mixtures, not claims about typical deployments; `blindset_census.py` and `opera-eval blindspot` produce them independently and agree.

The concrete case is more useful than the aggregate. `Qwen2.5-72B-Instruct-abliterated` ranks 5th of 620 on the headline and is unremarkable per task, 6th on IFEval, 11th on MATH, 18th on BBH, 41st on GPQA, 50th on MMLU-PRO, and *413th* on MUSR, which is what drags its average down. It is nevertheless the best available choice at the mixture (MATH 0.50, IFEval 0.25, GPQA 0.25), a plausible technical-assistant workload. Nothing the leaderboard prints reveals this. Table 1 shows the corresponding persona-level cost of taking the headline #1: zero for the two deployments whose mixture sits near the uniform one it evaluates at, and 3.8 and 5.5 points for the two that do not, with a 32B model beating the 78B headline leader by 5.5 points for a math tutor.

How fast does it shrink? Proposition 3 bounds the invisible gap by $2Lh(\Omega_0)$, so a suite of G conditions is η-complete once $G \gtrsim (2L/\eta)^d$. Figure 6(a) measures the decay on real pipelines. Reporting more conditions does shrink the blind set, and reaching the adjacent-rank resolution of 2.43×10^{-3} takes roughly **150 reported operating conditions**, about twenty-five times what a six-task leaderboard publishes.

We report a negative result alongside it. The measured decay exponent is -0.75, against the worst-case prediction $-1/d = -0.25$: real achievable sets are low-complexity and do not attain the adversarial bound, so the exponential-cost reading of Proposition 3(iii) is a worst case that real systems do not realise. What does survive measurement is the dimensional mechanism. Restricting Ω to faces of dimension $1, 2, 3, 4$ that contain the conflicting axis, and averaging over all such faces, the measured exponents are $-2.11, -1.38, -0.78, -0.75$ against predictions $-1.00, -0.50, -0.33, -0.25$: monotone, correlated at $r = 0.98$, and steeper by a factor of roughly two and a half (Figure 6(b)). The practical conclusion is unchanged, and now quantitative: suites in use are one to two orders of magnitude too small, and the requirement worsens with the dimension of the deployment space, while the dramatic reading of the worst-case rate is not supported by data, and we do not make it.

G.2 Persona sensitivity

Table 5 varies how deployments are drawn on the Open LLM Leaderboard, from authored personas of increasing sharpness to mixtures with no authored structure, and ends at the headline's own uniform mixture.

Table 4: The collapse boundary across five variants of the RLHF chart. "Agreement" is how often the persona-measure integral and the *barycentre-matched* point mass select the same system. "Reduction" compares the persona-measure integral against a *barycentre-matched* point mass, the comparison Proposition 1 makes diagnostic, since the two must coincide when profiles are affine, so any difference isolates curvature and nothing else. Frozen, they agree always and the reduction is exactly zero; adaptive, they do not. Curvature marks that boundary, but across these five adaptive rows it does not predict the size of the gain ($r = -0.004$). "Differ" is the reduction restricted to the shortlists on which the two rules disagree. Regret $\times 10^3$, 240 replications, 95% cluster bootstrap over replications. Shaded: the adaptive gain, positive in every variant; **bold**: frozen, it is exactly 0.0% in every variant.

		Agreement		Reduction vs. barycentre point mass			
Variant	Curv. $\times 10^3$	Frozen	Adapt.	Frozen	Adapt.	95% CI	Differ
UltraFeedback, TF–IDF RM	59.9	100%	89.2%	**0.0%**	**23.8%**	[13.4, 32.3]	65.9%
UltraFeedback, TF–IDF, seed 1	59.6	100%	84.6%	**0.0%**	**25.7%**	[15.6, 35.3]	53.1%
UltraFeedback, MiniLM RM	53.4	100%	81.7%	**0.0%**	**13.4%**	[3.0, 21.9]	35.6%
HelpSteer2, TF–IDF RM	13.5	100%	93.3%	**0.0%**	**15.6%**	[7.9, 22.0]	78.3%
HelpSteer2, MiniLM RM	11.7	100%	87.9%	**0.0%**	**27.8%**	[17.5, 37.5]	69.3%

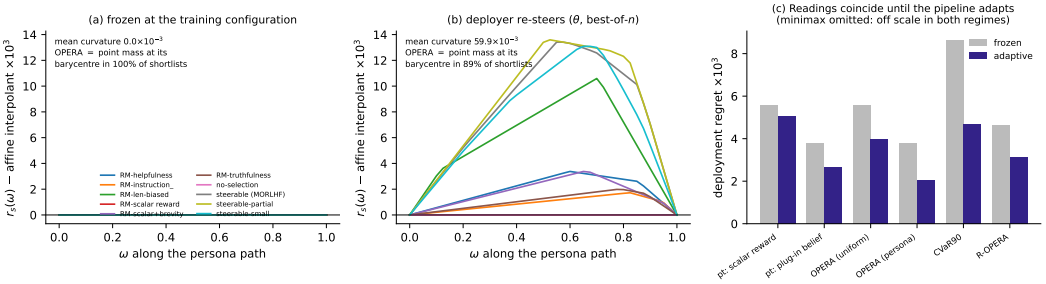


Figure 5: The collapse boundary, turned off and on. Each curve is a serving pipeline's risk profile minus the affine interpolant along a path between two deployment personas. **(a)** Frozen at the training configuration: flat at zero for every pipeline, and every linear aggregation rule coincides with the point mass at its barycentre. **(b)** The deployer re-steers θ and best-of-n: profiles bow, ordered by steerability, and the agreement between the uniform-measure integral and the scalar-reward point mass falls from 100% to 9%. **(c)** Deployment regret by aggregation rule in the two regimes.

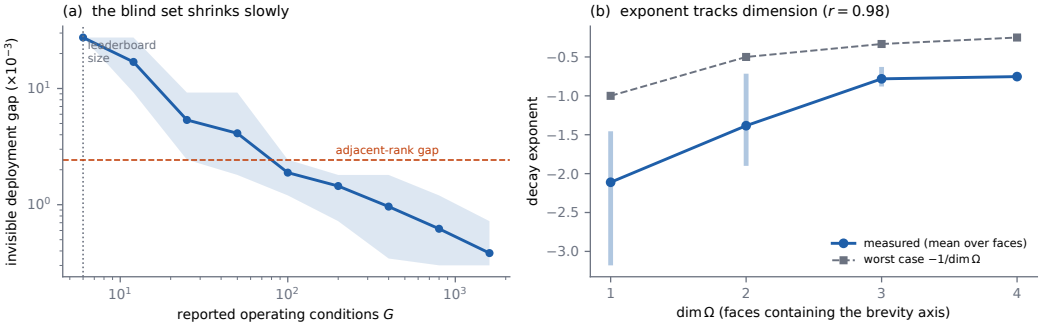


Figure 6: An exhibited blind set. **(a)** Largest deployment-risk gap between two realizable pipelines that a suite reporting G operating conditions cannot distinguish. At the suite size leaderboards actually publish (dotted), the invisible gap is $11.6\times$ the difference between adjacent ranks (dashed); reaching that resolution takes ~ 150 reported conditions. **(b)** The decay exponent tracks the dimension of Ω as covering theory predicts ($r = 0.98$), but real pipelines are about two and a half times more benign than the worst case. Bars span the faces of each dimension.

