

Overlap as a Design Variable: Optimal Interventions for Robust Reward Modeling and Downstream Alignment

Liz Lemma Future Detective

January 22, 2026

Abstract

Modern RLHF pipelines often train reward models on opportunistically collected preference data, where latent attributes of responses (truthfulness, helpfulness, formatting, length) are strongly correlated. The source material emphasizes that such limited latent positivity can yield reward misidentification: models fit in-distribution but fail under shifts that break training correlations, producing coherent yet misaligned behavior (goal misgeneralization). We formalize this problem as an experiment-design task. We build a tractable Bradley–Terry–Luce (BTL) preference model with a low-dimensional latent factorization Z , allow heterogeneous evaluator objectives C (observed), and let the platform choose targeted interventions—controlled response edits, prompt randomization, or counterfactual labeling—to reshape the training distribution over factor differences. We (i) quantify how limited overlap induces ill-conditioned Fisher information and high-variance reward estimates, (ii) propagate reward estimation error through downstream policy optimization to obtain welfare-regret bounds under distribution shift, and (iii) solve for cost-minimizing intervention allocations that achieve a target worst-case robustness guarantee. The headline result is an optimal design characterization: budget should be allocated to interventions that increase information in the highest-leverage latent directions (those that change which policy is selected), until a minimum overlap floor is reached. We complement theory with an empirical active-intervention loop on an open preference dataset using controllable text generation, demonstrating improved OOD robustness at equal labeling cost relative to naive data scaling.

Table of Contents

1. 1. Introduction: goal misgeneralization as a consequence of limited overlap and reward misidentification; why intervention design (not just scaling) is the 2026 bottleneck.

2. 2. Related work: goal misgeneralization; causal preference learning/positivity; reward hacking; experimental design for discrete choice/logistic models; active preference learning.
3. 3. Model: latent-factor BTL with observed objectives C ; designs as mixtures of intervention arms; definition of overlap via Fisher/restricted eigenvalues.
4. 4. Estimation: MLE for BTL/logistic regression on factor differences; asymptotic and finite-sample error bounds as a function of the design information matrix.
5. 5. Downstream decision and welfare: candidate-policy selection using learned reward; welfare regret bounds under policy misselection; link to coherent-but-wrong behavior under shifts.
6. 6. Optimal intervention allocation: convex program for cost-minimizing designs achieving a regret target; closed-form characterization in $d = 2$ and two-policy-contrast cases; discussion of when numerical methods are needed.
7. 7. Empirical demonstration: implement controllable interventions (attribute editing) on an open preference dataset; measure induced overlap, estimation variance, and OOD selection errors; compare to equal-cost naive scaling.
8. 8. Extensions and implications: heterogeneity in C , privacy/measurement costs, multi-objective rewards, adaptive (multi-round) design; practical recommendations (overlap floors) for RLHF pipelines and audits.

1 Introduction

A recurring failure mode in modern alignment pipelines is that systems can appear well-behaved on the training distribution yet pursue an unintended objective when deployed. In the language of *goal misgeneralization*, the learned policy has internalized a proxy that matches training-time supervision but diverges from the designer’s intent off-distribution. While this phenomenon is often discussed in terms of “inner alignment” or representation learning, we emphasize a complementary and, in practice, more operational lens: goal misgeneralization can arise because the *reward is not identified* from the available preference data. When preference comparisons concentrate in a narrow region of the latent space, many distinct reward functions induce nearly indistinguishable training likelihoods. A downstream optimizer can then select a policy that is optimal for an incorrect but statistically plausible reward, yielding coherent misoptimization rather than random error.

The key mechanism is limited overlap (or limited positivity) in the latent factors that actually determine preferences. Intuitively, preference data are informative about tradeoffs only when the data contain variation along the relevant dimensions. If two reward-relevant attributes are strongly correlated during training—say, helpfulness and harmlessness co-move because the baseline data source rarely exhibits conflict between them—then preference labels cannot reliably reveal how evaluators trade one attribute against the other. In such regimes, the platform can fit a reward model that performs well in-sample while placing essentially arbitrary weight on the “missing” direction. The risk becomes acute precisely when deployment breaks the training correlation structure: the policy optimizer exploits the misestimated tradeoff to move into regions where the proxy reward increases but the true evaluator welfare decreases.

This perspective clarifies why *scaling data* is not always the remedy. If the training distribution places negligible probability mass on the parts of factor space that distinguish competing reward hypotheses, then collecting more i.i.d. comparisons mostly repeats the same information. Statistically, the relevant curvature of the log-likelihood remains small, and the uncertainty in some reward directions decays slowly. In the logistic/BTL setting, this shows up as an ill-conditioned Fisher information matrix, where the smallest eigenvalues correspond to directions of weak excitation in the observed factor differences. In safety terms, “more RLHF” can reduce label noise and improve average-case calibration while leaving worst-case behavioral incentives essentially unconstrained along the dangerous directions.

We argue that this is becoming a practical bottleneck for alignment as of 2026. Frontier systems are increasingly optimized by strong downstream search: large policy classes, tool use, long-horizon planning, and preference-model-based rejection sampling all amplify small reward misspecifications

into large behavioral differences. Meanwhile, the marginal preference comparison is often drawn from a narrow, platform-convenient distribution (e.g. conversational prompts and common response styles), creating a mismatch between what is easy to label and what is safety-critical to identify. As a result, the limiting factor is not simply the number of comparisons n , but the *geometry* of the comparison distribution—which directions of latent variation are present, with what frequency, and at what cost.

This motivates treating data collection as a design problem rather than a passive logging exercise. Concretely, we consider that the platform can purchase targeted interventions: alternative prompting protocols, counterfactual pair generation, controlled perturbations, or curated evaluation tasks that alter the distribution of factor differences δ shown to evaluators. The interventions do not change the underlying preference parameter (the evaluator’s “true” tradeoffs), but they do change which tradeoffs are *revealed* in the data. From an economics viewpoint, this is analogous to choosing experiments to identify demand elasticities; from a learning-theory viewpoint, it is about ensuring sufficient exploration in the covariate space for consistent estimation; from an alignment viewpoint, it is about preventing the optimizer from exploiting unmeasured degrees of freedom in the reward.

A central theme of our framing is that overlap is not an abstract regularity condition: it is a safety-relevant resource that can be traded off against real costs. Interventions are expensive (they require expert labeling, bespoke test construction, or higher evaluator burden), but they can add information exactly where the baseline distribution is degenerate. This creates a concrete optimization frontier: we can either (i) spend heavily on more baseline data that may not reduce uncertainty in the missing directions, or (ii) spend selectively on interventions that change the support and correlations of latent factors, improving identifiability per unit cost. The latter option looks increasingly attractive as models become more capable optimizers: the welfare loss from a single misidentified tradeoff can dominate the average-case gains from marginally better in-distribution fitting.

A second theme is that the relevant notion of “identification” is downstream-dependent. In deployment, we typically do not use the reward model to predict absolute preference probabilities; we use it to choose among a finite set of candidate policies, fine-tuning checkpoints, decoding strategies, or system configurations. Thus, only certain directions in reward space matter: those that change the argmax over policies. If we let $\{\Delta\mu_m\}$ denote the contrasts in latent factor means between the optimal policy and alternatives, then reward uncertainty projected onto the span of $\{\Delta\mu_m\}$ is what drives policy regret. This suggests a design principle that is both theoretically crisp and operationally actionable: allocate intervention budget to increase information in the *high-leverage* directions where policies differ, rather than attempting to uniformly improve estimation in all d latent dimensions.

This downstream lens also sharpens how goal misgeneralization can look

“sudden.” If the best and second-best policies are separated by a small margin under the true objective, then even modest estimation error can flip their ordering. Moreover, as we expand the candidate set Π (more policies, more decoding options, more tools), we introduce more contrast directions that the design must be informative about. In practice, this means that capability progress can *increase* the required overlap: more sophisticated optimizers search harder for reward-model loopholes, and a richer policy class provides more ways to exploit misestimated tradeoffs. Absent intervention, the platform can see a paradoxical effect where training metrics improve while deployment welfare becomes more fragile.

There are immediate governance and verification implications. Overlap conditions can be audited: one can estimate the condition number of empirical information matrices under different evaluation protocols, or track whether certain factor tradeoffs are ever presented to evaluators. Intervention budgets can be justified in terms of worst-case regret targets rather than ad hoc “red teaming.” Regulators and internal safety review can demand evidence that the reward model is not merely accurate on a benchmark, but *identified* relative to the set of deployed policies. This is especially salient when deployment induces correlation shifts (new user populations, new tools, different prompt distributions) that preserve latent sufficiency but alter the joint distribution of factors; under such shifts, non-identified reward directions can become behaviorally dominant.

We also flag a failure mode that our formalism makes easy to miss if one thinks only in terms of average-case prediction error: a reward model can be *confident and wrong* in precisely the directions that matter. In ill-conditioned regimes, the MLE can have small uncertainty along well-excited directions and large uncertainty along the missing ones, yet standard evaluation might overweight the former. Downstream optimization then effectively extrapolates along the latter. This is a structural explanation for why reward hacking can persist even when the reward model seems well-calibrated on held-out comparisons: the held-out set is drawn from the same narrow design, so it cannot detect errors orthogonal to the observed variation.

