Optimal Overlap Design for Preference Learning: Quadrant-Filling Interventions and Condition-Number Sample Complexity

Liz Lemma Future Detective

January 22, 2026

Abstract

Modern preference learning pipelines (including DPO-style optimization) often fail out of distribution because the observational data exhibits limited overlap: latent response attributes such as length, format, and safety tone are strongly correlated with other quality-relevant factors. Building on the causal perspective on preference learning (confounding, overlap/positivity) and the observation that preference models only identify reward differences up to nuisance shifts, we formalize preference data collection as an optimal experimental design problem in a latent factor space. We introduce a clean latent-linear BTL model where interventions act as controllable transformations of responses that shift the distribution of latent attributes. We define an overlap condition number via the Fisher information of the BTL likelihood and show that (i) estimation error and downstream out-of-distribution regret scale with the inverse overlap, and (ii) targeted "quadrant-filling" intervention mixtures maximize overlap and sharply reduce required labels relative to passive logging when latent attributes are highly correlated. We give closed-form characterizations in a two-factor case and propose a practical bandit-style design algorithm that uses uncertainty over latent factors to adaptively choose interventions under a labeling budget. Empirically, we outline tests showing that actively decorrelating (length, quality) and (helpful, harmless) attributes improves OOD win rates and reduces reward hacking compared to standard preference data collection.

Table of Contents

causal preference learning (confounding, latent overlap); active preference learning; experiment design connections.

3. 3. A tractable latent BTL model with interventions: primitives, mapping text→latent factors, linear latent rewards; where positivity/overlap enters.

4. 4. Overlap metrics and why they matter: Fisher information, overlap condition number, and how it controls variance and OOD error; connection to limited latent positivity examples.

5. 5. The optimal overlap design problem: constrained optimization over intervention policies q(a|x); objective variants (A-optimal, E-optimal, minimax over environments); when closed form exists.

6. 6. Closed-form results in the 2-factor correlated-Gaussian case: passive vs active label complexity; quadrant-filling mixture structure; comparative statics in correlation ρ, noise, and budget.

7. 7. Algorithms for the general case: plug-in Fisher estimates, uncertainty-aware bandit design over interventions, feasibility constraints (controllable generation), and practical approximations.

8. 8. Empirical blueprint: (i) synthetic latent-factor stress tests (correlation flips), (ii) real preference datasets with controlled edits (length/format/safety), (iii) downstream policy learning (DPO) under OOD shifts.

9. 9. Discussion and 2026 implications: data governance, benchmark design, audit requirements, and how to operationalize overlap targets in production; limitations and extensions.

1 Related work

Our formulation sits at the intersection of modern preference optimization for language models, causal perspectives on preference data, and classical experimental design. We emphasize these connections because they clarify which parts of the alignment pipeline are "optimization" problems (improving a model given data) versus "identification" problems (acquiring data that makes the latent objective learnable in the first place), and because many observed failure modes in practice can be reinterpreted as overlap failures rather than optimizer failures.

DPO, RLHF, and KL-regularized preference optimization. The dominant practical approach to aligning large language models with human preferences is to (i) collect pairwise comparisons, (ii) fit a reward or preference model, and (iii) optimize the generator with a regularizer that keeps it near a reference policy. Canonical RLHF implementations use policy-gradient methods such as PPO with an explicit KL penalty to a reference model **???**. More recent work shows that, under a Bradley–Terry–Luce (BTL) style likelihood and a particular parameterization of the reward in terms of log-ratio to a reference policy, one can optimize preferences without an explicit reward model via *Direct Preference Optimization* (DPO) **?**. Related KL-regularized objectives and implicit reward formulations appear in a number of places, often motivated as stable alternatives to RL or as approximations to maximum entropy RL **??**.

These methods primarily address the *policy improvement* step given a fixed dataset of comparisons. Our focus is complementary: we treat the dataset itself as an object of design, and we make explicit how the induced distribution over latent differences Δz governs the Fisher information and thus the achievable accuracy of any downstream estimator (explicit reward model, DPO-style implicit model, or other). From this viewpoint, KL regularization controls how far the learned policy moves in model space, but it does not by itself guarantee that the preference signal spans the relevant latent directions. This distinction matters for safety: if the preference dataset under-excites certain directions (e.g., factuality vs. style), then increasingly powerful optimization—even if perfectly regularized—can amplify spurious correlations because the learner is forced to extrapolate off-support. Empirically observed "reward hacking" and "sycophancy" phenomena can be interpreted as such extrapolation errors: the optimizer reliably improves what is identifiable from the comparisons, which may be a proxy dimension correlated with true user welfare in-distribution but decoupled out-of-distribution **??**.

Causal views of preference data: confounding, support, and counterfactuals. Preference learning is naturally causal because labels are gen-

erated after the platform chooses which model outputs to show (or which prompts to elicit), and because the choice of what comparisons to collect can confound the relationship between latent factors and observed preferences. The causal inference literature emphasizes that identification of causal effects requires overlap/positivity: for each covariate configuration of interest, all treatments must have nonzero probability **?**. In our setting, the "treatments" are interventions $a \in \mathcal{A}$ that change the distribution of response-side features $z_T(x, y)$, and the analog of positivity is the requirement that the induced distribution of Δz span all directions needed to identify w. This perspective aligns with off-policy evaluation and counterfactual risk minimization in contextual bandits, where logging policies that fail to cover relevant actions yield unidentifiable counterfactual values **??**. It also resonates with recent work emphasizing that alignment data are *selected*—via prompt sourcing, filtering, or annotator instructions—and that selection can create systematic blind spots **??**.

A related line of work studies preference elicitation under hidden confounders or heterogeneous annotators. Models of annotator noise, rater bias, and context effects complicate the mapping from a latent utility to observed pairwise labels **??**. Our latent-factor decomposition $z(x, y) = (z_X(x), z_T(x, y))$ can be viewed as a structured way to discuss such heterogeneity: prompt-side factors capture task mix and user population, while response-side factors capture stylistic and substantive properties. An important limitation of our abstraction is that it keeps the label model conditionally logistic given $(x, \tilde{y}, \tilde{y}')$; in real deployments, annotators may strategically adapt, norms evolve, and preference judgments depend on framing. Extending experimental design objectives to these richer, potentially nonstationary label-generating processes remains open, and is likely essential for governance-relevant guarantees.

Active preference learning, dueling bandits, and preference-based RL. Choosing which comparisons to label is a classical active learning problem. In the online learning literature, *dueling bandits* study how to identify high-utility actions when feedback is pairwise comparisons rather than scalar rewards **?**. Preference-based reinforcement learning extends these ideas to sequential decision-making with human comparisons **?**. Bayesian approaches treat the latent utility as a posterior over functions and adaptively query informative comparisons, including in preference-based Bayesian optimization **?**. Our design problem is closest in spirit to these works, but differs in two ways that matter for language-model alignment. First, our "arms" are not only prompts or candidate policies but also *interventions* that transform responses (e.g., length, tone, safety constraints, tool use), which directly modulate feature overlap rather than merely selecting among existing candidates. Second, we explicitly incorporate per-intervention costs $c(a)$, reflecting that

some comparisons (expert domains, adversarial prompts, red-teaming, multilingual evaluation) are more expensive, and that practical pipelines must trade off coverage against budget.

Notably, much of the active preference learning literature optimizes information gain or regret under a fixed hypothesis class. Our E-optimal criterion—maximize $\lambda_{\min}(I_q)$—is a worst-direction notion of informativeness that is particularly aligned with safety concerns: the system fails in the direction that is least identified. This is analogous to robustness-motivated active learning, where one seeks guarantees against adversarial shifts or rare but high-stakes subpopulations. It also suggests a concrete failure mode: if data collection policies greedily optimize for immediate reward improvement (e.g., focusing on comparisons with large expected preference margins), they may reduce $\sigma(\cdot)(1-\sigma(\cdot))$ curvature and simultaneously collapse the support of Δz, yielding brittle learned objectives even if training loss decreases.

Connections to classical and modern optimal experimental design. Our formalization is deliberately close to optimal design in generalized linear models, where Fisher information and eigenvalue criteria (A-, D-, and E-optimality) provide principled data acquisition rules **??**. The observation that optimal designs often require only a small support (via Carathéodory-type arguments) is standard in that literature and motivates our claim that a small mixture over interventions can be sufficient even when $\mathcal{A}$ is large. In modern machine learning, similar ideas appear in adaptive data collection, curriculum learning, and dataset distillation, though typically without explicit overlap guarantees. There is also a close link to active sampling for logistic regression and to experimental design for bandits, where exploration policies are often justified by ensuring adequate information matrix conditioning **?**.

At the same time, language-model alignment introduces complications absent from textbook design. The "design points" are themselves generated by a model $\pi_0(\cdot \mid x)$ and transformed by T_a, so the platform cannot directly set Δz but only steer its distribution. This makes controllability of T_a (and its interaction with the generator) central: some interventions change superficial style without perturbing substantive content, while others (e.g., tool augmentation, refusal policies, or retrieval) can move the system into entirely different regions of the latent space. From a safety standpoint, this highlights a governance-relevant lever: auditing and standardizing the intervention set $\mathcal{A}$ and its costs $c(a)$ is tantamount to auditing the experimental design space available to the platform.

Overlap, distribution shift, and robustness in alignment. Finally, our emphasis on overlap connects to the broader literature on distribution shift and robustness for ML systems **??**. Alignment deployments face sys-

tematic shifts: user populations change, adversaries adapt, and the model itself changes the prompt distribution by shaping user behavior. In such settings, purely in-distribution objectives can be misleading, and worst-case or minimax perspectives become natural. Our minimax design framing can be seen as a specific instantiation: we choose q to control the least-identified directions of w under plausible shifts. This connects conceptually to robust RLHF proposals (e.g., adversarial data collection and red-teaming) and to evaluation methodologies that stress-test rare behaviors **?**. The open challenge is to make these connections operational: specifying credible shift sets $\mathcal{P}_{\text{shift}}$, measuring latent overlap in situ, and integrating the resulting design constraints into real data pipelines without prohibitive cost.

In summary, while DPO and KL-regularized methods tell us how to update a policy given preference data, the causal and experimental design literatures explain when those updates are *well-founded*. Our contribution is to translate those identification requirements into a tractable platform design problem over interventions, making explicit the tradeoff between budget, controllability, and safety-critical coverage.

2 A tractable latent BTL model with interventions

We now instantiate a minimal model in which "what we learn" (a reward parameter) and "what we can learn" (identifiability from collected comparisons) are separated cleanly. The key move is to treat the preference dataset as an *endogenous* object: the platform controls which kinds of model outputs are compared, via interventions that systematically perturb response properties. This lets us formalize a concrete safety tradeoff. If we collect only cheap, convenient comparisons, we may fit a very accurate preference model *on-support* while leaving some latent directions essentially unobserved; downstream optimization can then extrapolate in precisely those directions, producing brittle objectives under distribution shift.