Table 5: Sensitivity to the deployment family (Open LLM Leaderboard, 620 models, 300 replications). Randomly drawn deployments ($\mathrm{Dir}(\alpha)$, no authored structure) *raise* suboptimality above the authored personas, so the headline is not an artefact of the mixtures we chose. The last row is the degenerate boundary Proposition 1 predicts. Shaded: deployments drawn at random raise suboptimality above the authored personas; **bold**: at the headline's own condition the headline is near-optimal, as the theory requires.

Deployment family		Headline subopt.	OPERA gain
Authored personas, $c = 2$ (vague)	33	37.5%	+24.1%
Authored personas, $c = 8$ (default)		30.7%	+31.3%
Authored personas, $c = 100$ (sharp)		25.2%	+49.8%
$\mathrm{Dir}(0.3)$, sparse		**44.5%**	+71.1%
$\mathrm{Dir}(1)$, uniform on simplex		**39.0%**	+37.2%

G.3 COMPETITIVE SHORTLISTS

Figure 3 draws each shortlist uniformly from the whole pool, which puts weak candidates on some shortlists. A team choosing a model would shortlist plausible ones. We therefore repeat the protocol unchanged except for the pool: the ten candidates are drawn from the headline's own top 20. Table 6 reports both pools.

Table 6: Shortlists drawn from the whole pool, and from the headline's own top 20. "Subopt.": fraction of (shortlist, deployment) pairs where the headline pick is not the mixture-optimal candidate. "Headline regret" and "Best rule" are mean regret in each leaderboard's own units, Elo points for LMArena and loss on the $[0, 1]$ scale elsewhere. Brackets are 95% cluster bootstrap over replications. Shaded: on competitive shortlists the headline is wrong roughly twice as often.

Leaderboard	Pool	Subopt.	Headline regret	Best rule	Gain
LMArena	all	17.2% [14.1, 20.3]	2.453	0.448	82% [73, 88]
	top 20	**57.3% [53.4, 61.3]**	11.393	0.935	92% [89, 94]
HELM Lite	all	23.8% [21.0, 26.7]	.0121	.0076	38% [24, 49]
	top 20	**50.5% [47.5, 53.5]**	.0219	.0125	43% [35, 50]
Open LLM	all	32.2% [29.1, 35.3]	.0088	.0064	27% [11, 40]
	top 20	**64.5% [61.3, 67.7]**	.0293	.0096	67% [63, 71]
BRIDGE	all	26.3% [22.8, 29.6]	.0039	.0018	54% [39, 65]
	top 20	**47.8% [44.1, 51.7]**	.0075	.0023	69% [62, 75]

The competitive pool is harder on the headline in both senses. Suboptimality roughly doubles on every leaderboard, to between 47.8% and 64.5%, and the regret at stake rises by factors of 1.8 to 4.6. The mechanism is visible on the Open LLM Leaderboard, where the rank correlation between leaderboard order and deployment order falls from 0.87 on the full pool to 0.35 on the top 20. Among models that are close on the average, the average carries almost no information about which one suits a given deployment, while its mistakes stop being cheap. The objection that these results depend on shortlists no practitioner would assemble therefore runs the wrong way: the more plausible the shortlist, the more the headline costs.

G.4 THE CLINICAL LEADERBOARD: BRIDGE

BRIDGE (Wu et al., 2026) scores language models on 87 tasks built from electronic health records, case reports and doctor–patient consultations in nine languages, under zero-shot, chain-of-thought and five-shot prompting. Its headline, "Average Performance", is the equal-weight mean over the 87 tasks; we verified it against the per-task scores to 10^{-4}. The gap that motivated the benchmark is large: DeepSeek-R1 scores 92 on USMLE questions and 44.2 out of 100 here. We use the public leaderboard as downloaded on 15 September 2026: 109 models, every task scored for every model under all three strategies. Scores average heterogeneous task metrics, as BRIDGE's own headline does, and regrets below are in BRIDGE's units on the $[0, 1]$ scale.

Sites. A site places mass 0.85 uniformly on the tasks carrying one value of BRIDGE's own metadata: a language, a clinical stage or a task type. Each such value carries at least four tasks, giving 19 sites. The remaining mass is spread uniformly. No weight is authored. The protocol is that of Figure 3: 500 shortlists of $K = 10$, drawn from all 109 models or from the headline's top 20; score noise $\sigma = 0.02$; a noisy estimate $\hat{p} \sim \mathrm{Dir}(1.5\,Tp)$ of the site p; deployment uncertainty as a $\mathrm{Dir}(T\hat{p})$ cloud; realized mixtures $\mathrm{Dir}(0.8\,Tp)$, with $T = 87$ (HELM's concentrations rescaled per task); 95% cluster bootstrap over shortlists. In the *frozen* regime each site runs zero-shot and every profile is affine. In the *adaptive* regime each site also chooses its prompting strategy for its own workload, so $r_s(\omega) = \min_\tau [1 - \omega \cdot S_\tau(s)]$ is an envelope over three configurations.

Results. Over the full pool without noise, the headline leader (gemma-4-31B-it, 46.7) is not the best model at 11 of the 19 sites, by up to 3.3 points at question-answering sites and 2.2 at Spanish-language ones. Table 7 gives the shortlist protocol. The boundaries behave as Proposition 1 requires: at the headline's own uniform mixture the headline is near-optimal (5.0%, noise alone) and the gain is zero, and for near-uniform $\mathrm{Dir}(1)$ sites the gain is not distinguishable from zero, while sparse

Table 7: BRIDGE, 19 metadata sites, 500 shortlists. "Gain": mean-regret reduction of OPERA over the zero-shot headline. "vs. bary.": OPERA against the barycentre-matched point mass, with the share of shortlists on which the two return identical regret. Brackets are 95% cluster bootstrap. Shaded: even when sites choose a prompting strategy, OPERA and the barycentre-matched point mass agree, so the gain is relocation rather than a property of the rule.

Regime	Shortlists	Headline subopt.	Gain (regret)	vs. bary. (ident.)
Frozen	all 109	26.3% [22.8, 29.6]	54% [39, 65] (.0039 → .0018)	0.0% (100%)
Frozen	top 20	47.8% [44.1, 51.7]	69% [62, 75] (.0075 → .0023)	0.0% (100%)
Adaptive	all 109	37.2% [33.4, 41.4]	84% [79, 87] (.0081 → .0013)	**−0.1% (99.8%)**
Adaptive	top 20	72.2% [68.6, 75.6]	89% [87, 91] (.0161 → .0018)	**−1.9% (99.8%)**

$\mathrm{Dir}(0.1)$ sites give 26.4% suboptimality and a 73% [64, 80] gain. Frozen profiles bend by at most 2.3×10^{-15}.