Finally, we acknowledge limitations and open problems that are important for aligning the theory with deployed systems. First, the latent factorization Z is an idealization: in practice, we may only have proxies (automatic scorers, representation vectors, or hand-designed rubrics), and misspecification can interact with design choices. Second, evaluator objectives are heterogeneous and sometimes strategic; designing interventions per type c raises both statistical and fairness questions. Third, the BTL/logistic model is a stylized likelihood; real preference noise can be context-dependent and non-IID, affecting both identification and cost-effectiveness of interventions. Fourth, adaptive designs (active preference learning) can further reduce cost but introduce incentives, feedback loops, and potential Goodharting of the intervention protocol itself. We view these as opportunities: the core mes-

sage remains that the safety bottleneck is increasingly about *where* we collect preference data, not only *how much* we collect.

In sum, we study goal misgeneralization through the lens of limited overlap and experimental design. The formalism that follows makes precise a safety tradeoff that practitioners already face: when baseline data under-explores the latent dimensions that separate candidate policies, scaling alone may be economically inefficient and safety-insufficient, whereas targeted interventions can restore identifiability and reduce welfare regret at lower total cost.

2 Related work

Our framing sits at the intersection of (i) goal misgeneralization and inner alignment, (ii) causal and econometric treatments of preference learning under limited positivity, (iii) reward hacking and optimizer-induced distribution shift, and (iv) optimal experimental design for discrete choice models. The common thread is that apparent training-time success can coexist with a large set of statistically plausible reward hypotheses, and downstream optimization can systematically select policies that exploit the resulting ambiguity.

Goal misgeneralization and inner alignment. The alignment literature has long emphasized that a learned system may generalize the *wrong* objective even when it matches training feedback [??](#). Much of this discussion is representation-centric: internal mesa-objectives, proxy features, and inductive biases. We instead highlight a complementary failure mode that is closer to identification in statistics and econometrics: even if the model class and latent factorization were “correct,” the *data-generating process* can make the reward effectively non-identifiable in the directions that matter for deployment. This perspective resonates with recent work stressing that deployment often induces distribution shifts along precisely those axes that training-time oversight fails to cover, making downstream optimization brittle [??](#). In our setting, the “suddenness” of misgeneralization emerges from a policy-selection discontinuity: small changes in the learned weights can flip the argmax over a discrete policy set, producing a qualitatively different deployed behavior.

Preference learning, RLHF, and reward modeling as discrete choice. Learning reward models from pairwise preferences is a core ingredient of RLHF and related pipelines [??](#). The Bradley–Terry–Luce (BTL) / logistic formulation we use is standard in discrete choice and has been widely adopted as an idealized model of pairwise judgments. Prior work typically focuses on statistical efficiency, label noise, and downstream RL stability.

Our contribution is not to propose a new likelihood, but to foreground a particular *geometric* bottleneck: the curvature of the log-likelihood is concentrated along the directions of factor differences δ that the data actually explores. When those differences live near a low-dimensional manifold (e.g., because attributes are strongly correlated under the baseline data source), the reward is weakly constrained in orthogonal directions even if the training loss is low. This is precisely the regime where a strong policy optimizer can turn small modeling uncertainty into large welfare loss.

Causal preference learning and positivity/overlap. In causal inference, *positivity* (overlap) ensures that counterfactual effects are identifiable because each treatment has support across confounder strata [?](#). Analogous conditions appear in contextual bandits and offline policy evaluation, where insufficient exploration leads to high-variance or biased estimates. Recent work on causal preference learning similarly emphasizes that preferences observed under a narrow logging policy may not identify how users would respond under alternative options [?](#). Our setting differs in that the “treatment” is not the deployed policy directly, but the *comparison design* that generates pairs (Y, Y') and thus the distribution of δ . Nonetheless, the failure mode is structurally the same: without overlap in the covariate directions that determine the choice probabilities, multiple parameter vectors induce nearly identical observed likelihood. We operationalize this via the Fisher information (or restricted eigenvalues thereof), which plays the role of an overlap certificate tailored to the BTL/logistic model.

Reward hacking, Goodhart effects, and optimizer-induced shift. The observation that optimizing a learned reward can induce pathological behavior is often discussed under reward hacking and Goodhart’s law [??](#). Many papers emphasize misspecification (the reward fails to capture intent) or distribution shift (the policy visits out-of-distribution states). Our account isolates a third ingredient that is especially relevant for preference-based supervision: even when the reward model is *in-class* and the latent factors are sufficient for preferences, the data can leave some tradeoffs underdetermined, enabling “legitimate” exploitation of uncertainty. This connects to analyses of adversarial or worst-case generalization for learned objectives, where the optimizer searches along directions that are weakly supervised [?](#). The safety implication is that standard held-out evaluation drawn from the same baseline design can systematically miss the relevant error modes: the model may generalize well *within* the narrow subspace defined by the training comparisons while remaining unconstrained off that subspace.

Experimental design for discrete choice and logistic models. Our intervention model is closest in spirit to optimal experimental design for gen-

eralized linear models and discrete choice, where the experimenter chooses covariates to maximize information about parameters θ . In the BTL/logistic setting, classical criteria such as D -optimality (maximize $\log \det I$) or A -optimality (minimize $\text{tr}(I^{-1})$) motivate allocating samples to designs that improve identifiability. We adapt this lens to an alignment-driven objective: rather than optimizing generic parameter estimation, we care about *downstream welfare regret* under a finite policy set. This leads to a “task-aware” notion of information: the leverage directions are those spanned by policy contrasts $\Delta\mu_m$, and improving curvature in orthogonal directions may be economically wasteful. In this respect, our approach is closer to decision-theoretic experimental design, where one chooses experiments to reduce expected loss under a downstream decision rule π .

Active preference learning and adaptive data collection. A large literature studies active learning of preferences, including query synthesis for pairwise comparisons and adaptive designs that maximize expected information gain $\mathbb{E}[\text{info}]$. In RLHF practice, variants of active selection (e.g., sampling near uncertainty regions) are increasingly used, though often heuristically. Our paper is deliberately conservative: we first analyze *non-adaptive* mixtures of intervention arms, because this captures many operational interventions (e.g., fixed evaluation suites, curated counterfactual prompts) and is easier to audit. That said, the “high-leverage direction” principle naturally suggests adaptive extensions: if we can estimate which contrast directions are currently under-identified, we can allocate new comparisons to arms that most increase the corresponding restricted eigenvalues. A key open problem is to do so without creating feedback loops where the intervention protocol itself becomes a target for Goodharting or strategic behavior by either models or evaluators.

Distribution shift, robustness, and auditing overlap. Our overlap-based diagnosis complements robust RL and distributional robustness approaches that seek policies with guaranteed performance under shifts π . Those methods typically assume either access to a generative simulator or a specified uncertainty set over dynamics/outcomes. In preference learning for language models, the shift is often in the *joint* distribution of latent attributes, which is hard to specify *ex ante*. We therefore emphasize a more auditable intermediate object: the empirical geometry of the comparison distribution as summarized by information matrices. This connects to governance-oriented proposals to stress-test alignment pipelines using curated evaluations, red-teaming, and measurement of coverage across safety-relevant dimensions \mathcal{S} . Our contribution is to provide a simple theoretical bridge from such audits (“does the data excite the missing directions?”) to a welfare-relevant quantity (policy regret), making the case that overlap is

not only a statistical regularity condition but also a safety control knob.

Positioning and limitations relative to prior work. Compared to the inner-alignment and goal-misgeneralization literatures, our model is intentionally stylized: we assume latent sufficiency and a stable true parameter θ_c^* across interventions, and we treat candidate policies as finite. These assumptions let us isolate the identification–optimization interaction in a form where costs, overlap, and regret can be placed in a single optimization problem. Compared to the experimental design literature, our novelty is primarily in the alignment-motivated objective (worst-case welfare regret over types and policies) and in emphasizing that the “right” design depends on which policies are actually under consideration for deployment. In practice, $\mu(\pi, c)$ must be estimated, latent factors may be only partially observed, and interventions can change evaluator behavior; these issues are active areas where combining causal modeling, robust estimation, and mechanism design seems necessary.

Taken together, prior work suggests a clear lesson: when downstream optimization is strong and the training comparison distribution is narrow, identification failures can masquerade as benign generalization. Our next section formalizes this in a latent-factor BTL model with observed objective types and mixture designs over intervention arms, making precise how overlap enters through Fisher (and restricted) eigenvalues and how optimal intervention spending targets the policy-relevant directions.

3 Model: latent-factor BTL with observed objectives and mixture designs

We study a stylized preference-learning pipeline in which evaluators have heterogeneous objectives, training comparisons are generated under a controllable *design*, and deployment selects among a finite set of candidate policies by optimizing a learned reward. The purpose of the model is to make explicit (i) what it means for comparison data to have “overlap” in the latent factors that determine preferences, and (ii) how interventions can change that overlap without changing the underlying objective being learned.

Objects, objectives, and latent sufficiency. Each training example begins with a prompt X and two candidate responses Y, Y' (e.g., two completions from possibly different generation policies). An evaluator has an objective type $C \in \mathcal{C} = \{1, \dots, G\}$ that is *observed* by the platform (e.g., via user segment, task label, or rater instruction). The evaluator outputs a binary label

$$L = \mathbf{1}\{(X, Y) \succ (X, Y')\},$$

indicating a preference for (X, Y) over (X, Y') .