Primitives and data-collection protocol. A prompt is denoted $x \in \mathcal{X}$, drawn from a deployment-relevant distribution P_X (e.g., logged traffic, a curated mixture of tasks, or a worst-case mixture over user groups). A base generator $\pi_0(\cdot \mid x)$ proposes candidate responses $y \in \mathcal{Y}$. The platform has access to a finite intervention set $\mathcal{A}$; an intervention $a \in \mathcal{A}$ is a transformation

$$T_a : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}, \qquad \tilde{y} = T_a(x, y),$$

intended to modulate response-side properties (e.g., verbosity, tone constraints, tool-use, safety filters, retrieval augmentation, or domain-specific formatting). We allow interventions to depend on both the prompt and the base response because many realistic controls are post-processing or

decoding-time constraints that act on a candidate completion conditioned on x.

Data collection proceeds as follows for $t = 1, \ldots, N$: draw $x_t \sim P_X$; choose interventions $a_t, a'_t \sim q(\cdot \mid x_t)$ under a platform policy q; draw base candidates $y_t, y'_t \sim \pi_0(\cdot \mid x_t)$; form $\tilde{y}_t = T_{a_t}(x_t, y_t)$ and $\tilde{y}'_t = T_{a'_t}(x_t, y'_t)$; then request a preference label $L_t \in \{0, 1\}$ indicating whether the first transformed response is preferred. The distribution over observed tuples $(x_t, \tilde{y}_t, \tilde{y}'_t, L_t)$ is thus induced jointly by $(P_X, \pi_0, q, \{T_a\}_{a \in \mathcal{A}})$; in particular, the platform does not set comparison pairs directly, but only steers their distribution through q and the intervention operators.

Latent-factor map and linear latent reward. We posit a (possibly unknown) representation

$$z : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d, \qquad z(x, y) = \big(z_X(x), z_T(x, y)\big),$$

where $z_X(x) \in \mathbb{R}^{d_X}$ captures prompt-side/task-mix factors and $z_T(x, y) \in \mathbb{R}^{d_T}$ captures response-side factors, with $d = d_X + d_T$. This factorization is not a claim that prompts and responses are independent; rather, it enforces a bookkeeping distinction that becomes important under shift. Changes in user population move z_X, while decoding constraints and post-processing primarily move z_T. The platform's latent utility is assumed linear,

$$r(x, y) = w^\top z(x, y),$$

for an unknown $w \in \mathbb{R}^d$. We emphasize that linearity is a tractability assumption: it gives a transparent connection between dataset geometry and estimation error. In practice, one may treat z as a learned feature map and interpret w as the last-layer weights of a preference model; our analysis then describes which *directions* in feature space are weakly identified by the collected comparisons.

BTL label model and the role of margins. Given a prompt x and two transformed candidates $\tilde{y}, \tilde{y}'$, we define the latent difference

$$\Delta z \;=\; z(x, \tilde{y}) - z(x, \tilde{y}').$$

We assume a Bradley–Terry–Luce likelihood,

$$\mathbb{P}\big(L = 1 \mid x, \tilde{y}, \tilde{y}'\big) = \sigma\Big(w^\top \Delta z\Big), \qquad \sigma(u) = \frac{1}{1 + e^{-u}}.$$

This model captures a basic but deployment-relevant phenomenon: labels are most informative when comparisons are neither trivial nor impossible. Indeed, the curvature term $\sigma(u)(1 - \sigma(u))$ is maximized near $u = 0$ and

vanishes as $|u| \to \infty$. As a consequence, data-collection policies that preferentially sample "obvious wins" can decrease statistical efficiency even while increasing immediate agreement rates among annotators. For alignment, this creates a subtle failure mode: high inter-rater reliability does not imply that the resulting dataset identifies the underlying tradeoffs encoded in w.

Estimation objective and induced Fisher information. Let $\ell(w)$ denote the negative log-likelihood over N comparisons:

$$\ell(w) = \sum_{t=1}^{N} \Big(- \; L_t \log \sigma(w^\top \Delta z_t) - \; (1 - L_t) \log \big(1 - \sigma(w^\top \Delta z_t)\big) \Big).$$

The maximum-likelihood estimator $\hat{w}$ satisfies the score equation

$$\sum_{t=1}^{N} \Big(L_t - \sigma(\hat{w}^\top \Delta z_t) \Big) \Delta z_t = 0.$$

To connect data collection to achievable accuracy, we study the Fisher information under a design q:

$$I_q(w) = \mathbb{E}\Big[\sigma(w^\top \Delta z) \big(1 - \sigma(w^\top \Delta z)\big) \; \Delta z \Delta z^\top \Big],$$

where the expectation is taken over the random tuple $(x, \Delta z)$ generated by (P_X, π_0, q, T_a). Under standard regularity and boundedness assumptions (e.g., $\|\Delta z\| \le R$ almost surely and sufficient curvature on the realized support), the asymptotic covariance of $\hat{w}$ is $I_q(w_*)^{-1}/N$. This is the formal sense in which the platform's intervention policy shapes learnability: the spectrum of I_q is governed by which latent directions are excited by observed Δz vectors, and by whether preference outcomes saturate.

Positivity/overlap as an identification condition. The central identifiability requirement is an overlap (positivity) condition:

$$\lambda_{\min}\big(I_q(w_*)\big) \; \ge \; \underline{\lambda} \; > \; 0.$$

Intuitively, we need comparisons that vary the latent factors in linearly independent ways; otherwise, some components of w are unidentifiable no matter how large N is. When overlap fails, the problem is not that optimization is "hard"—it is that multiple distinct w explain the observed labels equally well because the dataset only explores a low-dimensional manifold of Δz. In alignment terms, this is precisely a blind-spot risk: the learned preference model may be forced to extrapolate how humans trade off, say, factuality versus style or helpfulness versus harmlessness, if the collected comparisons never independently vary those attributes. Subsequent policy optimization can then amplify whichever proxy direction is spuriously correlated with label outcomes on the observed support.

Intervention costs and the design problem. We incorporate a per-intervention cost $c(a)$ (generation overhead, evaluator expertise, latency, or operational risk). With a total labeling budget B, a natural per-sample constraint is

$$\mathbb{E}_{x\sim P_X,\ a\sim q(\cdot|x)}[c(a)] \ \le\ B/N.$$

The platform's design problem is then to choose q to make the worst-identified directions as identifiable as possible. A tractable robust surrogate is E-optimality:

$$\max_{q\in\mathcal{Q}}\ \lambda_{\min}\big(I_q(w_*)\big) \quad \text{s.t.} \quad \mathbb{E}[c(a)] \le B/N,$$

with $\mathcal{Q}$ encoding feasibility (e.g., $q(a \mid x) \ge 0$ and $\sum_a q(a \mid x) = 1$). The point is not that E-optimality is the unique correct criterion, but that it operationalizes a safety-relevant desideratum: we are optimizing the direction in which the system would otherwise fail under shift. This contrasts with designs that target average-case gains (e.g., trace or determinant criteria) and may neglect rare, high-stakes directions.

A two-factor specialization and "quadrant filling." To make the geometry explicit, consider a stylized case with $d = 2$ and a family of induced Δz distributions that are approximately elliptical, with correlation parameter $\rho(a)$ determined by the intervention. Passive collection (no meaningful intervention) often yields $|\rho| \approx 1$: many properties co-vary, so comparisons effectively lie near a one-dimensional curve. In this regime, $\lambda_{\min}(I_q)$ becomes small, and the sample size required to estimate both components of w grows like $1/(1-\rho^2)$. Interventions are valuable precisely insofar as they let us change this correlation structure by perturbing response-side factors without simultaneously moving all other factors.

In the extreme discrete analogue where $\Delta z \in \{-1, +1\}^2$, identifiability requires placing nonzero mass in each of the four quadrants; missing any quadrant yields $\lambda_{\min} = 0$ and hence non-identification. This motivates a practical heuristic: construct intervention mixtures that deliberately produce "crossed" comparisons (e.g., high factuality with low polish and vice versa), rather than only improving all dimensions simultaneously. Moreover, because information matrices live in a low-dimensional convex cone, optimal mixtures typically require only a small support over interventions, suggesting that a limited menu of well-chosen transformations can achieve most of the attainable overlap even when $\mathcal{A}$ is large.

Limitations and what must be monitored in practice. Two limitations matter for deployment. First, the platform rarely observes the true latent map z; it uses a proxy $\hat{z}$ (learned embeddings, heuristic attributes, or model-based scorers). Overlap must therefore be monitored in the proxy

space, and we should expect misspecification: good conditioning of an empirical information matrix in $\hat{z}$ does not guarantee conditioning in the true latent factors. Second, label generation can be non-stationary (annotator drift, changing norms, strategic behavior), violating the conditional logistic assumption. These issues do not eliminate the value of the design lens, but they shift the governance question: we should audit not only the trained preference model, but also the intervention set $\mathcal{A}$, the costing scheme $c(a)$, and the resulting empirical coverage diagnostics. In our view, making overlap a first-class metric is a concrete step toward verifiable guarantees that alignment training is not silently under-identifying safety-critical tradeoffs.

3 Overlap metrics and why they matter

The design objective in the previous section is phrased in terms of the Fisher information matrix $I_q(w_*)$, but in deployment we need to treat "overlap" as an *operational metric*: something we can estimate (even approximately) and use to predict when preference learning will generalize versus when it will produce brittle extrapolation. In our setting, overlap is not merely a support condition ("every action has nonzero probability") as in off-policy evaluation; it is a *geometric* property of the induced comparison distribution over latent differences Δz. Informally, good overlap means that the dataset contains comparisons that vary the salient latent factors in sufficiently independent directions, and at margins where labels retain curvature.

Fisher information as a curvature-weighted coverage matrix. Recall that under a fixed intervention policy q, each labeled comparison induces a random vector Δz, and the BTL likelihood contributes curvature proportional to $\sigma(u)(1-\sigma(u))$ at margin $u = w^\top \Delta z$. The population Fisher information is

$$I_q(w) = \mathbb{E}\Big[\sigma(w^\top \Delta z)\big(1-\sigma(w^\top \Delta z)\big)\ \Delta z \Delta z^\top\Big].$$

Two multiplicative effects matter. The matrix $\Delta z \Delta z^\top$ encodes *coverage of directions*: if Δz is nearly always aligned with a single vector, then $\mathbb{E}[\Delta z \Delta z^\top]$ is close to rank one regardless of sample size. The scalar factor $\sigma(w^\top \Delta z)(1-\sigma(w^\top \Delta z))$ encodes *label informativeness*: even if we cover many directions, comparisons at extreme margins are effectively deterministic and yield little curvature. Thus, overlap is inherently a joint property of (i) which differences we observe and (ii) where those differences land relative to the current tradeoffs w.

A useful approximation, when w is bounded and the induced margins satisfy $|w^\top \Delta z| \le M$ with nontrivial probability mass near 0, is that the curvature term is bounded away from 0 on the effective support. In that

regime we can sandwich

$$\underline{\alpha}\,\mathbb{E}[\Delta z \Delta z^\top] \;\preceq\; I_q(w) \;\preceq\; \overline{\alpha}\,\mathbb{E}[\Delta z \Delta z^\top],$$

for constants $0 < \underline{\alpha} \le \overline{\alpha} \le 1/4$ depending on M. This makes explicit that intervention design is, to first order, a problem of shaping the second-moment geometry of Δz, with curvature acting as a downweighting of "too-easy" comparisons.