Reconfiguration without curvature. Letting sites choose a prompting strategy makes the zero-shot headline worse, and using the best strategy's average instead recovers only 53% [43, 63] of its regret; the site's mixture recovers the rest. But the adaptive gain is still relocation, not aggregation: OPERA and the barycentre-matched point mass agree in 99.8% of shortlists. Five-shot prompting is the best strategy for nearly every model at nearly every site, so only 28 of the 109 envelopes deviate from affine by more than 10^{-3}, and where they do the kink lies off the sites' support. This is the local collapse of Figure 8(a) on third-party data, and the collapse diagnostic predicts it before any selection is run.

Standing. The sites are hypothesised workloads over BRIDGE's categories, not observed hospital case mixes, and the four assumptions of Appendix F.2 apply unchanged. Absolute regrets are small because the leading models sit within about two points of one another; the finding is that the headline is fragile to the site, not that the stakes of a single pick are large. Code and the downloaded leaderboard are in the supplement.

G.5 Social-choice aggregation baselines

A natural objection is that the mean is simply a poor aggregator, and that a better *fixed* rule would close the gap. Social-choice aggregation over tasks (Rofin et al., 2023) is the standard alternative. We add Borda count and Copeland's rule, computed on the same noisy per-condition scores, to the protocol of Figure 3 without changing any other rule's output. Table 8 reports how often each rule's pick is not mixture-optimal.

Table 8: Replacing the headline mean with social-choice rules, on the protocol of Figure 3 (whole-pool shortlists). "Subopt.": share of (shortlist, deployment) pairs where the rule's pick is not mixture-optimal, with 95% cluster-bootstrap intervals. Shaded: the only rule that improves on the headline on every leaderboard.

Rule	LMArena	HELM Lite	Open LLM	BRIDGE
Headline (mean or rating)	17.2 [14.4, 20.1]	23.8 [21.0, 26.6]	32.2 [29.2, 35.5]	26.3 [22.8, 29.3]
Borda count	11.1 [8.9, 13.4]	26.1 [23.3, 29.0]	45.7 [42.0, 49.3]	27.0 [23.5, 30.2]
Copeland's rule	11.9 [9.6, 14.3]	28.1 [25.0, 31.4]	45.9 [42.4, 49.5]	26.4 [23.2, 29.8]
OPERA	**7.5 [5.9, 9.4]**	**17.5 [15.3, 19.7]**	**26.1 [23.5, 28.8]**	**17.7 [15.0, 20.3]**

Social-choice rules improve on LMArena, whose headline rating is fixed to the arena's own traffic, but are worse than the mean on the other three leaderboards, markedly so on the Open LLM Leaderboard. Each is still one fixed rule evaluated at one implicit deployment; the profile rule is the only one of the four that improves on the headline everywhere.

G.6 A natural deployment study on logged Arena traffic

Every leaderboard deployment in Section 4.1 is a hypothesized mixture. Here the workload mixtures are observed.

Data and construction. We use `arena-human-preference-140k` (LMArena, 2025): 135,634 human votes on 53 models from 17 April to 24 July 2025, each conversation carrying task flags and a detected language. Only vote-level fields are read; no conversation text is used. Each conversation is assigned to one of five categories by precedence (math, then code, creative writing, instruction following, other), and models with fewer than 1,000 votes are dropped, leaving 48 models and 133,546 votes. Per category we fit a Bradley–Terry model, counting ties and "both bad" as half a win, and take a model's value in that category to be its win probability against a fixed reference model (`claude-opus-4-20250514`, the most voted). A model's value on mixture w is then its expected win rate against the reference on a query drawn from w, which is exactly linear in w, so single models have affine profiles. A *vendor*, meaning a deployer who adopts a provider and serves whichever of its models is best for its own workload, has as its profile the upper envelope over that provider's models; the nine providers with at least two models qualify.

Observed deployments. The pooled traffic mixture is 8.0% math, 26.9% code, 8.0% creative, 7.3% instruction following and 49.8% other. Eleven language markets with at least 1,500 conversations define observed deployments: code ranges from 14.8% (German) to 32.0% (English) and creative writing from 6.6% to 18.0% (Korean). We run two protocols. In the *market* protocol the deployer knows the other ten markets and enters the held-out one (11 deployments); in the *temporal* protocol it knows a market's earlier months and deploys into the next (25 deployments). Rules are the headline (the pooled mixture), a point mass at the barycentre of the known mixtures, and OPERA, the average over them. Shortlists are 500 random sets of ten single models and all 126 sets of four vendors.

Results. Observed markets stay close to the pooled mixture, within total variation 0.085 at the median and 0.135 at most, and there rankings rarely change: the headline pick is not mixture-optimal in 1.6% [0.5, 2.6] of single-model shortlists under the market protocol and 2.2% [1.3, 2.9] under the temporal one, and in no vendor shortlist. Model profiles do cross: at constructed deployments concentrating 50% or 80% of traffic on one category, with the rest at the pooled mixture, the headline is not optimal in 14.8% and 23.1% of single-model shortlists. The leaderboard effect of Section 4.1 is therefore a property of concentrated, product-like deployments, not of broad consumer markets, which is what the collapse boundary predicts: near the barycentre, the barycentre's ranking is right.

Vendors curve, but not where real deployments sit. Over the observed mixtures the Jensen gap between integrating a vendor's profile and evaluating it at the barycentre averages 0.0002 win-rate points, is at most 0.019, and is positive in only 3.4% of vendor–deployment pairs, because each provider's best model does not change across the observed range. OPERA and the barycentre point mass select identically in every shortlist under both protocols. This is the local collapse of Section 4.2 on natural data: the one adaptive system we can build from logged traffic has its kinks outside the region its deployments occupy. We note two limits. In this window one model family leads every category, and sessions are too short (median one vote) to recover individual users' workloads, so language markets are the finest observed deployments available.

G.7 Controlled stress tests and legacy benchmarks

Summary of the four controlled experiments (200 replications each, $c = 0.3$): (1) *Real-data prevalence shift* (GUSTO-I, UCI Heart, and a congenital cytomegalovirus screening cohort, cCMV): OPERA reduces point-mass regret by 77%/31%/23%, gains concentrated at deployment prevalences far from training; selection instability 0.47 vs. 0.54. (2) *Crossing-profile stress test* (five DGP families): OPERA's advantage increases monotonically with crossing magnitude (25× from lowest to highest quintile), overall 25% regret reduction, the empirical face of Proposition 2. (3) *Robustness sweep*: worst-case regret interpolates continuously in ε (Proposition 4 observed); even at $\varepsilon = 0.2$, R-OPERA improves 21% on point-mass worst case. Per-experiment tables, together with the 6×6 care-unit transfer matrix referenced in Section 4, are included in the reproducibility package.