We posit a latent factor vector $Z \in \mathbb{R}^d$ that summarizes the reward-relevant properties of the prompt-response pair (X, Y) , and analogously Z' for (X, Y') . The dimension d is not meant to be “the true complexity” of language behavior; rather, it is an analytical stand-in for a factorization induced by the model class, an evaluation rubric, or an interpretability layer. The key modeling assumption is *latent sufficiency* for preferences:

$$\Pr(L = 1 \mid X, Y, Y', C = c) = \Pr(L = 1 \mid Z, Z', C = c).$$

Operationally, this says that once we condition on the relevant latent attributes, all remaining dependence on the prompt and surface form is irrelevant for the preference probabilities. This assumption is strong, but it isolates the identification bottleneck we care about: even with the “right” features, the *comparison distribution* may fail to excite some directions of tradeoff.

BTL likelihood with objective-specific weights. Given objective type c , we model preferences via a Bradley–Terry–Luce / logistic link on factor differences. Let

$$\delta := Z - Z' \in \mathbb{R}^d,$$

and let $\theta_c^* \in \mathbb{R}^d$ denote the true (unknown) reward weights for type c . The preference model is

$$\Pr(L = 1 \mid \delta, C = c) = \sigma(\theta_c^{*\top} \delta), \quad \sigma(t) = \frac{1}{1 + e^{-t}}. \quad (1)$$

This specification captures a smooth stochastic choice rule and, crucially for our purposes, yields a likelihood whose curvature is governed by second moments of δ reweighted by the logistic variance term $\sigma(1 - \sigma)$.

We allow θ_c^* to vary across types because (in deployment) different users or evaluation protocols can place different relative value on the same latent factors. Observing C during training means we can estimate a separate reward model per type, which avoids conflating heterogeneity with noise; it also makes the design problem sharper, because overlap may be good for one type but not another.

Downstream policies and welfare. Deployment chooses among a finite set of candidate generation policies $\Pi = \{\pi^1, \dots, \pi^M\}$. For each type c and policy π , let

$$\mu(\pi, c) := \mathbb{E}[Z \mid \pi, C = c]$$

denote the mean latent factor vector induced by rolling out policy π for users of type c (treating μ as known in the clean model). True welfare is linear in these factors:

$$U(\pi, c) = \theta_c^{*\top} \mu(\pi, c).$$

This is the welfare quantity we ultimately care about controlling, because the learning pipeline uses an estimated $\hat{\theta}_c$ to select a deployed policy $\hat{\pi}(c) \in \arg \max_{\pi \in \Pi} \hat{\theta}_c^\top \mu(\pi, c)$. The finiteness of Π is deliberate: it captures common practice (choose among a small set of checkpoints, decoding rules, or safety-tuned variants) and makes the misselection failure mode crisp.

Designs as mixtures of intervention arms. Training data are generated under an experimentation design chosen by the platform. We model this via $J+1$ arms indexed by $j \in \{0, 1, \dots, J\}$, where $j = 0$ is a baseline (no special intervention) and $j \geq 1$ are targeted interventions. An arm should be interpreted broadly: it can change how prompts are sampled, how candidates (Y, Y') are constructed, which generation policies produce candidates, how responses are edited to induce controlled contrasts, or which subset of user contexts is routed to evaluation. What matters is that each arm induces a distribution over latent factor differences conditional on type:

$$\delta \sim P_j(\cdot \mid C = c).$$

We emphasize an invariance assumption that makes the intervention “identifying” rather than “preference-changing”: arms may change *what comparisons are shown*, but they do not change *how type- c evaluators trade off latent factors*. Formally, the same θ_c^* governs (1) across all j .

The platform chooses a mixture over arms. Let $w(c) \in \Delta^J$ be mixture weights for type c (where Δ^J is the simplex over $\{0, \dots, J\}$), and let $n(c)$ be the number of comparisons collected for type c , with $n_j(c) = n(c)w_j(c)$. Each arm has a per-sample cost c_j capturing labeling plus intervention overhead, and allocations are constrained by a budget $\sum_c \sum_j c_j n_j(c) \leq B$.

A convenient stylized generative process for one labeled comparison is:

$$C = c \text{ (observed); } J = j \sim w(c); \quad \delta \sim P_j(\cdot \mid c); \quad L \sim \text{Bernoulli}(\sigma(\theta_c^{*\top} \delta)).$$

We suppress X, Y, Y' in the analysis and treat $P_j(\delta \mid c)$ as the design primitive, since identification in (1) depends on the geometry of δ .

Limited positivity as ill-conditioned information. The central pathology we model is *limited latent positivity*: under the baseline arm $j = 0$, the realized factor differences δ occupy only a narrow region of \mathbb{R}^d , often close to a low-dimensional manifold. Concretely, this can occur when latent attributes are strongly correlated under the baseline data source (e.g., “helpfulness” and “harmlessness” move together in typical comparisons), or when the comparison generator rarely produces candidates that trade one factor off against another.

In logistic models, lack of coverage manifests as a near-singular Fisher information matrix. For a fixed type c and arm j , define

$$I_j(c) := \mathbb{E}_{\delta \sim P_j(\cdot \mid c)} \left[\sigma(\theta_c^{*\top} \delta) (1 - \sigma(\theta_c^{*\top} \delta)) \delta \delta^\top \right]. \quad (2)$$

Under a mixture design $w(c)$, the per-sample information is

$$I(w, c) = \sum_{j=0}^J w_j(c) I_j(c). \quad (3)$$

The logistic variance term $\sigma(1 - \sigma) \leq 1/4$ downweights comparisons that are almost surely decided (very large $|\theta_c^{*\top} \delta|$), reflecting that extremely easy comparisons are less informative about the precise tradeoffs.

Overlap statistics: eigenvalues and restricted eigenvalues. We use the information geometry to define overlap in a way that is tailored to preference learning. A natural global overlap certificate for type c is the minimum eigenvalue

$$\kappa(w, c) := \lambda_{\min}(I(w, c)).$$

When $\kappa(w, c)$ is small, there exists a direction v in parameter space such that $v^\top I(w, c)v$ is small, meaning that the likelihood is flat in that direction and θ_c^* is weakly identified from $n(c)$ samples. In the extreme case $\kappa = 0$, there is a nontrivial direction with zero curvature, corresponding to an unidentifiable tradeoff.

However, our downstream decision does not require uniform accuracy in all directions: policy selection depends on how $\hat{\theta}_c$ projects onto policy-contrast vectors. Let $\pi^*(c) \in \arg \max_\pi U(\pi, c)$ be the true best policy for type c , and define contrast directions

$$\Delta\mu_m(c) := \mu(\pi^*(c), c) - \mu(\pi^m, c), \quad m \neq *.$$

This motivates a *restricted* overlap statistic that measures curvature only on the subspace relevant for distinguishing candidate policies. Let

$$\mathcal{S}(c) := \text{span}\{\Delta\mu_m(c) : m \neq *\},$$

and define

$$\kappa_{\mathcal{S}}(w, c) := \min_{\substack{v \in \mathcal{S}(c) \\ \|v\|_2=1}} v^\top I(w, c)v.$$

In settings where d is large but the candidate policies differ meaningfully only along a small number of axes, $\kappa_{\mathcal{S}}$ can be a more accurate predictor of welfare regret than $\lambda_{\min}(I)$, and it also better captures why “buying overlap” should be targeted rather than uniform.

Interventions as geometric control. Interventions are useful precisely when they increase $\kappa(w, c)$ (or $\kappa_{\mathcal{S}}(w, c)$) per unit cost by changing the distribution of δ . A simple illustrative case is $d = 2$ with baseline δ concentrated along a near-diagonal direction (high correlation between components), which makes $I_0(c)$ ill-conditioned. An intervention that deliberately

constructs comparisons with opposing-factor changes (e.g., high on factor 1 but low on factor 2 versus the reverse) effectively rotates and spreads the support of δ , increasing curvature in the missing direction. Importantly, this can be cheaper than collecting many more baseline samples, because additional baseline data repeats the same near-collinear information.

Discrete latent factors and boundedness. For some applications it is natural to consider $Z \in \{0,1\}^d$ (attributes present/absent) or, more generally, bounded factors. Our analysis accommodates this variant as long as δ is bounded or sub-Gaussian under each $P_j(\cdot | c)$, since concentration of the MLE hinges on controlling tails. The design logic remains the same: arms are valuable when they ensure that $\delta\delta^\top$ has mass in the directions where the downstream decision is sensitive.

What the model sets up. This section defines the objects that connect data collection to downstream welfare: (i) the BTL likelihood (1) with type-specific parameters, (ii) a design space given by mixtures over intervention-induced δ distributions, and (iii) overlap quantified by eigenvalues of the resulting information matrix (3), optionally restricted to policy-relevant subspaces. In the next section we analyze estimation under this model, showing how the error of the logistic MLE scales with $n(c)$ and $\kappa(w, c)$ (or $\kappa_S(w, c)$), and how these estimation bounds translate into welfare regret under discrete policy selection.

4 Estimation: logistic MLE on factor differences and the role of information

We now analyze the estimation step of the pipeline: given preference labels generated from (1) under a mixture design, we fit $\hat{\theta}_c$ for each observed objective type c . The conceptual point is that “more data” is not, by itself, a guarantee of reward accuracy: the geometry of the observed factor differences δ —as summarized by the Fisher information induced by the design—determines which tradeoffs are learnable at a given sample size.