Overlap as a minimum-eigenvalue condition, and a condition number. We summarize overlap through the smallest eigenvalue $\lambda_{\min}(I_q(w_*))$. The identification requirement

$$\lambda_{\min}\big(I_q(w_*)\big) \ge \underline{\lambda} > 0$$

is a quantitative positivity condition: it rules out latent directions along which the data provide vanishing curvature. Since the direction of greatest danger is precisely the least excited one, E-optimality targets $\lambda_{\min}$ directly.

For diagnostics and comparative statics it is convenient to introduce an "inverse-overlap" measure (or overlap condition number)

$$\kappa(q) \;=\; \frac{\lambda_{\max}(I_q(w_*))}{\lambda_{\min}(I_q(w_*))}, \qquad \text{and in particular} \qquad \kappa_{\mathrm{inv}}(q) = \frac{1}{\lambda_{\min}(I_q(w_*))}.$$

Large $\kappa(q)$ means the dataset is informative about some directions in w but nearly silent about others. In alignment terms, this corresponds to learning some preference tradeoffs sharply (e.g., minor stylistic choices) while leaving safety-critical tradeoffs weakly identified (e.g., factuality versus persuasive framing). The danger is not merely statistical inefficiency; it is that downstream optimization will tend to move into precisely those weakly constrained directions because they admit the largest apparent gains under the learned model.

From overlap to variance: why the smallest eigenvalue dominates. Under standard M-estimation regularity (bounded $\|\Delta z\|$, well-specified logistic likelihood on the realized support, and strong convexity in a neighborhood of w_*), the MLE $\hat{w}$ concentrates at a rate governed by the curvature of the population risk, i.e., by $I_q(w_*)$. In particular, the asymptotic covariance scales as $I_q(w_*)^{-1}/N$, and finite-sample bounds yield the characteristic dependence

$$\mathbb{E}\big[\|\hat{w} - w_*\|_2^2\big] \;\lesssim\; \frac{d}{N\;\lambda_{\min}(I_q(w_*))}.$$

This makes the safety-relevant point sharp: improving the *worst-direction* curvature is multiplicatively more valuable than further improving already-well-identified directions. A design that increases $\lambda_{\max}$ without increasing

$\lambda_{\min}$ can reduce average error metrics while leaving the most dangerous extrapolation risk unchanged.

We can also see the role of margins here. Suppose interventions make responses "uniformly better" along the current reward direction so that $w_*^\top \Delta z$ is typically large in magnitude. Then $\sigma(1-\sigma)$ becomes small, shrinking all eigenvalues of I_q simultaneously. This is the statistical shadow of an intuitive labeling phenomenon: if comparisons are consistently obvious, we may obtain high agreement but low information about fine-grained tradeoffs. Put differently, overlap is not just about spanning $\mathbb{R}^d$; it is also about keeping a substantial fraction of comparisons in the informative margin regime.

From overlap to OOD error: why estimation geometry becomes a deployment risk. The overlap metric matters because we rarely deploy the preference model on the same distribution of comparisons used to train it. In our framing, deployment changes the distribution of (x, y), hence of $z(x, y)$ and of the relevant differences. A simple way to connect this to risk is to consider a downstream decision rule that compares candidates by the learned score $\hat{r}(x, y) = \hat{w}^\top z(x, y)$. When the test environment induces differences with second moment $\Sigma_{\text{test}} = \mathbb{E}[\Delta z_{\text{test}} \Delta z_{\text{test}}^\top]$, a generic plug-in bound yields

$$\text{Regret}_{\text{OOD}} \lesssim \|\Sigma_{\text{test}}\| \, \mathbb{E}\|\hat{w} - w_*\|_2 \lesssim \|\Sigma_{\text{test}}\| \sqrt{\frac{d}{N \, \lambda_{\min}(I_q(w_*))}}.$$

The dependence on $\|\Sigma_{\text{test}}\|$ captures the scale of latent variation encountered at deployment; the dependence on $\lambda_{\min}(I_q)$ captures whether the training set actually constrained those variations. When a rare user group or a rare task family activates a latent direction that training failed to identify, the regret can be dominated by that single direction even if in-distribution validation metrics look strong. This is the formal sense in which overlap is a governance-relevant quantity: it connects data-collection choices to worst-case downstream behavior under plausible shifts.

Limited latent positivity: correlation and missing quadrants. Two stylized failure modes illustrate why overlap can collapse even with large datasets.

First, consider a $d = 2$ setting where Δz is approximately elliptical with correlation ρ near 1 under passive collection. Then $\mathbb{E}[\Delta z \Delta z^\top]$ has eigenvalues proportional to $1 \pm \rho$, and the smaller one scales like $1 - \rho$. Since I_q is approximately a curvature scalar times this second moment (when margins do not saturate), we obtain the characteristic penalty

$$\lambda_{\min}(I_{\text{passive}}(w_*)) = \Theta(1 - \rho^2),$$

so sample complexity for fixed accuracy scales as $1/(1-\rho^2)$. This is the continuous analogue of the intuition that the dataset lives near a one-dimensional manifold: two latent factors move together, so we cannot disentangle their weights.

Second, in a discrete "quadrant" analogue with $\Delta z \in \{-1, +1\}^2$, missing any quadrant yields non-identification because $\mathbb{E}[\Delta z \Delta z^\top]$ becomes singular. This failure mode is easy to underestimate in practice because it can occur even when each *coordinate* appears to vary: if z_1 and z_2 only appear with the same sign, then both coordinates change but never independently. Interventions that produce crossed comparisons (high–low versus low–high) are precisely what repairs this, and E-optimality formalizes the "fill all quadrants" heuristic as maximizing worst-direction curvature.

Estimating overlap in practice: empirical information and proxy features. In a real pipeline we do not observe z and do not know w_*. Nonetheless we can approximate overlap diagnostics in a proxy space using learned embeddings or attribute estimators $\hat{z}(x, y)$ and a current model $\hat{w}$. A natural empirical analogue of Fisher information is

$$\widehat{I} = \frac{1}{N} \sum_{t=1}^{N} \widehat{s}_t \; \Delta\hat{z}_t \Delta\hat{z}_t^\top, \qquad \widehat{s}_t = \sigma(\hat{w}^\top \Delta\hat{z}_t)\big(1 - \sigma(\hat{w}^\top \Delta\hat{z}_t)\big),$$

and we can track $\lambda_{\min}(\widehat{I})$ (or a regularized version) over time, across prompt strata, and across intervention types. Doing so turns overlap from an abstract identifiability condition into a monitoring target: we can detect whether new interventions genuinely introduce novel directions of variation or merely rescale already-common ones, and whether our collection policy is drifting toward saturated comparisons.

We should be explicit about limitations. Conditioning in $\hat{z}$ does not guarantee conditioning in the true latent factors that humans use, and $\widehat{I}$ can be overly optimistic if the proxy collapses distinct concepts. Nevertheless, overlap monitoring remains valuable as a *necessary* condition for safety: if even the proxy space is ill-conditioned, the true space is unlikely to be better, and the resulting reward model is predictably under-identified.

Why this metric belongs in alignment and governance discussions. Overlap metrics give a concrete language for a familiar alignment tension: cheaper data collection tends to follow natural correlations in model outputs (highly correlated attributes, obvious comparisons), whereas safety demands deliberate exploration of rare or uncomfortable tradeoffs. The smallest eigenvalue $\lambda_{\min}(I_q)$ is a compact summary of whether we are paying that exploration cost. Because optimal designs can often be supported on a small mixture of interventions, the prescription is not "collect everything"

but rather "ensure that what we collect spans the dangerous directions." The open problem is to make these guarantees robust to representation shift and label non-stationarity; overlap is not the whole story, but it is one of the few quantities that transparently connects intervention policy, statistical identifiability, and downstream risk under shift.

4 The optimal overlap design problem

The previous section treated overlap as a *diagnostic*: a way to predict when preference learning will be statistically well-posed and when it will be brittle under shift. We now turn overlap into a *decision variable*. Concretely, the platform controls a randomized intervention policy $q(a \mid x)$ over a finite action set $\mathcal{A}$, subject to a labeling budget and heterogeneous per-intervention costs. The design problem is to choose q so that the induced labeled comparison distribution excites the latent directions that matter for safety and generalization.

From interventions to information contributions. Fix a prompt x and an intervention $a \in \mathcal{A}$. Under our data-collection protocol, we sample $y, y' \sim \pi_0(\cdot \mid x)$ and transform $\tilde{y} = T_a(x, y)$ (and analogously for a'). Let $\Delta z = z(x, \tilde{y}) - z(x, \tilde{y}')$ denote the induced latent difference. The expected Fisher contribution of choosing (a, a') at prompt x is

$$M_{a,a'}(x; w) = \mathbb{E}\Big[\sigma(w^\top \Delta z)\big(1 - \sigma(w^\top \Delta z)\big)\ \Delta z \Delta z^\top \ \Big|\ x, a, a'\Big],$$

where the expectation is over the base generator randomness (and any stochasticity in T_a). If we choose a, a' independently from the same policy $q(\cdot \mid x)$, the prompt-conditional information becomes

$$I_q(x; w) = \sum_{a \in \mathcal{A}} \sum_{a' \in \mathcal{A}} q(a \mid x) q(a' \mid x)\, M_{a,a'}(x; w),$$

and the population Fisher information is $I_q(w) = \mathbb{E}_{x \sim P_X}\big[I_q(x; w)\big]$. This decomposition is useful because it makes clear where convexity and tractability enter: the platform does not directly choose Δz, but it does choose mixture weights over a finite set of information-contributing matrices.

E-optimal and A-optimal objectives. We focus on objectives that directly control worst-direction uncertainty. The E-optimal criterion selects q to maximize the smallest eigenvalue of the information matrix:

$$\max_{q \in \mathcal{Q}}\ \lambda_{\min}\big(I_q(w_*)\big) \qquad \text{s.t.} \qquad \mathbb{E}_{x \sim P_X}\Big[\sum_{a \in \mathcal{A}} q(a \mid x) c(a)\Big]\ \le\ B/N,$$

with feasibility constraints $\sum_a q(a \mid x) = 1$ and $q(a \mid x) \geq 0$ for all x. E-optimality is the natural formalization of the safety-motivated requirement that *no* latent direction remains weakly constrained.

For comparison, the A-optimal objective targets average variance (trace of the inverse information):

$$\min_{q \in \mathcal{Q}} \ \mathrm{tr}\big(I_q(w_*)^{-1}\big) \qquad \text{s.t. the same constraints.}$$

A-optimality can be statistically attractive when we care about mean-squared error aggregated across directions, but it can underemphasize the particular directions that are rare yet safety-relevant. In practice, one can interpolate between these by optimizing a regularized spectral objective (e.g., maximize $\lambda_{\min}(I_q + \gamma I)$, or minimize $\mathrm{tr}((I_q + \gamma I)^{-1})$) to trade robustness against overly aggressive exploration.