H RLHF Serving Pipelines

H.1 Setup and additional checks

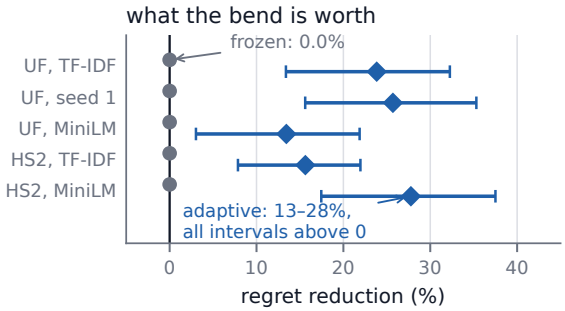


Figure 7: **What the bend is worth.** Deployment regret avoided by integrating the profile instead of scoring at the average condition, against a barycentre-matched point mass, across five variants (UF: UltraFeedback; HS2: HelpSteer2) with 95% cluster bootstraps over 240 replications. Frozen: 0.0%, exactly, in every variant. Adaptive: 13–28%, every interval above zero.

This appendix gives the full setup of the RLHF experiment of Section 4.2, the real adaptive system, the axis-rescaling study, and further robustness checks.

H.2 Full experimental setup

$\Omega = \Delta^4$ is the simplex of deployment *preference weightings* over five axes a served response is judged on. On UltraFeedback (Cui et al., 2024) (88,843 completions, 22,566 prompts, 17 generator models) these are the four annotated aspects (instruction-following, honesty, truthfulness, helpfulness), plus conciseness, one minus the fraction of a 512-token serving budget consumed. On HelpSteer2 (Wang et al., 2024b) (10,680 prompts $\times$ 2 responses) all five are human-annotated: helpfulness, correctness, coherence, brevity ($1 -$ verbosity$/4$) and depth (complexity$/4$); there the conflicting axis is annotated rather than constructed, and verbosity is essentially orthogonal to helpfulness within prompt ($\rho = 0.017$).

A candidate is a *serving pipeline*: a reward-model family $\{R_\theta : \theta \in \Theta_s\}$ together with a best-of-n budget, both reconfigurable at deployment. On a shared pool of real responses per prompt, the risk (3) of the body, so each configuration contributes a function affine in ω and the profile is the lower envelope of finitely many affine functions, curved exactly to the extent that the deployer can re-steer. Specialist reward models spend the whole annotation budget on one objective and cannot be re-steered; multi-objective ones split the same budget across all axes and can be re-weighted at deployment, so configuration sets are *not* nested and profiles cross rather than dominate. Prompts are split disjointly into reward-model training, selection (6,453 prompts) and deployment (6,424 prompts); reward heads are ridge regressions on response text (held-out Spearman 0.38–0.56), and we replicate every number with frozen MiniLM sentence embeddings in place of TF–IDF features. All criterion measures used here (uniform on the mixture lattice, and Dirichlet persona measures) have full support, so $\Lambda(r_s; \mu_0) = \mathrm{Lip}(r_s)$ exactly (Theorem 5) and the reported sensitivities are unaffected by the distinction; R-OPERA is computed as the first-order criterion $\widehat{\Phi}_\varepsilon$ throughout.

H.3 A real adaptive system

Design. 240 prompts are drawn from each of the select and deploy splits of Section 4.2 (same seed and cut). For each, Qwen2.5-7B-Instruct (NF4) generates eight responses at temperature 0.8: two under each of four system prompts (none, concise, thorough, careful), so the pool spans the quality–brevity trade-off that multi-model pools supply offline. ArmoRM's nineteen objective heads are evaluated on every response; the repository's remote code no longer imports under current `transformers`, so the model is reassembled exactly from its weights (Llama backbone, regression head on the last non-pad token, gating network on the assistant-header token) and checked to

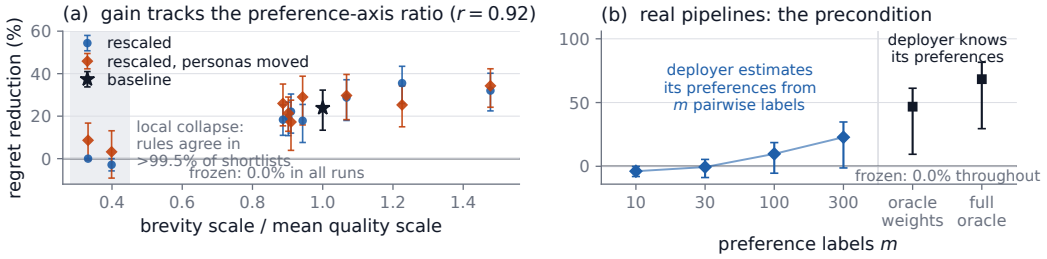


Figure 8: **The boundary, and its precondition.** Regret reduction over a barycentre-matched point mass (95% intervals; frozen, exactly 0.0%). **(a)** Under 18 axis rescalings the gain vanishes where the kink leaves the measure's support (Appendix H.4). **(b)** On real pipelines it needs a deployer who knows its preferences (Appendix H.3).

rank a correct answer above an incorrect one. The four UltraFeedback heads are calibrated to aspect units by per-axis linear fits on the select split only. OLMo-2-13B-Instruct (NF4, greedy) rates each response 1–5 on UltraFeedback's four rubrics; all 3,840 ratings parsed. The steerable pipeline re-weights ArmoRM's heads; the weak one uses the paper's TF–IDF ridge heads on the same fresh responses; the scalar pipeline uses ArmoRM's gated score. Everything downstream is the paper's code.

Estimated preferences. In the adaptive regime the deployer chooses its configuration by minimising estimated risk, $\hat{\tau} = \arg\min_\tau [1 - \hat{\omega} \cdot \hat{m}(\tau) + \lambda c(\tau)]$, with $\hat{m}$ the select-split means and $\hat{\omega}$ a Bradley–Terry fit to m pairwise comparisons simulated from the true weighting (logit scale 8), projected onto the simplex; it pays the true risk $1 - \omega \cdot m(\hat{\tau}) + \lambda c(\hat{\tau})$ on the deploy split. Frozen pipelines have one configuration, so these results are invariant to estimation by construction.