Per-type likelihood and the MLE. Fix an objective type c and consider the subset of comparisons with $C = c$. Writing these as $\{(\delta_i, L_i)\}_{i=1}^{n(c)}$, the (conditional) log-likelihood is

$$\ell_c(\theta) = \sum_{i=1}^{n(c)} \left(L_i \theta^\top \delta_i - \log(1 + e^{\theta^\top \delta_i}) \right), \quad \hat{\theta}_c \in \arg \max_{\theta \in \mathbb{R}^d} \ell_c(\theta). \quad (4)$$

Equivalently, the negative log-likelihood is convex, and the MLE is characterized by the score (first-order) condition

$$\nabla(-\ell_c)(\hat{\theta}_c) = \sum_{i=1}^{n(c)} \delta_i \left(\sigma(\hat{\theta}_c^\top \delta_i) - L_i \right) = 0, \quad (5)$$

matching the equilibrium condition stated earlier. This is just logistic regression with covariates δ_i and labels L_i ; the novelty is that the *distribution* of δ_i is an object we can influence through the experimental design.

Curvature, separability, and why overlap is an estimation issue. The Hessian of the negative log-likelihood is

$$\nabla^2(-\ell_c)(\theta) = \sum_{i=1}^{n(c)} \sigma(\theta^\top \delta_i) (1 - \sigma(\theta^\top \delta_i)) \delta_i \delta_i^\top. \quad (6)$$

Two structural features matter for safety-relevant behavior downstream. First, curvature is inherently *directional*: if the δ_i are nearly collinear, then (6) has a near-null direction, and the likelihood is almost flat along some reward tradeoff. Second, logistic models can exhibit (quasi-)separation: if the observed δ_i allow a hyperplane that (nearly) perfectly predicts L_i , then the MLE can have very large norm and become numerically unstable. In our setting, both phenomena are symptoms of limited latent positivity: the training comparisons do not contain enough “cross-cutting” tradeoffs to pin down θ_c^* .

Interventions are therefore not only about reducing variance constants; they can be necessary to make the problem well-conditioned and to avoid degenerate estimation regimes where the fitted reward is effectively arbitrary off the narrow training manifold.

Fisher information under a mixture design. To connect (6) to a population quantity, recall that under a mixture design $w(c)$, each sample draws $\delta \sim P_j(\cdot | c)$ with $j \sim w(c)$, and then draws $L \sim \text{Bernoulli}(\sigma(\theta_c^{*\top} \delta))$. The corresponding (per-sample) Fisher information at the truth is

$$I(w, c) = \sum_{j=0}^J w_j(c) I_j(c), \quad I_j(c) = \mathbb{E} \left[\sigma(\theta_c^{*\top} \delta) (1 - \sigma(\theta_c^{*\top} \delta)) \delta \delta^\top \mid \delta \sim P_j(\cdot | c) \right],$$

as defined in (2)–(3). Intuitively, $I(w, c)$ is the population analogue of $n(c)^{-1} \nabla^2(-\ell_c)(\theta_c^*)$; it is the object that interventions can improve by changing second moments of δ in the directions that matter.

A subtlety worth flagging is that $I_j(c)$ depends on the unknown θ_c^* through the logistic variance term. In practice, a platform may only be

able to optimize a proxy for information (e.g., second moments of δ alone, or information evaluated at a pilot estimate). This creates an additional exploration-exploitation layer: we want to buy overlap to learn θ_c^* , but we need a guess of θ_c^* to quantify which comparisons are informative.

Asymptotic normality and the role of $\lambda_{\min}(I)$. Under standard regularity conditions for logistic regression (e.g., correct specification, non-separation, and finite second moments), the MLE is consistent and asymptotically normal:

$$\sqrt{n(c)}(\hat{\theta}_c - \theta_c^*) \xrightarrow{d} \mathcal{N}(0, I(w, c)^{-1}). \quad (7)$$

This identifies the first-order dependence of estimation variance on the inverse information geometry. When $\lambda_{\min}(I(w, c))$ is small, there exists a direction in which the asymptotic variance is large, meaning that the learned reward can fluctuate significantly along a tradeoff direction even with large $n(c)$. This is the formal version of the empirical phenomenon that “collecting more of the same kind of preference data” can fail to resolve key value tradeoffs.

Finite-sample concentration via restricted strong convexity. For downstream decision-making, we typically need non-asymptotic guarantees. A convenient route is to use restricted strong convexity (RSC) of the empirical risk. Suppose that for a fixed c and all arms j , δ is bounded or sub-Gaussian; for concreteness, assume $\|\delta\|_2 \leq R$ almost surely.¹ Let

$$\kappa(w, c) = \lambda_{\min}(I(w, c)).$$

Then, with probability at least $1 - \alpha$, for $n(c)$ larger than a constant multiple of $d \log(d/\alpha)$, one can show

$$\|\hat{\theta}_c - \theta_c^*\|_2 \leq C \sqrt{\frac{d \log(1/\alpha)}{n(c) \kappa(w, c)}}, \quad (8)$$

for a constant C depending on R (and mild regularity parameters ensuring existence/uniqueness of the MLE). The key mechanism is that the empirical Hessian in (6) concentrates around its expectation, yielding a lower bound

$$\nabla^2(-\ell_c)(\tilde{\theta}) \succeq n(c) \frac{\kappa(w, c)}{2} I_d$$

uniformly over a neighborhood of θ_c^* (with high probability), which in turn implies that the negative log-likelihood is strongly convex around the truth and the score noise term concentrates.

¹The boundedness assumption can be relaxed to sub-Gaussian tails with slightly more bookkeeping; the qualitative dependence on $n(c)$ and $\kappa(w, c)$ is the same.

From a safety perspective, (8) makes precise why limited overlap is dangerous: if $\kappa(w, c)$ is tiny due to strong correlations in δ , then the sample size required to reduce reward error to an acceptable level grows as $1/\kappa(w, c)$, and may become prohibitive under realistic labeling budgets.

Policy-relevant accuracy and restricted eigenvalues. Uniform ℓ_2 accuracy can be overkill when deployment only chooses among a finite set of policies. As discussed earlier, only projections of $\hat{\theta}_c - \theta_c^*$ onto the policy-contrast span

$$\mathcal{S}(c) = \text{span}\{\Delta\mu_m(c) : m \neq *\}$$

directly influence which policy is selected. This motivates replacing $\kappa(w, c)$ by a restricted overlap statistic,

$$\kappa_{\mathcal{S}}(w, c) = \min_{\substack{v \in \mathcal{S}(c) \\ \|v\|_2=1}} v^\top I(w, c) v,$$

and proving an analogous bound on $\|P_{\mathcal{S}(c)}(\hat{\theta}_c - \theta_c^*)\|_2$ in terms of $\kappa_{\mathcal{S}}(w, c)$. Informally, if interventions increase curvature primarily in $\mathcal{S}(c)$, then we can guarantee policy selection stability without needing to learn every coordinate of θ_c^* precisely. This is one place where the formalism reveals a concrete tradeoff: we can design for welfare among a known set of candidate policies while still leaving the reward under-identified outside the policy-relevant subspace—a potential failure mode if the policy class later expands or if deployment shifts activate previously irrelevant directions.

Multiple objectives and allocation across types. Because C is observed, we estimate $\hat{\theta}_c$ separately for each c . The bounds above apply type-by-type with $n(c)$ and $I(w, c)$ determined by the type-specific allocation $w(c)$. Under a shared budget, this creates an explicit governance-relevant question: which user segments receive intervention-rich comparisons (high κ but higher cost) versus baseline comparisons (low cost but potentially low κ)? The model cleanly separates two issues: statistical feasibility (whether a type has enough overlap to learn) and normative prioritization (which types warrant tighter regret control due to higher stakes or smaller policy margins).

Practicalities, misspecification, and open problems. We have treated δ as observed through an interpretability layer and assumed correct logistic specification. In practice, δ is itself estimated (e.g., via a learned representation), and interventions may shift that representation; moreover, true preferences may be non-logistic or context-dependent even after conditioning on Z . These forms of misspecification can break the clean link between $I(w, c)$ and estimation error. A robust extension would treat the fitted $\hat{\theta}_c$

as a quasi-MLE and bound its error relative to the best-in-class parameter, with design criteria targeting robust curvature under model uncertainty. Another open problem is adaptive design: sequentially choosing arms based on interim $\hat{\theta}_c$ to focus information where it is most valuable, while ensuring exploration to prevent blind spots.

What we carry forward to welfare analysis. The estimation stage delivers a simple scaling law: reward-weight error decreases like $1/\sqrt{n(c)\kappa}$, where κ is an overlap certificate determined by the design mixture. In the next section we connect this estimation error to the discrete argmax over candidate policies, showing how small-but-structured reward errors can induce coherent-but-wrong deployment behavior, especially under shifts that change which policy-contrast directions are activated.

5 Downstream decision and welfare: how reward error becomes coherent policy error

We now turn to the final step of the pipeline: a downstream optimizer uses the learned reward to select a deployed policy from a finite candidate set. This is where the statistical geometry of the training design becomes welfare-relevant. The key distinction is that the downstream system does not “use” $\hat{\theta}_c$ in the abstract; it uses it through an arg max over Π , so small but structured estimation errors can induce discrete policy misselection and therefore large, systematic welfare losses.

Deployment rule and true welfare. Fix an objective type c and recall the per-policy mean factor vector $\mu(\pi, c) = \mathbb{E}[Z \mid \pi, c]$, which we treat as known (or accurately estimated) for the finite candidate set $\Pi = \{\pi^1, \dots, \pi^M\}$. The downstream optimizer deploys

$$\hat{\pi}(c) \in \arg \max_{\pi \in \Pi} \hat{\theta}_c^\top \mu(\pi, c), \quad (9)$$

while the true welfare of a policy is

$$U(\pi, c) = \theta_c^{*\top} \mu(\pi, c). \quad (10)$$

Let $\pi^*(c) \in \arg \max_{\pi \in \Pi} U(\pi, c)$ denote an optimal candidate under the true objective. The per-type welfare regret of deployment is

$$\mathcal{R}(w, n; c) = U(\pi^*(c), c) - \mathbb{E}[U(\hat{\pi}(c), c)], \quad (11)$$

where the expectation is over the training data (and therefore over $\hat{\theta}_c$ and the induced random choice $\hat{\pi}(c)$).