A minimax variant for distribution shift. Because the design choice is made at training time but evaluated at deployment, we often want robustness to plausible shifts. One stylized formulation is to posit an uncertainty set over environments, $\mathcal{P}_{\text{shift}}$, which may change the prompt distribution P_X and/or the generator-induced latent distribution given interventions. The robust E-optimal design is

$$\max_{q \in \mathcal{Q}} \ \min_{P \in \mathcal{P}_{\text{shift}}} \ \lambda_{\min}\big(I_{q,P}(w_*)\big) \qquad \text{s.t.} \qquad \mathbb{E}_P\big[c(a)\big] \leq B/N,$$

where $I_{q,P}$ emphasizes that the information depends on the environment through the induced distribution of $(x, \Delta z)$. This objective makes the governance interpretation explicit: we are choosing interventions to guarantee a minimum level of identifiability for the worst plausible deployment environment, rather than optimizing for the average user.

A useful special case is group robustness. If prompts are drawn from a mixture of groups $g \in \{1, \ldots, G\}$ with group-conditional distributions P_X^g, we can require

$$\max_{q} \ \min_{g \in [G]} \ \lambda_{\min}\big(I_q^g(w_*)\big),$$

which tends to allocate exploration mass toward groups for which passive overlap is worst. This formalizes the intuition that minority or edge-case usage should receive disproportionate measurement effort if it activates underidentified latent directions.

Cost-aware Lagrangians and KKT structure. The budget constraint is not cosmetic: it is the mechanism by which safety competes with operational realities. Introducing a multiplier $\eta \geq 0$, we can write an E-optimal Lagrangian

$$\mathcal{L}(q, \eta) = \lambda_{\min}\big(I_q(w_*)\big) \ - \ \eta \left(\mathbb{E}\Big[\sum_a q(a \mid x) c(a) \Big] - B/N \right).$$

At an optimum, the KKT conditions imply that interventions with positive mass must lie on the upper envelope of the tradeoff between marginal information gain (in the worst eigen-direction) and marginal cost. While $\lambda_{\min}$ is not differentiable everywhere, it has a well-defined subgradient: if v is a unit eigenvector corresponding to $\lambda_{\min}(I_q)$ (assume uniqueness for intuition), then locally the sensitivity of the objective to a perturbation δI is $v^\top(\delta I)v$. Thus, the design implicitly prioritizes interventions that increase curvature in the current worst-identified direction, not those that merely increase total variance explained.

When does the design reduce to a small mixture? Although $q(a \mid x)$ is, in principle, a policy over prompts, in many pipelines we can profitably study prompt-agnostic mixtures $\pi(a)$ as a baseline: choose a from π independently of x. This is justified when (i) prompts are high-dimensional and we lack reliable prompt-stratified estimates of information, or (ii) the main driver of overlap is the intervention-induced variation in response factors rather than prompt content. In this relaxation we solve

$$\max_{\pi \in \Delta(\mathcal{A})} \lambda_{\min}\Big(\sum_{a \in \mathcal{A}} \pi(a)\, \bar{M}_a\Big) \qquad \text{s.t.} \qquad \sum_a \pi(a) c(a) \le B/N,$$

where $\bar{M}_a$ is the average information contribution of intervention a (absorbing the a'-sampling convention into the definition). The feasible set is a convex polytope, and the mapping $\pi \mapsto \sum_a \pi(a)\bar{M}_a$ is linear. The consequence, familiar from classical optimal design, is that optimal mixtures are typically sparse: the optimizer can be supported on few interventions because it is selecting an extreme point of an information cone. Operationally, this matters because it turns an otherwise complex policy search into choosing a small "menu" of interventions and proportions.

Closed forms in two-factor models: decorrelation and quadrant filling. The strongest intuition emerges in $d = 2$, where geometry is visual. Suppose interventions primarily affect the correlation structure of $\Delta z = (\Delta z_1, \Delta z_2)$ while keeping marginal scales comparable. Passive collection may induce $\Delta z_1 \approx \Delta z_2$ (high ρ), making one eigen-direction nearly invisible. If the platform has access to at least two interventions whose induced correlations have opposite sign (or, more generally, span a range containing 0), then mixing them can drive the *effective* second-moment correlation toward 0, thereby maximizing the minimum eigenvalue of the moment matrix and, in the non-saturated regime, of $I_q(w_*)$ as well. In discrete analogues, the same phenomenon appears as "quadrant filling": the optimal design allocates mass so that all sign combinations of $(\Delta z_1, \Delta z_2)$ occur with comparable probability, eliminating missing-quadrant non-identification.

The key point is not the literal two-dimensionality but the mechanism: interventions should be selected to *break natural correlations* induced by the

base generator and by prompt distributions. In alignment-relevant terms, we should expect many harmful underidentification modes to be correlation-driven (e.g., "helpfulness" covarying with persuasive tone, or "harmlessness" covarying with refusal style), and therefore expect mixtures of deliberately chosen interventions to offer superlinear improvements over passive data.

Computation and implementation: from SDP ideals to practical heuristics. Optimizing $\lambda_{\min}$ subject to linear constraints admits standard convex-optimization reductions when the information depends linearly on the design (as in the mixture relaxation). One can introduce an auxiliary variable τ and impose the semidefinite constraint

$$\sum_a \pi(a)\bar{M}_a \;\succeq\; \tau I,$$

then maximize τ subject to cost and simplex constraints. This is an SDP and can be solved reliably at moderate $|\mathcal{A}|$ and d. For prompt-conditional $q(a \mid x)$ the problem becomes larger, but similar epigraph formulations apply if we discretize prompt strata.

In practice, two complications push us toward approximations. First, w_* is unknown, so $M_{a,a'}(x; w_*)$ is unavailable. Second, z is latent, so even the moment geometry must be estimated in a proxy space. A common approach is *sequential* design: maintain a current estimate $\hat{w}_t$ and proxy features $\Delta\hat{z}_t$, periodically re-estimate empirical information contributions $\widehat{M}_a$, and update the intervention mixture. This is a pragmatic compromise between full bandit-style adaptivity (which can be fragile under non-stationary labelers) and static designs (which can be wasteful early on).

Failure modes and open problems. Optimizing overlap is not synonymous with optimizing alignment. If the proxy representation collapses safety-critical distinctions, the design may confidently fill the wrong subspace. If labelers change criteria over time, the effective w is non-stationary, and information collected for yesterday's tradeoffs may not constrain tomorrow's. Moreover, robust (minimax) designs can over-allocate budget to adversarially unlikely shifts, harming average performance and potentially increasing exposure to harmful content during exploration.

These limitations point to a research agenda: coupling overlap design to verification mechanisms (audits on targeted slices, adversarial red-teaming as a structured intervention, and uncertainty-aware deployment constraints), and developing representations for which overlap metrics are not merely necessary but closer to sufficient. Nonetheless, treating overlap as an explicit optimization target is a substantive step: it forces us to encode, in the data-collection policy itself, which tradeoffs we are willing to pay to identify before we entrust downstream optimization with real-world decisions.

5 Closed-form results in the two-factor correlated-Gaussian case

To make the overlap story concrete (and to separate what is structural from what is an artifact of high-dimensional proxies), we now analyze a two-factor model in which the latent differences are approximately Gaussian and interventions primarily control correlation. This is the simplest setting where we can (i) compute Fisher-information eigenvalues in closed form up to a scalar, (ii) exhibit an explicit "decorrelating" optimal mixture, and (iii) read off label-complexity improvements and comparative statics in ρ, noise, and budget.

Model: intervention-indexed correlated Gaussians. Fix $d = 2$ and write $\Delta z = (\Delta z_1, \Delta z_2)$. For each intervention $a \in \mathcal{A}$, assume the induced latent difference is

$$\Delta z \mid a \;\sim\; \mathcal{N}(0, \Sigma_a), \qquad \Sigma_a = \begin{pmatrix} s_{1,a}^2 & \rho_a s_{1,a} s_{2,a} \\ \rho_a s_{1,a} s_{2,a} & s_{2,a}^2 \end{pmatrix},$$

with $|\rho_a| < 1$. We interpret ρ_a as a controllable "entanglement" between the two safety-relevant factors: passive data corresponds to a single baseline a_0 with $\rho_{a_0} \approx 1$, while active collection mixes interventions to reduce the *effective* correlation. For clarity, we first take $s_{1,a} = s_{2,a} = 1$ (correlation matrices) and later comment on unequal scales.

With the BTL/logistic label model, the Fisher information under a mixture $\pi(a)$ (prompt-agnostic for exposition) can be written as

$$I_\pi(w_*) = \mathbb{E}_{a\sim\pi}\, \mathbb{E}_{\Delta z\sim\mathcal{N}(0,\Sigma_a)}\Big[\sigma(w_*^\top \Delta z)\big(1-\sigma(w_*^\top \Delta z)\big)\; \Delta z \Delta z^\top\Big].$$

The nonlinearity $\sigma(\cdot)(1-\sigma(\cdot))$ couples w_* to the design. However, in the regime most relevant for learning (where comparisons are neither completely saturated nor completely random), this term behaves like a bounded scalar, letting us separate "margin/noise" effects from "overlap geometry" effects.

Decoupling geometry from the logistic curvature (a controlled approximation). Let $g(t) = \sigma(t)(1-\sigma(t)) \in (0, 1/4]$. If we assume $\|w_*\|_2 \le W$ and the design keeps $\mathbb{E}\|\Delta z\|_2^2$ bounded (true for Gaussians with bounded covariance), then $w_*^\top \Delta z$ is sub-Gaussian and $g(w_*^\top \Delta z)$ is typically bounded away from 0 on most mass. A convenient sufficient condition is to restrict to a "non-saturated" design class where

$$\underline{\alpha} \;\le\; \mathbb{E}\big[g(w_*^\top \Delta z) \mid a\big] \;\le\; \bar{\alpha} \qquad \text{for all } a \text{ in the support of } \pi,$$

for some constants $0 < \underline{\alpha} \le \bar{\alpha} \le 1/4$. Under this condition (or more formally, by sandwiching $g(\cdot)$ and using rotational symmetry of Gaussians), we obtain

the matrix inequality

$$\underline{\alpha}\,\Big(\mathbb{E}_{a\sim\pi}[\Sigma_a]\Big) \;\preceq\; I_\pi(w_*) \;\preceq\; \bar{\alpha}\,\Big(\mathbb{E}_{a\sim\pi}[\Sigma_a]\Big).$$

Thus, up to the scalar factor α induced by label noise and margin saturation, optimizing $\lambda_{\min}(I_\pi(w_*))$ reduces to optimizing $\lambda_{\min}(\Sigma_{\text{eff}})$ where

$$\Sigma_{\text{eff}} \;:=\; \mathbb{E}_{a\sim\pi}[\Sigma_a].$$

In 2D with unit variances, Σ_{eff} is fully described by its effective correlation $\bar{\rho} = \mathbb{E}_{a\sim\pi}[\rho_a]$:

$$\Sigma_{\text{eff}} = \begin{pmatrix} 1 & \bar{\rho} \\ \bar{\rho} & 1 \end{pmatrix}, \qquad \lambda_{\min}(\Sigma_{\text{eff}}) = 1 - |\bar{\rho}|.$$

Near $|\bar{\rho}| \approx 1$, this behaves as $\Theta(1-\bar{\rho}^2)$ up to constants, matching the overlap penalty in Proposition 2.