Deployer	Frozen	Adaptive
Oracle weighting and configurations (the paper's protocol)	0.0	68.3 [29.5, 81.6]
Oracle weighting, configurations estimated	0.0	46.8 [9.4, 61.2]
Bradley–Terry weighting, $m = 10$ labels	0.0	−3.9 [−8.0, −0.2]
Bradley–Terry weighting, $m = 30$	0.0	−0.6 [−8.8, 5.4]
Bradley–Terry weighting, $m = 100$	0.0	9.7 [−5.4, 18.6]
Bradley–Terry weighting, $m = 300$	0.0	22.8 [−1.3, 34.8]

Entries are the regret reduction (%) of the persona-measure integral over the barycentre-matched point mass, 240 replications, 95% cluster bootstrap. The two agree in 95.8% of adaptive shortlists, so the gain is concentrated in the 4.2% where they differ; absolute regrets are small (0.10 versus 0.31×10^{-3}). Held-out Spearman correlations with the judge are 0.55 (helpfulness), 0.28 (honesty), 0.27 (instruction-following) and 0.08 (truthfulness) for ArmoRM's heads, and 0.01 for its gated score; the weak TF–IDF heads reach 0.35 on helpfulness and are near zero elsewhere. R-OPERA's regret is the highest of any rule here (1.43×10^{-3}), the purchase-of-robustness effect of Section 4.2 under a well-specified measure.

H.4 Sensitivity to the preference axes

The full pipeline (reward-head training, pools, envelopes, aggregation rules and regret) is rerun with every aspect axis rescaled, $a_j \mapsto \alpha_j a_j$, applied after the $[0, 1]$ clip so that $\alpha_j > 1$ is not saturated, and feeding both the ground truth and the reward-head targets. Nine rescalings are drawn log-uniformly in three bands, each run with the nominal personas (U) and with the personas carried through the rescaling, $p \mapsto (p/\alpha)/\|p/\alpha\|_1$ (T). "Ratio" is the brevity scale over the mean quality scale; J is the spread across systems of the Jensen gap under the persona measure, and "Dis." the share of shortlists on which OPERA and the barycentre point mass disagree, both for U.

Rescaling band	Ratio	$J\times10^3$	Dis. (%)	Reduction U (%)	Reduction T (%)
Baseline	1.00	1.86	10.8	23.8 [13.4, 32.3]	n/a
$[0.25 - -4.0]$ #1	0.90	1.33	12.1	19.4 [10.8, 26.6]	21.3 [12.9, 29.0]
$[0.25 - -4.0]$ #2	0.33	0.43	0.0	0.0 [0.0, 0.0]	8.6 [−0.8, 16.7]
$[0.25 - -4.0]$ #3	0.94	1.78	10.8	17.9 [7.6, 25.5]	29.0 [15.9, 38.8]
$[0.5 - -2.0]$ #1	0.40	0.54	0.4	-2.9 [−5.8, 0.0]	3.2 [−9.2, 13.1]
$[0.5 - -2.0]$ #2	1.48	4.70	15.4	32.0 [22.5, 40.2]	34.3 [24.2, 42.3]
$[0.5 - -2.0]$ #3	0.89	1.83	9.6	18.3 [11.0, 25.3]	26.0 [15.4, 35.1]
$[0.8 - -1.25]$ #1	1.23	3.11	15.0	35.5 [26.0, 43.5]	25.3 [15.0, 34.2]
$[0.8 - -1.25]$ #2	1.07	1.85	12.5	28.8 [18.0, 37.1]	29.7 [18.4, 39.6]
$[0.8 - -1.25]$ #3	0.91	1.80	10.8	22.1 [12.0, 30.4]	17.2 [3.9, 27.5]

An additive shift $(0.3, -0.2, 0.1, 0.25, -0.15)$ gives 23.83% against the baseline's 23.84%. The frozen reduction is exactly 0.0% in all 19 rescaled runs. Across runs, the brevity-to-quality ratio predicts the adaptive reduction at $r = 0.92$, the local Jensen spread at Spearman 0.85, and global curvature at only 0.38.

H.5 Additional checks

Shortlist size does not drive the result either: repeating the whole experiment at $K = 4$ and $K = 9$ gives frozen agreement 100% in both, adaptive agreement 28.7% and 0.0%, and regret reductions of 64.8% and 58.3%, the collapse is exact regardless, and the adaptive agreement falls as the shortlist grows, which is what more opportunities to disagree should do.

The failure of the scalar-reward point mass is concentrated where the theory says it must be: it *beats* OPERA for the two personas sitting nearest the uniform weighting it evaluates at, and is $4.8\times$ worse for the one deployment that prices brevity heavily (18.6 versus 3.9×10^{-3}). Measured against the barycentre-matched point mass the two datasets agree throughout; measured against the scalar-reward point mass they appear to disagree (46% versus 60%), which is the conflation of barycentre relocation with curvature rather than a property of either dataset. Carrying the personas through an axis rescaling does not restore invariance (Appendix H.4): the constant and the serving cost do not rescale, so the rescaling is not a pure reparameterization of Ω.

Two side observations are worth recording because they discipline claims made elsewhere in this paper. First, minimax does *not* collapse in the frozen regime (53.761 versus 5.551×10^{-3}), which is why Proposition 1 is stated for linear aggregation rules only. Second, R-OPERA's regret rises monotonically in ε here ($1.98 \to 29.2 \times 10^{-3}$ as $\varepsilon : 0 \to 0.5$), because deployment uncertainty in this chart is well specified: robustness is a purchase, and Lemma 6 is a worst-case statement, not a promise of lower regret at a known measure.

H.6 The risk–sensitivity pair and the frontier

The pair, and what it summarises. The robust value (4) is first-order linear in the radius (Theorem 5(ii)), so the pair

$$\Gamma(s) \;=\; \big(\Phi_0(s),\ \Lambda(r_s;\mu_0)\big) \qquad (\textit{integrated risk}, \textit{sensitivity}) \tag{9}$$

which we call the *risk–sensitivity pair*, summarizes a system's deployment behaviour to first order. Plotting candidates at $\Gamma(s_k) \in \mathbb{R}^2$ gives the *deployment frontier*: the selections of $\widehat{\Phi}_\varepsilon$ over $\varepsilon \in [0, \infty)$ are exactly the extreme points of the lower-left convex hull (Proposition 8).

In words: score each candidate twice, for its average risk and for how fast that risk moves when the deployment shifts, then plot the two. Whatever your appetite for caution, the candidate you pick lies on the lower-left edge, so the interior can be dropped. Both numbers cost $O(nG + G^2)$ operations on top of ordinary cross-validation (Appendices H.6, E.6).

The geometry also yields a computable object per system. The robust value (4) is first-order linear in the radius (Theorem 5(ii)), $\Phi_\varepsilon(s) = \Phi_0(s) + \varepsilon\Lambda(r_s;\mu_0) + o(\varepsilon)$, so the pair the signature (9) of the body, summarizes a system's deployment behaviour to first order. We call it the *risk–sensitivity pair*. Both coordinates come from the cross-validated class-conditional estimates standard evaluation already produces, at a marginal cost of $O(nG + G^2)$ scalar operations (Appendix E.6).