Regret is driven by policy contrasts. Because welfare is linear in θ_c^* , regret under misselection can be expressed in terms of a *policy contrast vector*. For each $m \neq *$, define

$$\Delta\mu_m(c) := \mu(\pi^*(c), c) - \mu(\pi^m, c).$$

If the learned optimizer chooses $\hat{\pi}(c) = \pi^m$, then the realized welfare gap is exactly

$$U(\pi^*(c), c) - U(\pi^m, c) = \theta_c^{*\top} \Delta\mu_m(c). \quad (12)$$

Thus, welfare loss is controlled not by uniform accuracy of $\hat{\theta}_c$ in every coordinate, but by whether $\hat{\theta}_c$ ranks the finite set of inner products $\{\hat{\theta}_c^\top \mu(\pi^m, c)\}_{m=1}^M$ in the same order as $\{\theta_c^{*\top} \mu(\pi^m, c)\}_{m=1}^M$.

A margin condition for correct selection. A standard way to formalize stability of an $\arg \max$ under perturbations is via a *policy margin*. Define

$$\Delta(c) := \min_{m \neq *} \theta_c^{*\top} \Delta\mu_m(c), \quad (13)$$

the smallest true welfare advantage of $\pi^*(c)$ over the remaining candidates. If $\Delta(c)$ is large, then the best policy is robustly optimal; if it is small, then even slight reward error can flip the selection.

To connect this to estimation error, note that for any $m \neq *$,

$$\hat{\theta}_c^\top \mu(\pi^m, c) \geq \hat{\theta}_c^\top \mu(\pi^*(c), c) \Rightarrow (\theta_c^* - \hat{\theta}_c)^\top \Delta\mu_m(c) \geq \theta_c^{*\top} \Delta\mu_m(c). \quad (14)$$

By Cauchy–Schwarz,

$$(\theta_c^* - \hat{\theta}_c)^\top \Delta\mu_m(c) \leq \|\hat{\theta}_c - \theta_c^*\|_2 \|\Delta\mu_m(c)\|_2.$$

Therefore a sufficient condition for *no* policy flip is

$$\|\hat{\theta}_c - \theta_c^*\|_2 < \frac{\Delta(c)}{\max_{m \neq *} \|\Delta\mu_m(c)\|_2}. \quad (15)$$

When (15) holds, we have $\hat{\pi}(c) = \pi^*(c)$ and regret is zero (within the candidate class).

A Lipschitz regret bound under misselection. When (15) fails, we can still bound regret linearly in the estimation error. If $\hat{\pi}(c) = \pi^m$, then using (12),

$$U(\pi^*(c), c) - U(\hat{\pi}(c), c) = \theta_c^{*\top} \Delta\mu_m(c).$$

Add and subtract $\hat{\theta}_c$ and use that $\hat{\pi}(c)$ maximizes $\hat{\theta}_c^\top \mu(\pi, c)$:

$$\begin{aligned} \theta_c^{*\top} \Delta\mu_m(c) &= (\theta_c^* - \hat{\theta}_c)^\top \Delta\mu_m(c) + \hat{\theta}_c^\top \Delta\mu_m(c) \\ &\leq \|\hat{\theta}_c - \theta_c^*\|_2 \|\Delta\mu_m(c)\|_2 + 0, \end{aligned} \quad (16)$$

since $\hat{\theta}_c^\top \mu(\pi^m, c) \geq \hat{\theta}_c^\top \mu(\pi^*(c), c)$ implies $\hat{\theta}_c^\top \Delta\mu_m(c) \leq 0$. Taking the worst case over $m \neq *$ yields the deterministic bound

$$U(\pi^*(c), c) - U(\hat{\pi}(c), c) \leq \left(\max_{m \neq *} \|\Delta\mu_m(c)\|_2 \right) \|\hat{\theta}_c - \theta_c^*\|_2. \quad (17)$$

Combining (17) with an estimation bound such as (8) immediately gives a high-probability regret rate of order

$$\mathcal{R}(w, n; c) \lesssim \max_{m \neq *} \|\Delta\mu_m(c)\|_2 \sqrt{\frac{d}{n(c) \kappa(w, c)}},$$

up to logarithmic factors and problem-dependent constants. This makes the causal chain explicit: design \rightarrow information geometry \rightarrow reward error \rightarrow policy misselection \rightarrow welfare loss.

Policy-relevant subspaces and restricted overlap. The bound (17) is conservative when $\hat{\theta}_c - \theta_c^*$ is large in directions orthogonal to all contrasts. Define the policy-relevant subspace

$$\mathcal{S}(c) := \text{span}\{\Delta\mu_m(c) : m \neq *\},$$

and let $P_{\mathcal{S}(c)}$ denote the orthogonal projector. Since regret only depends on inner products with $\Delta\mu_m(c)$, we can refine (17) to

$$U(\pi^*(c), c) - U(\hat{\pi}(c), c) \leq \left(\max_{m \neq *} \|\Delta\mu_m(c)\|_2 \right) \|P_{\mathcal{S}(c)}(\hat{\theta}_c - \theta_c^*)\|_2. \quad (18)$$

This motivates designing for a restricted eigenvalue

$$\kappa_{\mathcal{S}}(w, c) := \min_{\substack{v \in \mathcal{S}(c) \\ \|v\|_2=1}} v^\top I(w, c) v,$$

rather than for $\lambda_{\min}(I(w, c))$ over all of \mathbb{R}^d . Operationally, we can often achieve large welfare gains by buying overlap in just the span of policy contrasts—but this comes with a safety caveat: if the policy class expands later (new capabilities, new prompts, or new post-training methods) then the relevant contrast directions may change, and previously “irrelevant” reward ambiguity can become decision-critical.

Coherent-but-wrong behavior as geometric exploitation. The phenomenon we ultimately care about is not merely random misclassification among policies, but systematic selection of a policy that is predictably bad under the true objective. In our setting, this happens when (i) the training design makes some direction u weakly identified, so $u^\top I(w, c)u$ is small, and (ii) the candidate set contains policies whose contrasts have substantial projection onto u . Then estimation error can be large in the u direction, and the

optimizer in (9) can select a policy whose apparent advantage comes almost entirely from that poorly identified tradeoff.

This is “coherent” because the downstream optimizer is correctly maximizing the learned reward; it is “wrong” because the learned reward is unconstrained off the training manifold. In fact, if baseline data make δ nearly collinear, then many θ vectors produce nearly identical likelihood; the MLE (or its regularized variant) picks one, and the policy optimizer reliably pushes in the direction that looks best under that arbitrary choice. Interventions matter here because they change which tradeoffs are actually observed during training, shrinking the set of θ that fit the data and preventing the optimizer from exploiting ambiguity.

Deployment shift: when missing directions become activated. A particularly safety-relevant failure mode arises under shifts between training-time and deployment-time factor geometry. Even if Z remains a sufficient latent summary for preferences (so θ_c^* is stable), deployment can change the set of available policies and therefore change the contrast vectors $\Delta\mu_m(c)$. Moreover, the mapping from policy to factor means can shift: a new model version, a new inference-time steering mechanism, or a new content domain can alter $\mu(\pi, c)$ and effectively rotate the policy-relevant subspace $\mathcal{S}(c)$ toward directions that were weakly identified during training.

Under such a shift, a design that was adequate for the original $\mathcal{S}(c)$ can fail catastrophically: $\kappa_{\mathcal{S}}(w, c)$ for the *new* contrast span can be small, so the learned reward extrapolates in precisely the directions the optimizer now uses. This formalizes a common governance concern: passing offline preference benchmarks does not certify safe optimization under distribution shift if the benchmark does not cover the high-leverage tradeoffs induced by future deployments.

Implications for auditing and conservative deployment. The above suggests two complementary mitigations. First, we can audit overlap geometrically: given an estimated information matrix $\hat{I}(w, c)$ (or even just empirical second moments of δ), we can evaluate whether the design is informative along currently relevant contrast directions and along plausible future directions (stress tests). Second, we can make deployment conservative when margins are small. For example, if the estimated policy margin

$$\hat{\Delta}(c) := \hat{\theta}_c^\top \mu(\hat{\pi}(c), c) - \max_{\pi \neq \hat{\pi}(c)} \hat{\theta}_c^\top \mu(\pi, c)$$

is tiny relative to an uncertainty radius for $P_{\mathcal{S}(c)}(\hat{\theta}_c - \theta_c^*)$, then the system can defer, randomize, or choose a robust alternative that maximizes worst-case utility over a confidence set for θ_c^* . We treat such gating and robust optimization mechanisms as extensions, but the core point remains: the

welfare impact of preference learning is mediated by discrete downstream choices, so we must reason about information geometry and policy margins together, not in isolation.

6 Optimal intervention allocation: cost-minimizing designs for a regret target

Having related welfare regret to the information geometry induced by the training design, we can now ask the natural systems question: given a menu of intervention arms with known per-sample costs, how should we allocate our finite budget to achieve a specified regret guarantee at minimum cost? The answer is an optimal experimental design problem, but with a deployment-driven objective: we do not seek “uniformly good” estimation of θ_c^* ; we seek to buy information in the few directions that can flip the downstream arg max over candidate policies.