Passive label complexity blows up as $|\rho| \to 1$. If we collect comparisons passively under a single intervention a_0 with correlation $\rho = \rho_{a_0}$, then $\Sigma_{\text{eff}} = \Sigma_{a_0}$ and, in the non-saturated regime,

$$\lambda_{\min}(I_{\text{passive}}(w_*)) \;\asymp\; \alpha\,(1-|\rho|) \;\approx\; \alpha\,(1-\rho^2),$$

where $\alpha \in [\underline{\alpha}, \bar{\alpha}]$ summarizes label noise/margins. Combining with the generic MLE scaling (Proposition 1) specialized to $d = 2$ yields the passive label complexity

$$N_{\text{passive}}(\epsilon) \;=\; \tilde{O}\!\left(\frac{1}{\alpha\,(1-\rho^2)\,\epsilon^2}\right).$$

This makes the failure mode explicit: when the base generator (and prompt mix) entangles the two latent factors so that $\rho \approx 1$, one direction becomes nearly unidentifiable. If $1-\rho$ is on the order of 2^{-b} (i.e., b bits of "near-collinearity"), then N_{passive} grows on the order of 2^b for fixed ϵ, which is the sense in which correlation can induce an exponential-in-precision data burden.

Active decorrelation: a closed-form optimal mixture. Now suppose the platform can choose between two interventions a_+ and a_- whose induced correlations satisfy $\rho_+ > 0$ and $\rho_- < 0$, and (for now) both have unit marginal variances. Consider a mixture π_θ putting mass θ on a_+ and $1-\theta$ on a_-. Then

$$\bar{\rho}(\theta) = \theta\rho_+ + (1-\theta)\rho_-, \qquad \lambda_{\min}(\Sigma_{\text{eff}}(\theta)) = 1 - |\bar{\rho}(\theta)|.$$

The E-optimal choice is immediate: pick θ so that $\bar{\rho}(\theta) = 0$, i.e.,

$$\theta^* = \frac{-\rho_-}{\rho_+ - \rho_-} \in (0, 1),$$

which yields $\Sigma_{\text{eff}} = I$ and hence $\lambda_{\min}(\Sigma_{\text{eff}}) = 1$. In words, we mix the two interventions just enough to cancel the correlation induced by each, "filling" the missing eigen-direction. Under the same non-saturation condition, this implies

$$\lambda_{\min}(I_{\pi_{\theta^*}}(w_*)) \asymp \alpha, \qquad N_{\text{active}}(\epsilon) = \tilde{O}\left(\frac{1}{\alpha\,\epsilon^2}\right).$$

The improvement factor relative to passive collection is therefore on the order of

$$\frac{N_{\text{passive}}(\epsilon)}{N_{\text{active}}(\epsilon)} = \tilde{\Omega}\left(\frac{1}{1-\rho^2}\right),$$

which diverges as $|\rho| \to 1$. This is the simplest analytic instance of the general "quadrant-filling" intuition: we do not need to know w_* perfectly to know that correlation collapse destroys identifiability, and that mixing interventions that induce different correlation signs repairs it.

Quadrant filling as a sign-coverage statement. Although the Gaussian model is continuous, its geometry can be understood discretely by looking at signs. When $\rho \approx 1$, most mass lies near the diagonal $\Delta z_1 \approx \Delta z_2$, so the sign patterns $(+,-)$ and $(-,+)$ are exponentially rare in the tails, and any estimator that needs those contrasts (to distinguish weights on factor 1 vs factor 2) is starved. Interventions with negative ρ rotate mass toward the anti-diagonal, increasing the frequency of those "off-diagonal" sign patterns. The decorrelating mixture makes these sign quadrants comparably likely, which is precisely what is needed to keep $\lambda_{\min}$ bounded away from 0.

Comparative statics: correlation, noise/margins, and budget. The closed forms also let us cleanly state three comparative statics that are operationally important.

(1) Correlation $|\rho|$. Holding α fixed, $\lambda_{\min}$ decreases as $|\bar{\rho}|$ increases, and the passive sample complexity increases like $1/(1-\rho^2)$. The key point is that *variance is not enough*: one can have large total second moment $\text{tr}(\Sigma)$ while still having $\lambda_{\min}(\Sigma) \approx 0$.

(2) Noise and saturation through α. The factor α is maximized when comparisons are "just hard enough": if $w_*^\top \Delta z$ is typically near 0, then $g(\cdot)$ is near its maximum $1/4$; if $|w_*^\top \Delta z|$ is typically large, then $g(\cdot)$ is near 0 and information collapses even if overlap is good. This yields a design tension: interventions that increase variance (e.g., by making responses more extreme) can *reduce* information by pushing labels into a near-deterministic regime. In practice, we should therefore treat overlap maximization as coupled to

a "difficulty calibration" problem: we want to fill directions while keeping margins in the informative band.

(3) Budgeted mixing with heterogeneous costs. Let c_+ and c_- be the per-sample costs of a_+ and a_-. Under an average cost constraint $\theta c_+ + (1-\theta)c_- \leq \bar{c}$, the decorrelating mixture θ^* may be infeasible. In the unit-variance case, the constrained optimum is to choose the feasible θ that minimizes $|\bar{\rho}(\theta)|$; equivalently, "spend as much budget as possible" on whichever intervention moves $\bar{\rho}$ toward 0 most efficiently per unit cost. This induces a threshold phenomenon: once $\bar{c}$ is large enough to permit $\theta = \theta^*$, additional budget no longer improves $\lambda_{\min}$ through correlation (though it may still help via richer intervention sets or higher-quality labelers that increase α). Below the threshold, $\lambda_{\min}$ improves approximately linearly with budget because $|\bar{\rho}(\theta)|$ is affine in θ.

Unequal marginal scales and three-point mixtures. If interventions also change the marginal variances ($s_{1,a} \neq s_{2,a}$), then Σ_{eff} depends on both scale and correlation, and the optimal design may trade off "whitening" against decorrelation. In 2D, a useful robust heuristic is to allow a third intervention a_0 (often the cheap/passive one) and solve for a three-point mixture that simultaneously (i) keeps α large (avoids saturation), (ii) balances effective variances to prevent one coordinate from dominating, and (iii) drives effective correlation toward 0. This is the continuous analogue of the discrete result that, in $d = 2$, a small support (often ≤ 3 actions) suffices to achieve the E-optimal point on the cost–information frontier.

Taken together, the two-factor Gaussian case provides a clean "mechanistic" picture: passive preference data can be arbitrarily sample-inefficient when latent factors are naturally entangled, while a cost-aware mixture of a few targeted interventions can restore overlap, control worst-direction uncertainty, and yield unbounded gains as $|\rho| \to 1$, provided we avoid regimes where labels saturate and Fisher curvature vanishes.

Why the two-factor Gaussian case is the right "toy" for overlap. In many alignment-relevant datasets, what we ultimately need to learn is not a single monolithic notion of "quality," but relative weights on multiple partially-confounded desiderata (e.g., truthfulness versus politeness, harmlessness versus helpfulness). The failure mode we worry about is not that responses have low variance overall, but that the variance lives in (approximately) a one-dimensional manifold: the generator and prompt distribution jointly move factors together. In that regime, pairwise comparisons are plentiful yet systematically uninformative about some directions of w_*. The two-factor correlated-Gaussian model isolates exactly this geometry: correlation plays the role of an "entanglement knob," and E-optimal design corresponds to actively creating comparisons that break the confounding.

A slightly more explicit Fisher-information decomposition. While the logistic curvature term $g(t) = \sigma(t)(1 - \sigma(t))$ complicates exact closed forms, we can still make the separation between (i) geometry of Δz and (ii) saturation/noise effects more formal than a heuristic scalar bound. Let $\Delta z \sim \mathcal{N}(0, \Sigma)$ and write $u = w_*^\top \Delta z$. Since $(\Delta z, u)$ is jointly Gaussian, Stein identities yield that for any twice-differentiable scalar function f with integrable derivatives,

$$\mathbb{E}\Big[f(u)\,\Delta z \Delta z^\top\Big] = \mathbb{E}[f(u)]\,\Sigma + \mathbb{E}[f''(u)]\,\Sigma w_* w_*^\top \Sigma. \tag{1}$$

Taking $f = g$ gives an exact representation of $I(w_*)$ (for a fixed Σ) as a sum of a "baseline" term proportional to Σ and a rank-one correction aligned with Σw_*. Two observations follow. First, regardless of w_*, the smallest eigenvalue obeys

$$\lambda_{\min}(I(w_*)) \;\geq\; \mathbb{E}[g(u)]\,\lambda_{\min}(\Sigma), \tag{2}$$

because the second term in (1) is positive semidefinite when $\mathbb{E}[g''(u)] \geq 0$ (and even without that sign, it is rank-one and cannot repair a singular Σ in directions orthogonal to Σw_*). Second, the only way to make $\lambda_{\min}(I(w_*))$ uniformly large over w_* is to keep both $\mathbb{E}[g(u)]$ away from 0 (avoid saturation) *and* $\lambda_{\min}(\Sigma)$ away from 0 (ensure overlap). This clarifies why correlation collapse is structurally dangerous: no amount of label quality can compensate for $\lambda_{\min}(\Sigma) \approx 0$.

Passive data: correlation induces an exponential-in-precision burden. Specializing to unit variances, $\Sigma = \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$ has eigenvalues $1 \pm \rho$, so $\lambda_{\min}(\Sigma) = 1 - |\rho|$. Plugging into (2) yields

$$\lambda_{\min}(I_{\text{passive}}(w_*)) \;\gtrsim\; \mathbb{E}[g(u)]\,(1 - |\rho|).$$

Thus Proposition 1 implies $N(\epsilon)$ scales like $1/\big(\mathbb{E}[g(u)](1 - |\rho|)\epsilon^2\big)$ up to logarithms and constants. The "exponential in bits of precision" interpretation is worth stating carefully because it is the practical governance concern: if $\rho = 1 - 2^{-b}$ (near-collinearity at b bits), then $1 - |\rho| \asymp 2^{-b}$ and the required labels scale as $\Omega(2^b)$ for fixed target error. In other words, the data burden explodes precisely when the generator makes factors move together so reliably that users rarely elicit countervailing tradeoffs. This is a plausible real-world scenario for safety: models may be trained to be simultaneously "more helpful" and "more aligned" in ways that hide the boundaries where helpfulness and harmfulness diverge, leaving the platform with little information about how to set (or even identify) the correct tradeoff weights.

Active mixing as "effective whitening" and why two interventions can suffice. Now consider a mixture over interventions, inducing a mixture

over covariances. Under the same unit-variance simplification and prompt-agnostic mixing, $\Sigma_{\text{eff}} = \mathbb{E}_{a\sim\pi}[\Sigma_a]$ again has the form

$$\Sigma_{\text{eff}} = \begin{pmatrix} 1 & \bar{\rho} \\ \bar{\rho} & 1 \end{pmatrix}, \qquad \bar{\rho} = \mathbb{E}_{a\sim\pi}[\rho_a],$$

so $\lambda_{\min}(\Sigma_{\text{eff}}) = 1 - |\bar{\rho}|$. The key point is that we can change $\bar{\rho}$ by changing the intervention mix even if we do not directly observe Δz. If we have two controllable interventions with opposite-sign correlations, $\rho_+ > 0$ and $\rho_- < 0$, then the mixture weight $\theta^* = -\rho_-/(\rho_+ - \rho_-)$ achieves $\bar{\rho} = 0$, i.e., $\Sigma_{\text{eff}} = I$. Combining with (2) gives

$$\lambda_{\min}(I_{\text{active}}(w_*)) \;\gtrsim\; \mathbb{E}[g(u)] \cdot 1,$$

so the improvement factor in label complexity scales like $(1 - |\rho|)^{-1}$ (or equivalently $(1 - \rho^2)^{-1}$ up to constants when $|\rho| \approx 1$). This is the continuous analogue of "quadrant filling": rather than hoping the passive generator occasionally produces $(+,-)$ and $(-,+)$ contrasts in the tails, we proactively choose transformations that shift mass toward those missing sign patterns.