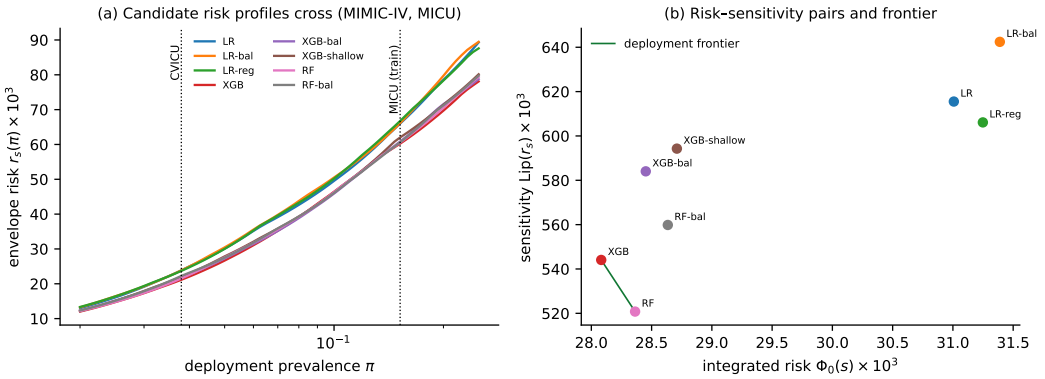


Figure 9: The two objects, on the clinical chart of Appendix I.1 (MIMIC-IV, selection on MICU). **(a)** Envelope risk profiles of the eight candidates cross along the deployment prevalence range; the best model at the training prevalence (dashed) is not the best at the CVICU deployment (dotted). **(b)** Risk–sensitivity pairs $\Gamma(s) = (\Phi_0, \Lambda)$; the first-order robust selection path traces the deployment frontier: these are exactly the *first-order* robust selections (Proposition 8); Section D.5 verifies that exact R-OPERA makes the same selections throughout the operating range.

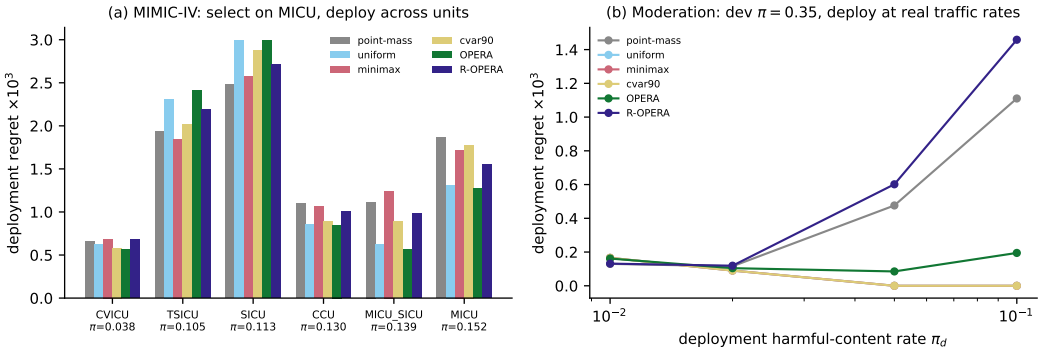


Figure 10: Deployment regret by aggregation rule. **(a)** MIMIC-IV: select on MICU, deploy across the six care units (50 replications). **(b)** Moderation: develop at enriched $\pi = 0.35$, deploy at real traffic rates (30 replications).

Plotting each candidate at $\Gamma(s_k) \in \mathbb{R}^2$ gives the *deployment frontier*: the selections of $\widehat{\Phi}_\varepsilon$ as ε ranges over $[0, \infty)$ are exactly the extreme points of the lower-left convex hull supported by normals $(1, \varepsilon)$, so a candidate strictly inside the hull is never selected at any robustness level (Proposition 8, Appendix F.1). This is the paper's operational deliverable: one scatter plot on which the robust selection path traces the hull and the point-mass pick is visibly interior whenever profiles cross. Signature error obeys the same concentration bounds as selection, requiring $\min(n_0, n_1) \gtrsim \log(KGn/\delta)/\gamma^2$ with *no* dependence on the ratio of deployment to evaluation prevalence, no importance weighting, hence none of its variance (Appendix E.6).

I Clinical and Safety Replications

I.1 MIMIC-IV: care-unit and temporal shift at one hospital

One row per first ICU stay ($n = 76{,}627$; MIMIC-IV, Johnson et al., 2023), outcome in-hospital mortality, features demographics + first-24h labs and vitals. Deployment shift is real, not resampled: mortality prevalence ranges from 0.038 (CVICU) to 0.152 (MICU) across the six major care units of the *same* hospital, and from 0.106 (2008–2010) to 0.154 (2020–2022) across anchor-year groups. Eight modern candidates (regularized/balanced logistic regression, gradient boosting, random forests) are trained and selected on MICU, then deployed across units; candidate envelope profiles cross 3–4 times along the prevalence range, so the setting is squarely in the curved regime

of Remark 12. All numbers come from one pinned software environment (scikit-learn 1.6.0, XGBoost 2.1.4), in which the script is deterministic; tree-ensemble candidates are not reproducible across library versions, so we report 95% cluster-bootstrap intervals over replications and read only effects that survive them.

Across 50 replications with leakage-free held-out evaluation we compare the integral against a *barycentre-matched* point mass, the comparison Proposition 1 makes diagnostic, rather than against the point mass at the development prevalence, which would confound the choice of rule with the choice of condition. The two rules select differently on 34% of (replication, unit) cells, so the curvature is real and consequential. The benefit is not: pooled over units the reduction is -4.0% $[-16.3, 6.6]$, and *every* unit's interval covers zero, from CVICU at -16.8% $[-42.7, 3.0]$ to CCU at $+14.5\%$ $[-8.8, 29.4]$ (MICU $+5.2\%$, MICU/SICU $+6.9\%$, SICU -14.1%, TSICU -6.1%). Against the development point mass the integral appears to gain $+5.4\%$, but that number rewards relocating to the deployment range and says nothing about the aggregation rule.

The losses have a measurable cause. A profile estimated on MICU transfers only under label shift, when class-conditional score distributions are shared across units. Black-box shift estimation (Lipton et al., 2018), which is valid under exactly that premise, estimates the surgical units' prevalence at 0.003 and 0.005 against true values of 0.113 and 0.105, and CVICU's at 0.18 against 0.038, so the premise fails where integration loses. With profiles this noisy the estimation term of Corollary 3 dominates the curvature the diagnostic finds, which is the regime moderation (Appendix I.3) escapes on two orders of magnitude more data.

Label-shift baselines. A deployer with unlabeled data from its own unit can estimate that unit's prevalence and select there, which uses information OPERA does not. Over the same 50 replications, whose point-mass and OPERA rows reproduce the run above exactly, selecting at the BBSE estimate is *worse* than the development point mass (-10.3% $[-20.2, -1.5]$), and so is selecting at the *true* deployment prevalence (-6.5% $[-14.0, 0.2]$); both comparisons are against that point mass, as is the $+5.4\%$ they should be read beside. Knowing where deployment sits does not help when the profile being read there no longer describes it. Figure 9 shows the crossing profiles and the deployment frontier, and Figure 10(a) the per-unit regrets.