From regret targets to information constraints. Fix an objective type c and suppress c in notation. Let the intervention mixture be $w \in \Delta^J$ over arms $j \in \{0, \dots, J\}$, with per-sample costs $c_j > 0$ and arm-specific Fisher information matrices $\{I_j\}_{j=0}^J$. Under a mixture w , the per-sample information is

$$I(w) = \sum_{j=0}^J w_j I_j,$$

and with n total comparisons the information scales as $nI(w)$. The key observation is that many regret proxies can be expressed as monotone functions of the quadratic forms

$$\Delta\mu_m^\top I(w) \Delta\mu_m, \quad m \neq *,$$

where $\Delta\mu_m = \mu(\pi^*) - \mu(\pi^m)$ are the policy contrast directions. Intuitively, if we only become very certain about θ along those directions, the policy ranking becomes stable even if θ remains ambiguous elsewhere.

One convenient large-sample proxy comes from approximating the MLE as Gaussian,

$$\hat{\theta} \approx \mathcal{N}\left(\theta^*, (nI(w))^{-1}\right),$$

and then bounding the probability that a suboptimal policy π^m overtakes π^* under $\hat{\theta}$. A standard large-deviation calculation yields a proxy of the form

$$\mathcal{R}(w, n) \lesssim \sum_{m \neq *} (\theta^{*\top} \Delta\mu_m) \exp\left\{-\frac{n}{2} \Delta\mu_m^\top I(w) \Delta\mu_m\right\}, \quad (19)$$

up to constants that depend on the logistic curvature term and boundedness of δ . While (19) is not a theorem as stated, it captures the correct design

dependence: regret decays exponentially in n times an information quantity computed *only* along the contrast directions.

Alternatively, if we use the more conservative Lipschitz-style bound $\mathcal{R} \lesssim \max_m \|\Delta\mu_m\|_2 \|\hat{\theta} - \theta^*\|_2$ together with a concentration bound $\|\hat{\theta} - \theta^*\|_2 \lesssim \sqrt{d/(n\kappa)}$, we obtain the simpler sufficient condition

$$n \kappa_{\mathcal{S}}(w) \gtrsim \frac{d}{R_{\text{target}}^2}, \quad \kappa_{\mathcal{S}}(w) := \min_{\substack{v \in \mathcal{S} \\ \|v\|_2=1}} v^\top I(w) v, \quad (20)$$

where $\mathcal{S} = \text{span}\{\Delta\mu_m : m \neq *\}$. This formulation makes explicit the safety tradeoff: if we under-invest in overlap in \mathcal{S} , the optimizer can reliably exploit the resulting ambiguity.

A convex program for cost-minimizing designs. Let $\bar{c}(w) := \sum_{j=0}^J c_j w_j$ denote the expected per-sample cost under mixture w . The total expected data-collection cost is $n \bar{c}(w)$. A regret target can be enforced either via the proxy (19) (exponential constraints) or via the sufficient eigenvalue condition (20). Both yield tractable optimization problems.

A particularly transparent convex formulation comes from enforcing per-contrast information lower bounds. Choose a required information level $\tau > 0$ (which can be backed out from R_{target} via either (19) or (20)) and solve

$$\begin{aligned} \min_{w \in \Delta^J, n \geq 0} \quad & n \bar{c}(w) \\ \text{s.t.} \quad & n \Delta\mu_m^\top I(w) \Delta\mu_m \geq \tau, \quad \forall m \neq *. \end{aligned} \quad (21)$$

Because $\Delta\mu_m^\top I(w) \Delta\mu_m = \sum_j w_j q_{mj}$ with $q_{mj} := \Delta\mu_m^\top I_j \Delta\mu_m$, the constraints in (21) are linear in w once n is fixed (and jointly convex in (w, n) after standard transformations). In fact, eliminating n gives an equivalent fractional program

$$\min_{w \in \Delta^J} \frac{\bar{c}(w)}{\min_{m \neq *} \sum_{j=0}^J w_j q_{mj}}, \quad (22)$$

and then setting $n = \tau / \min_{m \neq *} \sum_j w_j q_{mj}$. Introducing an auxiliary variable t for the minimum, we obtain a convex (indeed, linear) reformulation:

$$\begin{aligned} \max_{w \in \Delta^J, t \geq 0} \quad & \frac{t}{\bar{c}(w)} \\ \text{s.t.} \quad & \sum_{j=0}^J w_j q_{mj} \geq t, \quad \forall m \neq *. \end{aligned} \quad (23)$$

This is a design principle we can interpret operationally: we buy a mixture w that maximizes the *worst-case* contrast information per unit cost.

When we instead want to target the restricted eigenvalue in (20) directly, we can work with a semidefinite program over the contrast span. Let $S \in \mathbb{R}^{d \times r}$ be an orthonormal basis for \mathcal{S} (with $r = \dim(\mathcal{S}) \leq M - 1$). Then

$$\kappa_{\mathcal{S}}(w) = \lambda_{\min}(S^\top I(w) S).$$

Maximizing $\kappa_{\mathcal{S}}(w)$ per unit cost is equivalent to

$$\begin{aligned} \max_{w \in \Delta^J, \kappa \geq 0} \quad & \kappa \\ \text{s.t.} \quad & S^\top \left(\sum_{j=0}^J w_j I_j \right) S \succeq \kappa I_r, \\ & \bar{c}(w) \leq 1. \end{aligned} \tag{24}$$

The constraint is linear matrix inequality (LMI), so (24) is convex. This formulation is attractive when M is large and we prefer to summarize policy relevance by a subspace rather than enumerate all contrasts.

Closed-form in the simplest case: $d = 2$ and a single contrast. The design problem admits a particularly clean characterization when there is only one decision-critical direction. Concretely, suppose $d = 2$ and $M = 2$, so there is a unique contrast vector $\Delta\mu \in \mathbb{R}^2$ (up to scaling). Then the relevant information under arm j is the scalar

$$q_j := \Delta\mu^\top I_j \Delta\mu,$$

and under mixture w it is $\sum_j w_j q_j$. In this one-dimensional reduction, (22) becomes

$$\min_{w \in \Delta^J} \frac{\sum_j c_j w_j}{\sum_j q_j w_j}.$$

Because the objective is a ratio of two linear functions over a simplex, an optimum is achieved at an extreme point: we place all mass on a single arm

$$j^* \in \arg \max_{j \in \{0, \dots, J\}} \frac{q_j}{c_j} = \arg \max_j \frac{\Delta\mu^\top I_j \Delta\mu}{c_j}. \tag{25}$$

The minimum-cost way to reach a target information τ is then to sample exclusively from j^* with

$$n = \frac{\tau}{q_{j^*}}, \quad \text{total cost} = \frac{\tau c_{j^*}}{q_{j^*}}.$$

This clarifies the logic behind targeted interventions: if the baseline arm has poor overlap in the contrast direction (small q_0 because δ is nearly collinear), then q_0/c_0 can be dominated by an intervention that changes the geometry even if it is more expensive per sample.

Two contrast directions in $d = 2$: when mixing becomes necessary.

The next nontrivial regime is still $d = 2$ but with $M \geq 3$ such that \mathcal{S} is two-dimensional and at least two contrasts matter. In that case, concentrating on a single arm may be suboptimal because an arm can be informative along one contrast while remaining weak along another. In the linear-constraint form (23), we must satisfy multiple inequalities $\sum_j w_j q_{mj} \geq t$ simultaneously; the optimum may require mixing arms to balance the weakest contrast. Geometrically, we are seeking a mixture that makes all relevant quadratic forms large enough, and the optimum often lies at a mixture of at most two arms (by standard results on linear programs over simplices), though which two depends on the full matrix of q_{mj} and costs.

This regime is also where safety intuitions become sharper: if we optimize only for a single “headline” tradeoff, we may leave other policy contrasts under-identified, enabling coherent-but-wrong selection among a richer candidate set. In practice, expanding Π (new model variants, new steering knobs) tends to increase the number of active contrasts, pushing us away from single-arm closed forms and toward systematic numerical optimization.

When we need numerical methods (and what can go wrong). Outside these low-dimensional special cases, two factors force us into computation rather than closed form. First, for $d > 2$ the restricted eigenvalue objective in (24) is genuinely matrix-valued, and designs that look good along each individual $\Delta\mu_m$ can still yield a poorly conditioned $S^\top I(w)S$ (e.g., information concentrated in a narrow cone within \mathcal{S}). Second, for large M the set of contrasts can be large or even time-varying as the product evolves; it is then more stable to solve subspace SDPs or to optimize a smooth surrogate such as $\log \det(S^\top I(w)S)$ (a D -optimal criterion on the policy-relevant subspace), which remains convex in w .

There is also an estimation layer we cannot ignore: I_j and even $\Delta\mu_m$ are rarely known exactly. Empirically, we must estimate q_{mj} from pilot data or from intervention metadata, and then solve a *robust* variant of (21) that accounts for uncertainty (e.g., q_{mj} lying in confidence intervals). This is more than a statistical nicety: if we overestimate an intervention’s informativeness, we may under-collect data in exactly the missing direction, leading to overconfident policy optimization. Robust design is convex in many practical uncertainty sets (boxes or ellipsoids), but it increases cost; the governance-relevant question is how much robustness margin we should require before deploying optimization that can amplify any remaining misspecification.

Finally, when we scale to multiple objective types $c \in \mathcal{C}$, the platform faces a coupled allocation problem across types under a shared budget. The same convex structure persists (we solve (21) per type and then allocate $n(c)$ across c), but the safety stakes rise: types with smaller margins or more severe downside risk may warrant disproportionate intervention spend.