Interpreting "quadrants" as safety-relevant counterfactuals. The sign-quadrant picture is not merely geometric; it corresponds to counterfactual evaluations that are often missing in practice. If Δz_1 encodes a safety attribute (e.g., harmfulness) and Δz_2 encodes a utility attribute (e.g., helpfulness), then $\rho \approx 1$ means we mostly see "better on both" versus "worse on both," which is easy for labelers but nearly useless for learning the tradeoff. The off-diagonal quadrants correspond to precisely the difficult comparisons we need: "more helpful but less safe" versus "less helpful but safer." Interventions that create such contrasts include: adversarial prompt variants that elicit borderline unsafe content while preserving task utility; formatting or style constraints that preserve content but change perceived politeness (to decouple politeness from correctness); and controlled refusals that keep safety fixed while varying helpfulness within safe bounds. The design lens here flags a safety failure mode: collecting only easy comparisons may systematically entrench confounding, yielding a model that appears well-aligned on aggregate metrics while being poorly identified on the very tradeoffs that matter under distribution shift.

Noise, saturation, and the "difficulty calibration" constraint. The preceding gains assume $\mathbb{E}[g(u)]$ does not collapse. Yet $g(u)$ is maximized near $u = 0$ and decays when $|u|$ is large (near-deterministic preferences). Interventions that aggressively increase the spread of Δz can therefore backfire: they may improve $\lambda_{\min}(\Sigma_{\text{eff}})$ while simultaneously shrinking $\mathbb{E}[g(u)]$ by pushing comparisons into a saturated regime. From (1), this is not a second-order nuance; $\mathbb{E}[g(u)]$ scales the *entire* baseline term. Operationally,

we can think of a "Goldilocks" constraint: we want Δz to explore directions broadly, but with magnitudes such that $w_*^\top \Delta z$ remains in an informative band. In deployment terms, this argues for interventions that *recombine* factors (change correlation) rather than simply making outputs extreme, and for labeler protocols that encourage fine-grained judgments (to avoid near-deterministic labels). It also motivates the algorithmic focus of the next section: we need to estimate information online and adaptively steer toward comparisons that are both diverse and non-saturated.

Budgeted mixing: a simple cost–overlap frontier. When interventions have heterogeneous per-sample costs, the decorrelating mixture may be infeasible. Suppose we again mix a_+ and a_- with cost constraint $\theta c_+ + (1-\theta)c_- \le \bar{c}$. Since $|\bar{\rho}(\theta)|$ is convex and piecewise-linear in θ, the constrained E-optimal solution is achieved at the feasible θ closest to θ^*, i.e.,

$$\theta^*_{\text{bud}} \in \arg\min_{\theta\in[0,1]} \; |\theta\rho_+ + (1-\theta)\rho_-| \quad \text{s.t.} \quad \theta c_+ + (1-\theta)c_- \le \bar{c}.$$

This yields a threshold phenomenon: if $\bar{c}$ is high enough to allow $\theta = \theta^*$, then correlation-driven overlap cannot be further improved by budget (though other benefits, like higher-quality labelers improving $\mathbb{E}[g(u)]$, remain). Below the threshold, $\lambda_{\min}(\Sigma_{\text{eff}}) = 1 - |\bar{\rho}(\theta)|$ increases approximately linearly with additional budget because $\bar{\rho}(\theta)$ is affine in θ. This is an experimentally actionable prediction: small incremental spend on targeted, expensive interventions can yield disproportionate reductions in worst-direction uncertainty when passive data are highly entangled.

Beyond unit variances: why three-point mixtures appear and what breaks. If interventions also change marginal scales, Σ_a is no longer determined by ρ_a alone, and the E-optimal design seeks to "whiten" Σ_{eff} rather than merely set $\bar{\rho} = 0$. In 2D, the cost–information optimum often lies on a face of the convex hull of $\{\Sigma_a\}$, so a mixture supported on a small number of actions suffices (consistent with Proposition 4). Practically, this suggests maintaining (i) a cheap baseline action to ensure coverage and avoid brittleness, plus (ii) one or two targeted actions that independently modulate correlation and scale. There are also clear limitations of the Gaussian abstraction: real Δz may be heavy-tailed, multimodal, and prompt-dependent, and labelers may deviate from BTL (especially on ambiguous safety content). Nonetheless, the main takeaway is robust: identifiability is governed by worst-direction overlap, and interventions should be evaluated by their ability to populate missing tradeoff comparisons under realistic cost and saturation constraints. This sets up the algorithmic problem we turn to next: in general d and without direct access to Σ_a, how do we approximate I_q from data and adaptively allocate budget across interventions to maximize $\lambda_{\min}$ subject to feasibility?

6 Algorithms for the general case: plug-in Fisher estimates and uncertainty-aware intervention design

In the general setting, we do not observe the latent factors $z(x, y)$, the intervention-to-factor map T_a, or the true weight vector w_*. What we *do* control is the data-collection loop: for each prompt x we can choose an action $a \in \mathcal{A}$ (often with prompt-dependent feasibility), generate candidate responses, apply T_a, and purchase a comparison label. The algorithmic problem is therefore a coupled estimation–design task: we want to learn w_* while simultaneously steering the intervention mixture toward regions of the latent space that make w_* identifiable, under cost and safety constraints.

A feature-based proxy and a plug-in Fisher estimator. Since z is latent, we work with a learned feature map $\phi(x, y) \in \mathbb{R}^d$ (e.g., a frozen encoder, a reward-model penultimate layer, or a task-specific representation) and treat $\Delta\phi_t = \phi(x_t, \tilde{y}_t) - \phi(x_t, \tilde{y}'_t)$ as a proxy for Δz_t. Under the BTL/logistic model, the (population) Fisher information for w takes the form

$$I(w) = \mathbb{E}\Big[g\Big(w^\top \Delta z\Big)\ \Delta z \Delta z^\top\Big], \qquad g(u) = \sigma(u)\big(1 - \sigma(u)\big),$$

so a natural empirical plug-in estimate is

$$\widehat{I}_t = \frac{1}{t}\sum_{s=1}^{t} g\Big(\hat{w}_t^\top \Delta\phi_s\Big)\ \Delta\phi_s \Delta\phi_s^\top \ +\ \lambda I_d, \tag{3}$$

where $\hat{w}_t$ is the current MLE (or regularized MLE) on the collected comparisons and $\lambda > 0$ is a small ridge term used both for numerical stability and to encode a prior lower bound on unmodeled overlap. This estimator is cheap to maintain online: we can update the sum in (3) incrementally and recompute $\hat{w}_t$ either in batch or via stochastic Newton/gradient steps using the score equation. The ridge term is not merely a numerical trick; it corresponds to an explicit stance that, when our feature proxy is misspecified, we should avoid overconfidently declaring a direction unidentifiable.

Design as an online optimization over intervention mixtures. Given $\widehat{I}_t$, the design objective suggested by the earlier overlap analysis is to increase $\lambda_{\min}(\widehat{I}_t)$ as quickly as possible per unit cost. The simplest abstraction is prompt-agnostic mixing: choose a distribution $\pi \in \Delta(\mathcal{A})$ and sample $a \sim \pi$ each round. If we had access to the per-action information contributions $M_a(w) = \mathbb{E}[g(w^\top \Delta z)\Delta z \Delta z^\top \mid a]$, the E-optimal design would solve

$$\max_{\pi \in \Delta(\mathcal{A})}\ \lambda_{\min}\left(\sum_{a \in \mathcal{A}} \pi(a)\, M_a(w_*)\right) \quad \text{s.t.} \quad \sum_a \pi(a) c(a) \le \bar{c}.$$

In practice $M_a(w_*)$ is unknown, so we replace it with an estimate $\widehat{M}_{t,a}$ constructed from the subset of samples collected under action a (again using the plug-in curvature term with $\hat{w}_t$). We then repeatedly solve the surrogate convex program

$$\pi_{t+1} \in \arg\max_{\pi \in \Delta(\mathcal{A})} \lambda_{\min}\left(\sum_a \pi(a)\, \widehat{M}_{t,a}\right) \quad \text{s.t.} \quad \sum_a \pi(a)c(a) \le \bar{c}, \tag{4}$$

and sample the next intervention accordingly. This is a direct analogue of classical optimal design, but with two alignment-relevant complications: (i) the curvature term $g(\hat{w}_t^\top \Delta\phi)$ couples the objective to the current estimator, and (ii) feasibility and safety constraints make $\mathcal{A}$ effectively prompt-dependent.

Uncertainty-aware selection: optimism, Thompson sampling, and safe exploration. A purely greedy plug-in strategy based on (4) can fail early: if $\widehat{M}_{t,a}$ is inaccurate, the algorithm may prematurely commit to interventions that appear informative under the current (wrong) $\hat{w}_t$, starving the dataset of the counterfactual comparisons needed to correct that mistake. We therefore treat intervention choice as a bandit-like problem where the "reward" is information gain.

One practical approach is *optimism under uncertainty*. For each action a, we maintain a confidence set $\mathcal{C}_{t,a}$ for M_a (e.g., via matrix concentration applied to the empirical second moments of $\Delta\phi$ reweighted by $g(\hat{w}_t^\top \Delta\phi)$). We then choose a mixture π that maximizes a lower confidence bound on the smallest eigenvalue:

$$\pi_{t+1} \in \arg\max_{\pi} \min_{M_a \in \mathcal{C}_{t,a}} \lambda_{\min}\left(\sum_a \pi(a) M_a\right) \quad \text{s.t. cost/feasibility.}$$

This "robust E-optimal" step explicitly allocates samples to reduce worst-direction uncertainty in the information matrices, not just to exploit current estimates.

A complementary alternative is *Thompson sampling* over designs: sample plausible M_a (or a low-dimensional parametrization of M_a) from an approximate posterior and solve the corresponding E-optimal design. This tends to induce natural randomization, which is desirable for overlap/positivity. In safety-critical settings, we can further require *safe exploration* constraints—for instance, a minimum mass $\gamma > 0$ on a baseline intervention that is known to be benign and cheap, ensuring that the collection policy never collapses to a narrow slice of behavior that might hide harmful corner cases.