Under temporal shift (select on 2008–2010, deploy on later year groups; 50 replications), integration is significantly better only at 2014–2016 ($+39\%$ $[15, 58]$) and loses at 2020–2022 (-35% $[-83, 0.3]$), the pandemic-era period in which case mix changes the class-conditionals themselves. When the *profile* estimated at development no longer describes deployment, no aggregation rule can compensate. R-OPERA is worse than the point mass at every year group (for example 2.36 against 1.30×10^{-3} at 2020–2022): with a well-specified deployment measure, hedging is a purchase rather than a gain, the same effect recorded at the end of Section 4.2.

I.2 eICU-CRD: NINETY-SIX HOSPITALS

Data and cohort. eICU-CRD (Pollard et al., 2018) pools ICU stays from US hospitals. We take one row per patient, the first unit stay (`unitvisitnumber` $= 1$), requiring a recorded discharge status and a matching APACHE physiology record; this joins `patient.csv` to `apacheApsVar.csv` on `patientunitstayid`. Of 208 hospitals, 96 have at least 500 such stays and are retained, giving 128,169 stays. The eligibility threshold was fixed before any outcome was examined.

Features, outcome, missingness. Features are the 24 APACHE Acute Physiology variables plus age and sex (26 in total); eICU codes missing physiology as -1, which we map to missing and impute at the median of the *source* pool only, never using any row from a deployment hospital. Age is numeric with "$>$ 89" mapped to 90. The outcome is in-hospital mortality (`hospitaldischargestatus` $=$ Expired), at hospital discharge rather than a fixed horizon. Pooled prevalence is 0.0923; per hospital it runs from 0.024 to 0.178 with median 0.085.

Protocol. Deployments are hospitals. Hospitals are partitioned at random into six folds; for each fold the source pool is every hospital outside it, split in half into training and validation, and each hospital inside it is a separate deployment. No deployment hospital contributes to the profiles it is evaluated against, and imputation is fitted on training rows only. Candidates are the eight families of

Appendix I.1, unchanged, with unchanged hyperparameters. Profiles are envelopes over thresholds on the validation split, on the same log-odds grid of 60 points over $\Omega = [0.02, 0.20]$, which covers every hospital's prevalence. Rules are those of Table 2, with $\varepsilon = (b - a)/4$. At deployment every rule is granted the same threshold re-optimisation at the hospital's realized prevalence, so the comparison is between selections, not between thresholds. Regret is measured against the best candidate at that hospital. Eight replications differ in fold assignment and candidate seed. The source caps of prediction P2, the dose-response prediction of the protocol below, are 9,330 training rows, matching the MIMIC study exactly, then 20k and 40k, then the full 53,389; each cap draws an independent random subsample per replication rather than a nested sequence. The fold design is one departure from the protocol, which held hospitals out one at a time; six folds bound compute and still keep every deployment hospital out of the source pool it is evaluated against.

What was fixed in advance, and when. The forecast, design, comparator, and the four predictions are recorded in `eicu/PREREGISTRATION.md` in the supplement, written and hashed (SHA-256 `ff89367d...`) before any eICU outcome was examined; the analysis script `eicu/analyze_eicu.py` was written before the runs completed. This is a timestamped internal protocol, not a public registry entry. It fixed the barycentre-matched comparator, the hospital eligibility rule, the candidate set, the observed-data estimand, and a commitment to report the falsifier as the headline.

Uncertainty. The resampling unit is the replication ($n = 8$); intervals are 95% two-sided percentile cluster bootstraps over replications. The interval for the full-source gain, $[0.0, 4.5]$, is not clipped: the bootstrap distribution has an *atom* at exactly zero carrying 10% of its mass, and no draw is negative. The reason is visible in the per-replication gains, which are 0.00 in six of the eight replications and $+5.85$ and $+11.00$ in the other two. At full source the two rules agree on every hospital in most replications, so most replications produce no difference at all; the headline $+2.36\%$ is an average over that mixture, not a typical value.

Why a source-prevalence point mass wins. It gains 3.2% against the barycentre-matched point mass, beating the integral's 2.4%, and the reason is relocation rather than rule choice. The barycentre of our log-odds-uniform default on $[0.02, 0.20]$ is 0.054, while the median deployment hospital sits at 0.085 and the pooled source prevalence at 0.089. The source prevalence is therefore close to the middle of the deployment population and our default's barycentre is not. This is the paper's own thesis applied to itself: where the differences barely bend, what remains is the question of *where* you evaluate, and a better-placed point mass wins.

I.3 Safety moderation under prevalence shift

Moderation classifiers are developed on class-enriched data, here Civil Comments (Borkan et al., 2019) enriched to $\pi_{\text{dev}} = 0.35$, standard practice since harmful content is rare, then deployed on traffic with $\pi \in [0.01, 0.10]$, varying by surface and time. The operating range is the *deployment* range $\Omega = [0.005, 0.15]$; the development prevalence lies outside it, which is precisely the point-mass rule's failure mode. Heterogeneous candidate families (word- and character-level linear models, naïve Bayes, calibrated SVM, gradient boosting, and a lexicon blocklist) produce envelope profiles that cross inside the deployment zone: character-level and margin-based models dominate below $\pi \approx 0.02$, boosting above. An earlier version of this experiment evaluated the point mass at the grid point nearest 0.35 on a grid spanning only $[0.005, 0.15]$, that is, at 0.15, inside the deployment range. The numbers below place it correctly at 0.35, and are rebuilt in the pinned environment of Appendix I.1; every other rule's selections are unaffected by the correction.

Two comparisons are worth separating. Against a point mass at the *development* prevalence, any improvement mixes the choice of rule with relocating into the deployment range. Against a *barycentre-matched* point mass, which is where Proposition 1 says a linear rule must land if difference profiles are affine, the improvement isolates the rule. Here the two comparisons nearly coincide, because both point masses miss: over 30 replications the integral and the barycentre point mass select differently on 30% of cells, and the integral reduces regret by $+72\%$ $[38, 91]$, while a uniform measure reaches $+87\%$ $[76, 92]$ and minimax ties it. This is the paper's clearest aggregation gain, and our own default measure is not the one that earns it.