This makes explicit a tension that often remains implicit in practice: cost-efficient preference learning is not just about collecting more labels, but about deciding *which counterfactual tradeoffs* we are willing to pay to observe so that downstream optimization remains well-grounded.

6.1 Empirical demonstration: controllable interventions via attribute editing

We now outline an empirical demonstration that mirrors the design logic above while staying close to what an RLHF pipeline can actually implement today: we take an open preference dataset, construct several *controllable intervention arms* by editing responses to manipulate specific attributes, and then measure (i) the induced overlap/conditioning in the latent difference distribution, (ii) the resulting estimation variance of a BTL reward model, and (iii) downstream *out-of-distribution* (OOD) policy-selection errors when a learned reward is used to choose among a finite menu of candidate policies. The goal is not to claim that attribute editing perfectly instantiates a causal intervention on ground-truth latent factors, but to test the concrete prediction of our formalism: when baseline data exhibits strong factor correlations (limited positivity), *equal-cost* naive scaling yields much smaller gains than targeted interventions that improve information geometry in policy-relevant directions.

Dataset and factorization. We begin with a public pairwise preference dataset such as Anthropic `hh-rlhf` or Stanford SHP, which provides tuples (X, Y, Y', L) where L indicates which response is preferred. To connect to the latent-factor model, we require a map from prompt-response pairs to factor vectors $Z \in \mathbb{R}^d$. In practice we implement this in one of two ways. First, we can use a small set of interpretable, evaluator-facing attributes (e.g., helpfulness, harmlessness, verbosity, refusal style, factuality) and train lightweight attribute predictors on a small annotated subset; Z is then the vector of predicted attribute scores. Second, we can use a representation-based factorization (e.g., principal components of a frozen encoder embedding of (X, Y)) and treat the resulting coordinates as latent factors. The first approach offers clearer safety interpretation (we know what direction we are buying overlap in), while the second reduces reliance on potentially noisy attribute labels; we recommend running both as a robustness check.

Given Z and Z' we compute $\delta = Z - Z'$ for each comparison. Baseline limited overlap typically shows up immediately: some components of δ have very low variance, and more importantly, δ concentrates near a low-dimensional subspace due to correlations (e.g., responses that are more helpful are also more verbose, or refusals are also more harmless). This is the empirical analogue of an ill-conditioned Fisher information matrix.

Constructing intervention arms via controlled edits. We define arms $j \in \{0, 1, \dots, J\}$ where $j = 0$ is the unmodified dataset distribution and $j > 0$ are *attribute-edited* distributions. Each intervention sample is constructed by taking an existing (X, Y) and producing an edited response \tilde{Y} that aims to shift one attribute while minimally perturbing others. Concretely, we implement edits using a separate instruction-following model (an “editor”) with prompts such as: “Rewrite the response to be *equally helpful* but *more concise*” or “Rewrite to preserve content but change tone to be more polite.” This yields counterfactual pairs (Y, \tilde{Y}) (or (\tilde{Y}, Y')) which we then send to the same preference labeling process as the baseline (or, in an offline study, we use the dataset’s labels only for baseline and collect new labels for edited pairs).

We operationalize several arm templates that are designed to change the geometry of δ : (i) *single-attribute toggles* (increase/decrease one attribute while constraining others), intended to add mass along a coordinate direction; (ii) *decorrelation edits* (e.g., “increase helpfulness but keep verbosity fixed”), intended to break observed baseline correlations; (iii) *contrast-amplifying edits* targeted to estimated policy contrasts $\Delta\mu_m$ (e.g., edit responses specifically along the direction that distinguishes two candidate policies). Each arm has an estimated per-sample cost c_j equal to labeling cost plus editing overhead (editor inference, filtering, and quality control).

Because edits can fail (changing multiple attributes or degrading coherence), we include automated and manual filters: we reject edited responses that violate basic constraints (toxicity, nonsensicality) and we estimate realized attribute shifts $\widehat{\Delta Z} = Z(X, \tilde{Y}) - Z(X, Y)$ to verify that an arm actually moves the intended factor distribution. Importantly, these checks also serve as a safety diagnostic: interventions that systematically induce unintended attribute changes are precisely the ones that can mislead downstream optimization.

Measuring induced overlap and information geometry. For each arm j , we estimate an empirical Fisher information matrix using the logistic curvature weights. Given a fitted parameter $\bar{\theta}$ (e.g., from a pilot model trained on a small balanced mixture), define

$$\widehat{I}_j := \frac{1}{n_j} \sum_{i \in \mathcal{D}_j} \sigma(\bar{\theta}^\top \delta_i) (1 - \sigma(\bar{\theta}^\top \delta_i)) \delta_i \delta_i^\top,$$

and for any mixture w , $\widehat{I}(w) = \sum_j w_j \widehat{I}_j$. We then report (a) $\lambda_{\min}(\widehat{I}(w))$ as a global conditioning proxy and (b) the policy-relevant restricted eigenvalue $\widehat{\kappa}_{\mathcal{S}}(w) = \lambda_{\min}(S^\top \widehat{I}(w) S)$ for S spanning $\mathcal{S} = \text{span}\{\Delta\mu_m\}$. The core empirical prediction is that certain edits dramatically increase $\widehat{\kappa}_{\mathcal{S}}(w)$ at modest additional cost, especially in regimes where the baseline \widehat{I}_0 is nearly rank-deficient.

To make this diagnostic interpretable to practitioners, we also visualize the projected δ clouds onto the top two eigenvectors of \widehat{I}_0 and onto key contrast directions $\Delta\mu_m$, showing directly whether interventions populate previously empty regions. This “overlap plot” is the empirical analogue of checking positivity in causal inference, and it can be audited without trusting the full reward model.

Estimation variance under equal cost. Next, we compare reward-model estimation quality under (i) baseline-only scaling and (ii) mixtures that include interventions, holding total expected cost fixed. For each design we train a BTL/logistic reward model $\hat{\theta}$ and estimate uncertainty via the plug-in covariance $(n\widehat{I}(w))^{-1}$ and via nonparametric bootstrap over comparisons. We then evaluate two quantities: the global parameter error $\|\hat{\theta} - \hat{\theta}_{\text{ref}}\|_2$ relative to a high-data reference fit $\hat{\theta}_{\text{ref}}$, and, more importantly, the *contrast-direction error*

$$\max_{m \neq *} \frac{|(\hat{\theta} - \hat{\theta}_{\text{ref}})^\top \Delta\mu_m|}{\|\Delta\mu_m\|_2},$$

which directly proxies the probability of flipping the arg max over policies. The predicted pattern is that intervention mixtures reduce contrast-direction variance much faster than baseline scaling, even if the overall ℓ_2 error improves only modestly.

Downstream OOD policy-selection errors. To test the deployment-relevant failure mode (coherent-but-wrong optimization), we instantiate a finite candidate set $\Pi = \{\pi^1, \dots, \pi^M\}$ using either (a) different decoding/steering settings of a fixed generator (temperature, system prompt, refusal threshold) or (b) different checkpoints/variants. For each π^m we estimate $\mu(\pi^m, c)$ by rolling out on a prompt set and computing average factors Z . We then select $\hat{\pi} = \arg \max_m \hat{\theta}^\top \mu(\pi^m, c)$.

We evaluate regret and misselection on two test distributions: an in-distribution (ID) holdout from the dataset and an OOD shift constructed to change factor correlations while keeping attributes meaningful (e.g., prompts requiring terse factual answers; prompts eliciting refusals; or a filtered subset where verbosity and helpfulness decouple). Because true welfare $U(\pi, c) = \theta^* \top \mu(\pi, c)$ is not directly observable, we approximate it using a held-out panel of preference labels between policy rollouts (treating the panel aggregate as ground truth for evaluation) or using an expensive, high-quality evaluator model calibrated on human data. We then report: (i) the frequency with which $\hat{\pi}$ differs from the best policy under the evaluation labels, and (ii) the realized welfare gap. The key comparison is whether intervention-designed training reduces OOD misselection at the same cost.

Equal-cost baselines and what would falsify the story. The central comparison is a cost curve: for a grid of budgets B , compare baseline-only training (all mass on $j = 0$) against a design that allocates some weight to targeted edits. The formalism predicts a regime of diminishing returns for baseline scaling when $\hat{\kappa}_{\mathcal{S}}(w)$ is small: additional baseline samples mostly reduce uncertainty in already-identified directions. Conversely, if interventions do not improve $\hat{\kappa}_{\mathcal{S}}$, we should not expect downstream gains; indeed, a clean falsification is that edited arms fail to increase restricted eigenvalues yet still reduce OOD misselection, which would suggest our overlap-based proxy is missing the relevant mechanism. Another falsification is that interventions increase overlap metrics but worsen OOD behavior, indicating that editing changes the effective objective (violating the assumption that θ^* is invariant) or introduces systematic label noise.

Safety and governance takeaways from the empirical protocol. Even as a research prototype, this demonstration produces auditable artifacts that map cleanly onto governance needs: a documented menu of intervention arms, their realized attribute shifts, per-arm costs, and measured overlap floors $\hat{\kappa}_{\mathcal{S}}$. The main safety implication is that we can turn a vague requirement (“collect diverse preference data”) into a measurable control target (“achieve a minimum policy-relevant overlap floor at a stated confidence”), and we can check whether a proposed intervention genuinely improves identifiability rather than merely adding more labels. At the same time, the experiment highlights an open problem: attribute editing is only a *proxy* intervention, and ensuring invariance of the underlying preference parameter is nontrivial. In practice, the same tools used for overlap improvement must be paired with invariance tests (e.g., checking for systematic shifts in pairwise preferences on anchor comparisons) before we allow the downstream optimizer to treat the learned reward as authoritative.