Contextual feasibility: prompt-dependent action sets and controllable generation. Most interventions are not universally applicable. Some

edits require that the base response contain certain structures; some adversarial variants are only meaningful for particular tasks; and some safety-sensitive transformations are disallowed for certain prompts or jurisdictions. Formally, we can model this by an action set $\mathcal{A}(x) \subseteq \mathcal{A}$ and optimize a contextual policy $q(a \mid x)$. A tractable approximation is to cluster prompts into finitely many buckets $b(x) \in [K]$ (by topic, risk level, or embedding) and learn a separate mixture $\pi^{(k)}$ per bucket, solving a version of (4) with data restricted to that bucket. This yields a design that is both implementable and auditable: we can report, per bucket, the implied worst-direction uncertainty via $\lambda_{\min}(\widehat{I}_t^{(k)})$.

On the generation side, interventions often correspond to *controllable decoding* (length caps, style tags, refusal templates), *structured edits* (paraphrases that preserve semantics), or *counterfactual elicitation* (asking for alternative solutions, edge-case analyses, or safety disclaimers). The key algorithmic constraint is that the effective transformation T_a is stochastic and sometimes fails (the model refuses, ignores constraints, or drifts semantically). We therefore treat each attempted intervention as producing either a valid transformed output or a null outcome; the design must account for the acceptance probability $\alpha(a, x)$, since the true per-labeled-sample cost is closer to $c(a)/\alpha(a, x)$. Empirically, incorporating acceptance rates prevents the policy from over-allocating to "theoretically perfect" actions that rarely succeed.

Difficulty calibration: avoiding saturation while increasing overlap. Because $g(u)$ shrinks when comparisons become too easy (large $|u|$) or too noisy (effectively random labels), we want interventions that generate *informative margins*. A useful operational heuristic is to target comparisons with predicted $|\hat{w}_t^\top \Delta\phi| \approx \tau$ for a moderate τ, which can be implemented by rejection sampling over generated candidate pairs or by generating multiple candidates and selecting the pair with a near-threshold predicted margin. This resembles active learning: we spend generation budget to shape the labeled set toward regions where the logistic model has high curvature, while still ensuring geometric diversity in $\Delta\phi$.

Computational approximations for large d and large $\mathcal{A}$. Directly optimizing $\lambda_{\min}$ can be expensive when d is large. Two pragmatic relaxations are common. First, we can replace the E-optimal objective with a smooth surrogate such as $\log\det(\widehat{I}_t)$ (D-optimality) or $\mathrm{tr}(\widehat{I}_t^{-1})$ (A-optimality), which admit stable gradients and work well with first-order methods. Second, we can approximate $\lambda_{\min}$ using a few iterations of the power method on $\widehat{I}_t^{-1}$, enabling online selection without full eigendecomposition.

When $\mathcal{A}$ is large or continuous (e.g., prompt perturbation strength, decoding temperature), we can parameterize $q_\theta(a \mid x)$ and perform stochastic

gradient ascent on a differentiable surrogate of the design objective, using implicit differentiation through the estimator $\hat{w}_t$ only approximately (e.g., treating $\hat{w}_t$ as fixed for several steps). While this breaks the clean separation between estimation and design, it is often the only scalable route.

What can go wrong and what we can still certify. All of the above relies on a proxy ϕ and on approximate adherence to a BTL-like label model. Misspecification can induce "illusory overlap": $\widehat{I}_t$ looks well-conditioned in feature space while the true latent tradeoff remains under-identified. To mitigate this, we can (i) track multiple feature maps (ensemble encoders) and optimize for worst-case conditioning across them, (ii) periodically run targeted audits that directly test missing-quadrant comparisons, and (iii) enforce minimum exploration across a diverse action set as a governance constraint rather than an optimization outcome.

These algorithmic components—plug-in information estimation, uncertainty-aware intervention allocation, and feasibility-aware controllable generation—define an implementable pipeline for actively *creating* the comparisons that reveal safety–utility tradeoffs. The remaining question is empirical: how do these strategies behave under controlled correlation stress tests, and do they translate into better downstream policies under distribution shift? We turn next to a concrete blueprint for answering that question.

7 Empirical blueprint: stress tests, controlled edits, and downstream robustness

Our theoretical claims ultimately live or die on an empirical question: can an intervention policy *reliably* create the missing counterfactual comparisons that determine safety–utility tradeoffs, without quietly trading away robustness through proxy misspecification or feasibility failures? We therefore recommend a three-part evaluation program that escalates from fully controlled synthetic environments (where ground truth overlap is measurable) to real preference data with controlled interventions (where overlap must be tracked in proxy space) and finally to downstream policy optimization under distribution shift (where the costs of non-identification become operational).

7.1 (i) Synthetic latent-factor stress tests: correlation sweeps and "correlation flips"

Goal. We want a testbed in which the latent representation $z(x, y)$ is *known*, the true reward weights w_* are known, and the platform can choose from a finite intervention set $\mathcal{A}$ that induces tunable correlations among the components of Δz. This lets us verify, without ambiguity, whether the

active design pipeline increases $\lambda_{\min}(I_q)$ and achieves the predicted sample-complexity gains as correlations approach $\rho \to 1$.

A minimal generative model. A convenient construction is:

1. Sample a prompt type $x \sim P_X$ (optionally a mixture over clusters to mimic heterogeneous traffic).

2. Sample a base response $y \sim \pi_0(\cdot \mid x)$ by generating a latent feature vector $z(x, y) \in \mathbb{R}^d$ and then (optionally) a surface form. For stress tests we can skip surface forms entirely and treat z as observed by the simulator but not by the platform.

3. For each intervention $a \in \mathcal{A}$, define a stochastic map on features, e.g.

$$z_T(x, \tilde{y}) = A_a z_T(x, y) + b_a + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \Sigma_a),$$

   so that each a yields a different second-moment structure for Δz and, in $d = 2$, a different effective correlation $\rho(a)$.

4. Generate labels from the BTL model:

$$\mathbb{P}(L = 1 \mid x, \tilde{y}, \tilde{y}') = \sigma\Big(w_*^\top \Delta z\Big), \qquad \Delta z = z(x, \tilde{y}) - z(x, \tilde{y}').$$

The key controllability knob is that some actions should *increase* correlation (collapsing support) while others should *decorrelate* factors (filling quadrants), but at higher cost or lower feasibility.

Correlation sweeps. For $d = 2$, we recommend an explicit sweep over $\rho \in \{0, 0.5, 0.8, 0.9, 0.95, 0.99\}$ for the passive design, and a parallel suite in which the platform can mix actions with different $\rho(a)$. For each setting, report:

$$\mathrm{MSE}(N) = \mathbb{E}\|\hat{w}_N - w_*\|_2^2, \qquad \widehat{\lambda}(N) = \lambda_{\min}(\widehat{I}_N),$$

as functions of N, and verify whether the passive curve exhibits the predicted blow-up as $\rho \to 1$ while the intervention-mixture curve remains stable (up to cost constraints). A useful diagnostic is to plot the empirical distribution of Δz and explicitly quantify "quadrant mass" in $d = 2$:

$$\widehat{p}_{s,t} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}\{\mathrm{sign}(\Delta z_{n,1}) = s,\ \mathrm{sign}(\Delta z_{n,2}) = t\}, \qquad s, t \in \{-1, +1\},$$

since nontrivial mass in all four quadrants is the simplest witness of identifiability in the discrete-factors intuition.

Correlation flips as an OOD stressor. A particularly alignment-relevant failure mode is that training data induces one correlation structure while deployment induces another. We can model this by training under $\Delta z \sim \mathcal{D}_{\text{train}}$ with correlation ρ_{train} and evaluating under $\mathcal{D}_{\text{test}}$ with correlation ρ_{test} (including sign flips). The empirical question is whether an intervention policy that maximizes $\lambda_{\min}$ under training prompts also produces a $\hat{w}$ that transfers when Σ_{test} changes. Concretely, we measure an OOD ranking loss or regret proxy such as

$$\text{Err}_{\text{test}} = \mathbb{P}_{(x,\Delta z)\sim\mathcal{D}_{\text{test}}}\left[\text{sign}(\hat{w}^\top \Delta z) \neq \text{sign}(w_*^\top \Delta z)\right],$$

and relate it to $\widehat{\lambda}(N)$ to test whether "more overlap during collection" actually predicts "less brittleness under shift."

Algorithmic ablations. Synthetic environments also let us separate conceptual mechanisms: (i) plug-in E-optimal vs random mixing; (ii) uncertainty-aware vs greedy plug-in; (iii) margin targeting (keeping $g(\hat{w}^\top \Delta z)$ away from saturation) vs no targeting; (iv) cost-aware vs cost-ignorant. The point is not to win a benchmark but to confirm which components are necessary to prevent early lock-in to a misleading design—a concrete safety concern when the system prematurely concludes that a delicate tradeoff is "already learned."

7.2 (ii) Real preference datasets with controlled edits: length/format/safety as manipulable factors

Goal. In real data we cannot observe z, and interventions T_a can fail or drift semantically. The objective therefore shifts: we want to show that controlled edits create *measurable* increases in proxy overlap and yield downstream gains without introducing confounds that invalidate labels.

Dataset construction around explicit intervention families. We recommend starting from prompts drawn from multiple risk regimes (benign assistance, policy-sensitive questions, and adversarial/jailbreak-style prompts) and generating base candidates from a fixed generator (or a small set of generators to increase diversity). Then define a small, auditable action set $\mathcal{A}$ whose transformations target interpretable factors:

- **Length / verbosity:** shorten vs expand while preserving core answer.
- **Format:** free-form vs structured (bullets, step-by-step plan, citations).
- **Safety posture:** add/remove a cautionary disclaimer; add policy-compliant refusal framing; add benign-alternative suggestions.

- **Uncertainty calibration:** confident vs hedged; include/exclude explicit assumptions.

Each T_a should come with an automatic validation check (e.g., constraint satisfaction, refusal template presence, toxicity filter thresholds) and a manual audit rate, since a frequent empirical pitfall is that the intervention intended to isolate one factor silently changes another.

Measuring overlap in proxy space. With a fixed feature map ϕ, we can track overlap and missing-direction risk via bucketed Fisher surrogates. For prompt buckets $b(x) \in [K]$, report:

$$\widehat{\lambda}^{(k)}(t) = \lambda_{\min}\left(\widehat{I}_t^{(k)}\right), \qquad \widehat{I}_t^{(k)} = \frac{1}{|\mathcal{S}_k|} \sum_{s \in \mathcal{S}_k} g(\hat{w}_t^\top \Delta\phi_s)\, \Delta\phi_s \Delta\phi_s^\top + \lambda I_d,$$

where $\mathcal{S}_k$ are samples in bucket k. The empirical claim we want is not merely that $\widehat{\lambda}$ increases, but that it increases *uniformly across buckets*, because safety failures often concentrate in underrepresented regions.

Audits for "illusory overlap." Proxy-based overlap can be misleading. We therefore propose targeted audits that are explicitly designed to falsify the "we covered the space" story: (i) sample pairs from directions corresponding to the smallest eigenvector of $\widehat{I}_t^{(k)}$ and have expert labelers judge whether the comparison is meaningful and on-policy; (ii) hand-construct "quadrant checks" for key factor pairs (e.g., long & safe vs short & unsafe, long & unsafe vs short & safe) and verify that the collected data contains all combinations at nontrivial rates; (iii) estimate intervention acceptance rates $\alpha(a, x)$ and report effective costs $c(a)/\alpha(a, x)$, since a design that relies on rarely-successful edits can look good on paper while failing operationally.