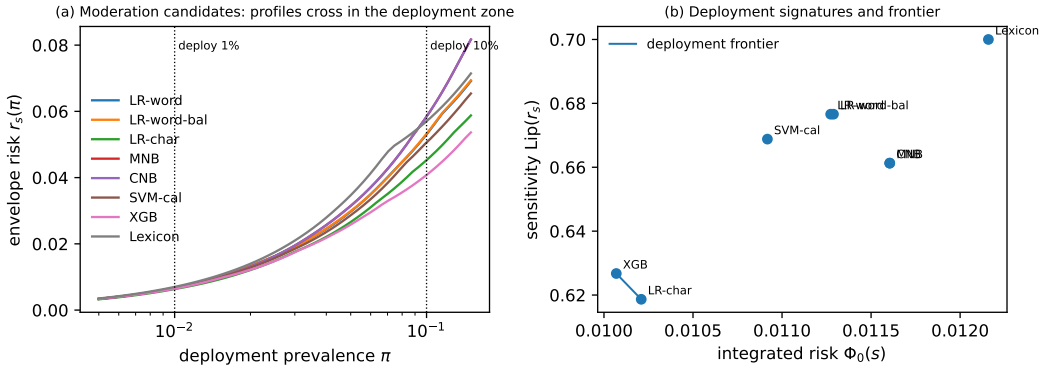


Figure 11: Moderation candidates (one replication), computed by our reference implementation (anonymized). **(a)** Envelope profiles over the deployment range; heterogeneous families cross in the deployment zone. **(b)** Risk–sensitivity pairs; the frontier's extreme points (XGB, char-level LR) are exactly the first-order robust selections as ε sweeps.

Across 30 replications, selecting at the development prevalence is costly, and every rule that reads the deployment range improves on it. Uniform weighting, minimax and CVaR, which here select the same candidate, reduce pooled regret by 86% $[75, 91]$. OPERA reduces it by 70% $[38, 90]$: it gains 82% at $\pi_d = 0.05$ and 0.10 but loses 24% $[-53, -2]$ at $\pi_d = 0.01$, because the log-odds-uniform weight favours low prevalences, where a different candidate looks best on validation data. R-OPERA hedges toward a candidate that is poor at moderate prevalence and does not improve on the point mass (-26% $[-98, 20]$), the Bayesian-to-minimax trade of Proposition 4 landing on the wrong side here. The replication therefore supports the *where* claim strongly and shows that the log-odds default is not the best reading on this task, consistent with its sensitivity to the lower endpoint (Appendix I.5); the earlier claim that R-OPERA's gains concentrate at the deepest shift came from the misplaced point mass and does not survive the correction. Figure 11 shows the candidate profiles and the resulting deployment frontier, computed end-to-end by the reference implementation: the frontier has two extreme points (gradient boosting and the character-level model), and the first-order robust selection path moves from the first to the second as the robustness radius grows, the frontier-sweep proposition (Proposition 8) realized on live candidates.

I.4 LLM ROUTING ON ROUTERBENCH

Routing chooses which model answers a query, and deployments differ by orders of magnitude in what they pay for accuracy. On RouterBench (Hu et al., 2024), $36{,}497$ logged queries answered by 11 models with each answer's cost recorded, the operating condition is that willingness to pay. A candidate is a portfolio of three models: serving one fixed model its profile is affine, routing by task category it is an envelope. Frozen, the two rules agree in every shortlist (0.0%). Routed, they differ on 20%, so the curvature is real and consequential for *which* system is chosen. The benefit is not stable. The integral's advantage over a barycentre-matched point mass runs from $+0.3\%$ to $+41.7\%$ depending on where the willingness-to-pay interval is placed (table below), and at the interval we would have chosen on the grounds that it contains every switch of the cheapest adequate model, $[50, 5000]$, it is absent. A cluster bootstrap over task names gives $+7.1\%$ on average, but that resamples the workload as well as the noise; at the *observed* task composition the integral loses 2.2%, and 23% of resamples are negative. The bootstrap interval is a correct interval for the wrong estimand, so we report the observed-workload number, and report routing as a system where difference profiles bend, not one where aggregation pays. Routing by task category also presumes the category is known before the query is answered.

Willingness-to-pay range	Rules agree	Diff./spread	Gain at observed workload
$[10, 1000]$	63.7%	0.580	$+29.8\%$
$[20, 500]$	71.0%	0.388	$+29.5\%$
$[50, 5000]$	82.0%	0.546	$+\ 0.3\%$
$[100, 2000]$	91.3%	0.355	$+\ 9.9\%$
$[200, 20000]$	61.3%	0.201	$+41.7\%$
$[500, 50000]$	68.7%	0.083	$+39.7\%$
$[1, 100000]$	65.7%	0.164	$+28.0\%$

I.5 Sensitivity to the default measure's endpoints

On a truncated deployment interval $[a, b]$ the log-odds-uniform measure is a default whose normaliser depends on a and b (Appendix E.4). We measure how much that matters. For each replication of an experiment the candidates are fitted once, and selection is then repeated under a grid of intervals around the default, on one logit grid covering all of them, so every difference is due to the endpoints alone.

Clinical (MIMIC-IV). The default interval is $[0.02, 0.25]$; the grid multiplies each endpoint by up to two in either direction and adds a very wide and a tight interval, both still covering every deployment unit's prevalence. Over the 50 replications of Appendix I.1 we record how often OPERA's pick differs from its pick under the default interval, the pooled regret reduction over the point mass at the *development* prevalence, and the number of care units at which OPERA beats it; the eleven-interval table is in the supplement. That baseline mixes relocation with the choice of rule, so the reduction columns here are not comparable with the barycentre-matched numbers of Appendix I.1; what this table isolates is the sensitivity of the *pick* to the assumed interval, which no choice of baseline affects.

Moving an endpoint by up to a factor of two changes OPERA's pick in 4 to 18% of replications, and 22% for a very wide interval. The pooled reduction is $+2.7\%$ $[-6.1, 10.4]$ at the default and ranges from -0.7% to $+7.4\%$ across intervals, every interval includes zero, and OPERA beats the point mass at four of six units under every interval. The endpoints therefore move the size of the clinical effect but not its sign pattern, and not the conclusion of Appendix I.1 that the effect is not significant when pooled. Discretisation matters on the same scale: the default interval gives $+2.7\%$ on this 240-point grid and $+5.4\%$ on the 60-point grid of the main run.

Moderation (Civil Comments). The default interval is $[0.005, 0.15]$ and the point mass sits at the development prevalence 0.35. Over the 30 replications of Appendix I.3, pooled over the four deployment prevalences, the reduction at the default is $+70\%$ $[34, 90]$; it stays at $+70\%$ for every lower endpoint from 0.005 to 0.01 and rises to $+86\%$ at $[0.01, 0.15]$ and $[0.01, 0.25]$ (eleven-interval table in the supplement).

Here the default is robust to its upper endpoint and to lower endpoints between 0.005 and 0.01, and sensitive only when the lower endpoint drops below half a percent, where the weight $1/[\pi(1-\pi)]$ concentrates on very low prevalences: picks then change in 10 to 20% of replications and the pooled gain is no longer distinguishable from zero. Across both experiments, then, the endpoints matter at the extremes of the interval and are otherwise second order, and an elicited deployment distribution should replace the default whenever one exists.