6.2 Extensions and implications: heterogeneity, costs, multi-objective rewards, and adaptive design

The clean story above treated a single objective type c with a fixed factorization Z and a one-shot (non-adaptive) design mixture w . In practice, RLHF systems deviate from each of these assumptions in ways that matter for safety: evaluators are heterogeneous, some measurements are expensive or privacy-sensitive, reward is intrinsically multi-objective, and platforms can (and often should) run multi-round data collection. Our main claim in this section is that these complications do not invalidate the overlap-first perspective; rather, they sharpen it. When we account for realistic constraints, the object we want to *control* becomes a family of policy-relevant overlap floors, together with explicit tests for invariance and robustness.

Heterogeneity in evaluator objectives C . If we observe objective labels c (e.g., annotator group, jurisdiction, or user segment), the natural extension is to treat each type as its own BTL problem with parameter θ_c^* and design mixture $w(c)$. The platform then faces an allocation problem across types under a shared budget:

$$\min_{\{n_j(c)\}} \sum_{c \in \mathcal{C}} \sum_{j=0}^J c_j n_j(c) \quad \text{s.t.} \quad \sup_{c \in \mathcal{C}} \mathcal{R}(w, n; c) \leq R_{\text{target}}.$$

Two deployment-relevant regimes are worth separating. In an *average-welfare* regime, the platform cares about $\sum_c p(c) \mathbb{E}[U(\hat{\pi}(c), c)]$ and will rationally spend less on rare types. In a *worst-case* or *rights-based* regime (common in safety and governance discussions), we instead enforce the constraint for every c (or every protected subset of types). The overlap framing makes the tradeoff explicit: if some types have smaller margins $\Delta(c)$ or more severe baseline correlations (smaller $\kappa(w, c)$), they dominate the required $n(c)\kappa(w, c)$ and thus dominate cost. This is not merely an estimation issue; it is a fairness and safety issue, because the downstream optimizer can be reliably correct for the majority while being systematically wrong for a minority type.

If c is *unobserved* (or only partially observed), we can still use the same geometry but at a price: the effective design must ensure overlap for a mixture distribution, and the reward model becomes either a single pooled θ^* (which is misspecified under heterogeneity) or a latent-mixture model. The failure mode here is familiar: pooling can yield a reward that is accurate on average but wrong in precisely the regions of Z that distinguish types. A pragmatic mitigation is to define a conservative policy-relevant subspace \mathcal{S} that includes contrasts across *both* candidate policies and suspected type differences, and require a minimum restricted eigenvalue on that enlarged subspace. This turns “unknown heterogeneity” into an auditable coverage requirement rather than a vague concern.

Privacy and measurement costs as design constraints. Our empirical protocol implicitly assumed that we can compute Z (or a proxy) cheaply and store it for overlap diagnostics. In deployed systems, however, factor measurement may be privacy-sensitive (e.g., attributes inferred from user data), or expensive (e.g., high-quality model-based evaluators used to score factuality). There are two distinct costs: (i) the per-sample intervention/labeling cost c_j that already appears in the budget, and (ii) a *measurement cost* for producing the factorization itself (or for producing reliable proxies used in design and auditing).

One extension is to treat the factor pipeline as another arm choice: for each comparison we may choose a measurement mode m (cheap proxy vs. expensive audit-grade scoring), which changes both the observed covariates

$\tilde{\delta}$ and the implied information matrix. If measurement noise satisfies a classical errors-in-variables model, then naively plugging in $\tilde{\delta}$ can underestimate uncertainty and inflate $\hat{\kappa}$, creating a dangerous false sense of overlap. The design implication is that overlap floors must be certified with respect to a measurement process whose error is bounded, ideally by occasionally paying for “gold” measurements to calibrate proxies. The governance implication is that an overlap audit should report not only $\hat{\kappa}_S$ but also the measurement procedure and its calibration error bars.

Privacy adds another layer. If we must use differentially private (DP) estimates of Z or of $\hat{I}(w)$, then the overlap floor itself becomes a random quantity with privacy-induced noise. This does not make overlap auditing impossible, but it shifts the question: we should require that the lower confidence bound on $\kappa_S(w)$ (accounting for both sampling and DP noise) exceeds a threshold. Put differently, DP does not eliminate the need for overlap; it forces us to budget for a larger safety margin.

Multi-objective reward models and policy menus. Many RLHF deployments are intrinsically multi-objective: helpfulness, harmlessness, truthfulness, style, and user satisfaction are not cleanly reducible to a single scalar without normative choices. In our notation, this corresponds to either (i) multiple types c with different θ_c^* (a normative pluralism view), or (ii) a single deployment objective that is a function of several latent components (a constrained optimization view). Both can be handled by expanding the policy-relevant subspace.

For example, suppose deployment selects a policy subject to safety constraints, such as maximizing helpfulness while keeping expected harm below a bound. If we model constraints as additional linear functionals of $\mu(\pi)$, then the relevant directions are not only the unconstrained contrasts $\Delta\mu_m$, but also the gradients of the active constraints. The design target becomes a restricted eigenvalue over the span of those directions. Similarly, if we deploy via a scalarization that varies across contexts (effectively randomizing over θ), then worst-case regret involves a supremum over a set of plausible θ vectors, and the design must provide overlap in a larger cone rather than a single direction. The core safety message is that “more objectives” typically means “more high-leverage directions”; without interventions, baseline correlations can silently delete some of them.

Adaptive (multi-round) intervention design. A one-shot mixture w is attractive analytically, but platforms can often run multi-round collection: gather a pilot dataset, fit $\hat{\theta}$, estimate \hat{I}_j , then allocate the remaining budget toward the most informative arms. This is essentially an experimental design or bandit problem with a safety twist: exploration is costly and may expose evaluators to unusual outputs, yet insufficient exploration risks coherent-but-

wrong optimization.

A simple adaptive template is:

$$w^{(t+1)} \in \arg \max_{w \in \Delta^J} \min_{v \in \mathcal{V}} v^\top \widehat{I}^{(t)}(w) v - \lambda \sum_j c_j w_j,$$

where \mathcal{V} is a set of unit vectors spanning the policy-relevant subspace (e.g., normalized $\Delta\mu_m$ and any constraint directions), and $\widehat{I}^{(t)}(w) = \sum_j w_j \widehat{I}_j^{(t)}$ is the plug-in information estimate after round t . One can add optimism bonuses (upper confidence bounds on \widehat{I}_j) to avoid prematurely collapsing onto a seemingly good arm when estimates are noisy. The technical open problem is to obtain regret guarantees that compose (estimation error \rightarrow policy error) under adaptive data collection, since classical MLE asymptotics can fail under heavy adaptivity. The practical recommendation is more modest: adaptivity should be used to *target missing directions* identified by overlap diagnostics, and the resulting design should be re-audited after each round to ensure that the realized δ distribution actually improved conditioning.

Practical recommendations: overlap floors as pipeline controls and audit artifacts. We can distill the above into implementable controls for RLHF pipelines.

First, define a policy-relevant subspace \mathcal{S} before large-scale training. Concretely, compute (or bound) $\Delta\mu_m$ for the finite policy menu under consideration (including foreseeable variants such as different refusal thresholds or decoding settings), and set $\mathcal{S} = \text{span}\{\Delta\mu_m\}_{m \neq *}$ (optionally augmented with constraint directions and suspected heterogeneity directions).

Second, enforce an *overlap floor* requirement:

$$\kappa_{\mathcal{S}}(w, c) := \lambda_{\min}(S^\top I(w, c) S) \geq \underline{\kappa},$$

where S is an orthonormal basis for \mathcal{S} . In practice we certify this with a lower confidence bound on $\widehat{\kappa}_{\mathcal{S}}$ using bootstrap (and, if applicable, DP noise accounting). The threshold $\underline{\kappa}$ should be set by a welfare regret target via the bound in Proposition 2, i.e., by requiring $n(c) \underline{\kappa}$ large enough given the observed policy margins.

Third, couple overlap improvements with *invariance checks*. Any intervention that changes the effective preference parameter (violating the assumption that θ^* is invariant across arms) can create a misleadingly well-conditioned design that optimizes the wrong thing. A lightweight invariance test is to include a fixed set of “anchor comparisons” drawn from baseline and periodically re-label them; systematic shifts in predicted vs. observed preferences by arm are a red flag. This is also the right place to incorporate governance requirements: an auditor can demand documented arm

definitions, realized attribute shifts, anchor-comparison stability, and certified overlap floors, rather than relying on informal claims of “data diversity.”

Finally, adopt a conservative deployment rule: if the certified overlap floor on \mathcal{S} is not met for a protected type c , then either (i) defer deployment for that type, (ii) restrict the policy menu to a smaller set with larger margins, or (iii) allocate additional budget to targeted interventions. This makes the safety tradeoff explicit: we either pay to restore identifiability where it matters, or we constrain optimization to avoid confident errors.

These extensions emphasize a broader point. In realistic RLHF systems, the main risk is not that we lack data in aggregate, but that we lack *geometric coverage* in the directions that determine downstream decisions. Interventions, measurement choices, and adaptive design are all levers for controlling that coverage; overlap floors are the mechanism by which we can make those levers auditable and enforceable.