7.3 (iii) Downstream policy learning under OOD: DPO with stress-tested evaluation suites

Goal. Ultimately we care about the policies trained from these comparisons, not just about $\hat{w}$ or $\widehat{I}$. The core hypothesis is that improving overlap during collection yields downstream policies that are *less* sensitive to distribution shift and do not overfit to spurious correlations between, say, verbosity and perceived helpfulness or between refusal templates and perceived safety.

Training protocol. Using matched budgets, we train DPO-style policies on (a) passively collected comparisons and (b) actively designed comparisons (including controlled edits). We recommend holding constant: the base generator class, total labeled comparisons, and the labeler pool, so that differences can be attributed to the collection design rather than hidden capacity or annotation effects.

OOD evaluation axes. OOD should be multi-dimensional:

- **Prompt shift:** new domains, new jurisdictions/policies, different languages, and higher-adversariality prompts.
- **Factor shift:** systematic changes in correlations (e.g., deployment prompts where safe answers are necessarily longer, versus training prompts where length is cosmetic).
- **Evaluator shift:** alternative labelers or rubric changes, reflecting governance realities where standards evolve.

We then report both preference metrics (win rate under held-out comparisons, calibrated reward-model scores) and safety metrics (policy-violation rate, jailbreak success rate, harmful instruction compliance). The safety-relevant evaluation question is whether the active design reduces *tail* failure rates, not just average wins.

Linking overlap metrics to downstream outcomes. To make the results actionable, we should empirically connect overlap proxies to deployment behavior: regress OOD failure probability on bucket-level $\widehat{\lambda}^{(k)}$ (and on acceptance-adjusted action diversity) to test whether low-overlap buckets predict where the policy breaks. If such a link holds even weakly, it motivates a governance-friendly operational rule: specify minimum overlap targets per risk bucket as a precondition for deployment, rather than treating data collection as an unstructured scaling exercise.

Limitations and open empirical risks. Even a clean win on these experiments would not imply that $\lambda_{\min}$ is the only relevant design criterion: real labelers are nonstationary, preferences can be context-dependent, and high-overlap data can still encode the wrong objective if the action set systematically excludes morally salient counterfactuals. Conversely, a failure to see gains may reflect representation misspecification in ϕ rather than a flaw in overlap-driven design. This is precisely why we advocate the staged blueprint above: it separates "the principle is wrong" from "the proxy is wrong" and makes the remaining gaps explicit enough to govern.

8 Discussion and 2026 implications: governance, benchmarks, audits, and operational overlap targets

The empirical blueprint above treats overlap as a measurable bottleneck for identifying safety–utility tradeoffs. The 2026-relevant question is how to translate that observation into *institutional* and *operational* practice: what

should be logged, what should be audited, what should be guaranteed before deployment, and what should be benchmarked so that "we collected more preference data" does not become a substitute for demonstrating identifiability and robustness.

Data governance as experimental-design governance

Preference datasets are increasingly regulated and internally controlled not only as privacy-sensitive artifacts but as *decision-critical evidence* about normative objectives. From that perspective, the platform is not merely sampling i.i.d. comparisons; it is running a controlled experiment whose design determines which value-laden tradeoffs are learnable. This framing suggests two concrete governance upgrades.

First, we should treat the intervention policy $q(a \mid x)$ as a governed object, akin to a training-time "policy lever" that requires review. In 2026 deployments, it is common to iterate on prompts, rubrics, and model policies weekly; without explicit control, these iterations can silently change the induced Δz distribution and invalidate prior conclusions about identifiability. A minimal governance artifact is an *intervention registry* that records, for each action family: (i) its semantic intent (which factor it is supposed to vary), (ii) its known confounds, (iii) its acceptance rate $\alpha(a, x)$ and failure modes, and (iv) its cost model $c(a)$ including expert-labeler requirements. This is not busywork: if we cannot say which interventions are responsible for filling which missing directions, we cannot later attribute robustness (or failures) to controllable causes.

Second, we should make overlap a first-class dataset contract. Today, dataset documentation often lists domains, languages, and toxicity prevalence. For preference learning, we additionally want a statement of *identifiability coverage*: in each risk-relevant bucket (e.g. benign assistance, self-harm, bio, cyber, jailbreak), what is the minimum overlap achieved, operationalized as a proxy for $\lambda_{\min}(I_q)$ (or a conservative surrogate) and tracked over time? In other words, we want dataset "nutrition labels" that include not only *what* content exists but *which tradeoffs are learnable* from it.

Benchmarks that punish non-identification, not only low averages

A benchmark that only reports average win rates can be satisfied by collecting comparisons along a narrow manifold (highly correlated factors) and then fitting a reward model that performs well on that manifold while failing under modest correlation shifts. If we accept the premise that non-identification is an alignment risk, then benchmarks should be designed to detect it.

One practical direction is to incorporate *counterfactual edit suites* as benchmark primitives. Rather than only testing on naturally sampled model

outputs, the benchmark itself can include standardized controlled edits (length, format, refusal framing, uncertainty calibration) that intentionally generate "quadrant" comparisons. A system that genuinely learned the intended tradeoff should behave consistently across these edits; a system that relied on spurious correlations should fail systematically (e.g. overvaluing verbosity, overvaluing template-like refusals, or penalizing calibrated uncertainty).

A second direction is to formalize *correlation-shift evaluation* as a requirement. Benchmarks can include paired test distributions whose marginal task content is similar but whose factor correlations differ (including sign flips). Passing then requires not only high average preference on the in-distribution split but bounded degradation under the shifted split. This makes explicit what is otherwise an informal fear: that the learned objective is "right" only under a particular training-time entanglement of factors.

Finally, we should expect benchmark designers to publish *coverage diagnostics* alongside scores, in the same way robustness benchmarks publish per-slice results. For preference benchmarks, that means reporting proxy overlap metrics (e.g. eigenvalues of second-moment matrices in representation space) and ensuring that benchmark construction does not itself collapse to a single style axis.

Audit requirements: from "label quality" to "design adequacy"

Current audits focus on labeler agreement, policy compliance, and demographic harms. Those remain necessary, but they do not address the design-level failure mode emphasized here: collecting many labels that are uninformative about key directions in w_*. We therefore see a need for a new audit category: *design adequacy audits*.

A design adequacy audit answers questions of the form: "Could we have learned the tradeoff we claim to have learned, given the comparisons we collected?" Concretely, an audit can (i) compute bucket-level overlap surrogates, (ii) identify the minimum-eigenvector directions (the "missing" directions), and (iii) sample or construct comparisons that are intended to load on those directions. Expert evaluators can then judge whether those comparisons are meaningful and whether the platform has a feasible path to collecting them at scale. Importantly, these audits are not purely statistical: they test whether interventions preserve semantics sufficiently that a factor-isolating comparison is actually interpretable by labelers.

We also expect audits to incorporate *effective-cost accounting*. In production, what matters is not nominal $c(a)$ but $c(a)/\alpha(a, x)$ and the induced latency and operational complexity. A design that achieves excellent overlap only by relying on interventions that fail 70% of the time (or require scarce expert review) is fragile in the same way that a security control is fragile if it is too expensive to use consistently. Reporting acceptance-adjusted costs, per bucket, makes this fragility legible to governance stakeholders.

Operationalizing overlap targets as production SLOs

To make overlap actionable, we need an operational analogue of "maintain $\lambda_{\min}(I_q) \geq \underline{\lambda}$" that fits modern ML production: online monitoring, alerting, and rollbacks. We propose treating overlap as a service-level objective (SLO) with three layers.

(1) Pre-deployment gating. Before training a new reward model or preference policy on a dataset slice, require that each risk bucket meets a minimum overlap threshold in proxy space, e.g. $\widehat{\lambda}^{(k)} \geq \tau_k$, where τ_k is stricter for high-stakes buckets. This is an explicit commitment to "no training on non-identifying data" for safety-critical domains. Where thresholds cannot be met due to feasibility, the system should document the missing directions and avoid claims of robustness that depend on them.

(2) Online drift monitoring. After deployment, monitor proxies for correlation drift: changes in second moments of $\Delta\phi$ and in the distribution of margins $\hat{w}^\top \Delta\phi$ (to detect saturation). When drift is detected, trigger targeted recollection using interventions that are known to fill the relevant directions. This turns what is often a reactive incident response into a controlled maintenance loop.

(3) Budget allocation as a control problem. If we accept that overlap is costly, then the core operational question becomes allocation: where should we spend the next unit of labeling budget? A governance-friendly implementation is a "knapsack with coverage" planner that prioritizes buckets whose overlap proxy is lowest relative to threshold, adjusted by estimated harm. This is a natural place to embed cost-aware E-optimal design ideas without forcing every product team to reason about Fisher information directly.

Limitations and extensions

Several limitations are important for correct interpretation and for setting 2026 research priorities.

Proxy misspecification remains the central risk. All practical overlap metrics rely on a representation ϕ (or an approximate $\hat{z}$). If ϕ omits morally salient factors, then maximizing proxy overlap can create a false sense of security. The right response is not to abandon overlap metrics, but to treat them as *auditable hypotheses*: we should regularly probe whether the "missing directions" in proxy space correspond to genuine semantic tradeoffs, and we should diversify ϕ (e.g. multiple encoders, evaluator models, and rule-based features) to reduce single-point failure.

Nonlinear and nonstationary preferences. Linear reward models are useful because they make identifiability and information geometry explicit. Real labelers exhibit context dependence, nonstationarity, and occasionally strategic or policy-conditioned behavior. Extending the design logic to generalized linear or nonparametric reward models is conceptually straightforward but operationally harder: overlap must then be defined relative to local curvature or to function-class complexity, not just $\lambda_{\min}$ of a fixed matrix. A promising intermediate step is to adopt local linearization (piecewise models) with bucket-level overlap targets.

Safety constraints can conflict with quadrant filling. Some "quadrants" correspond to content we do not want to generate or show to labelers (e.g. unsafe long-form instructions). This creates a real tension: identification may demand counterfactuals that are themselves hazardous. In 2026 governance regimes, we expect this to be handled by constrained design: only fill quadrants within an approved action set, and treat the remaining missing directions as an explicit uncertainty that must be compensated by other controls (policy rules, refusal enforcement, or expert-only red teaming). Put differently, overlap is not a license to collect anything; it is a way to precisely state what we *cannot* learn under safety constraints.

From single w to heterogeneous objectives. Different user groups, jurisdictions, and policy regimes imply different effective w's. The minimax framing suggests overlap targets should be met for the worst-off group, but this can be expensive. A practical extension is to treat group membership (or jurisdiction) as part of x and impose per-group overlap SLOs, with explicit budget tradeoffs and escalation paths when a group cannot be covered sufficiently.

Overall, the key 2026 implication is that preference-data collection should be governed like an experiment whose design determines what is learnable. Overlap metrics provide a concrete bridge between abstract identifiability and the day-to-day realities of intervention engineering, labeling budgets, and deployment audits. If we make that bridge explicit—in benchmarks, in monitoring, and in governance artifacts—we reduce the probability that "more RLHF" becomes an unexamined substitute for demonstrating robustness where it matters.