VALIDATION GEOMETRY FOR REWARD OPTIMIZATION UNDER PROXY MISSPECIFICATION

Mu Chen

ABSTRACT

Finite validation does not reveal which unmeasured proxy errors a specified optimizer will amplify. We characterize this dependence through an optimizer-conditioned geometry. For any fixed endpoint, including one returned by a nonlinear or proxy-dependent optimizer, the centered density ratio relative to the reference distribution is the exact sensitivity direction for proxy overstatement. The worst undetectable bias equals the error budget times the unmeasured component of this sensitivity. It vanishes for every admissible residual exactly when validation spans that direction. Under a finite validation budget, average risk, restricted worst case, and operational measurement design lead to different optimization problems. In our experiments, validators chosen for a training distribution transfer under interpolation but fail after covariance rotation or the appearance of unseen optimizer directions. Feasible dictionary designs also remain far from the unrestricted spectral lower bound. These results certify a declared endpoint; they do not guarantee true-reward recovery or universal safety.

1 INTRODUCTION

Reward optimization replaces objectives such as human preference, factual correctness, safety, and task utility with a tractable proxy. This substitution underlies preference-based reinforcement learning and language-model alignment (Christiano et al., 2017; Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022; Rafailov et al., 2023). It also creates a familiar failure mode: further optimization may improve the proxy after the intended objective has stopped improving, or has begun to worsen (Manheim & Garrabrant, 2018; Pan et al., 2022; Skalse et al., 2022; Gao et al., 2023).

Existing work asks whether a proxy is hackable, how overoptimization scales, or how an optimizer should be regularized. We instead ask: *what has a finite validation protocol ruled out for a specified optimizer?* Answering this requires the geometry between two families of directions: proxy errors left invisible by validation and directions to which the optimizer's endpoint distribution is sensitive. A scalar validation score does not contain this information.

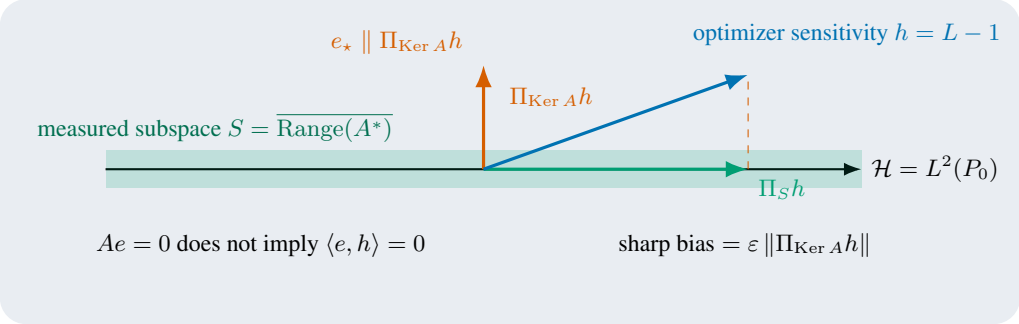

Figure 1: Finite validation certifies only the component of optimizer sensitivity in its measured subspace. The orthogonal component is the exact direction along which an admissible proxy residual can create worst-case bias.

Our formulation separates four questions that are often conflated. First, for a *fixed optimizer endpoint*, how large can the proxy overstatement be after all validation constraints pass? Second, when is that

bias identifiable? Third, if only m measurements are affordable, which subspace should be measured? Fourth, do measurements designed for one optimizer family transfer to held-out optimizers? Figure 1 gives the answer to the first two and foreshadows why the latter two require explicit quantifiers.

We make four contributions:

- The exact endpoint sensitivity of an arbitrary proxy-conditioned optimizer is $dP_R/dP_0 - 1$. This identity reduces proxy overstatement to a residual-sensitivity pairing without linearizing the optimizer. The sharp projection radius and identifiability condition then follow directly.
- For validation design, we distinguish an eigenspace solution for average squared bias from a Kolmogorov-width problem over a restricted worst-case class. Neither result solves measurement selection under dictionary, cost, or noise constraints.
- Best-of-N (BoN), exponential tilting, and local policy-gradient directions admit the same sensitivity representation. We also give the exact atomwise BoN likelihood ratio for discrete scores and state the variance-only role of action-independent baselines.
- Experiments test the equalities and their failure boundaries. They include exact kernel witnesses, identifiability transitions, held-out optimizer shifts, noisy certificates, four frozen candidate-pool cells, and a multi-checkpoint trajectory. Negative and reverse-alignment cells remain in the reported results.

The projection argument is classical. The modeling step is to identify $dP_R/dP_0 - 1$ as the exact sensitivity of a realized proxy-conditioned endpoint. Once this direction is explicit, validating a proxy and validating the optimizer that acts on it become different questions. Average, worst-case, and feasible-design claims also require different quantifiers.

2 Problem Setup: Validation Meets Optimization

Let $(\mathcal{X}, \mathcal{F}, P_0)$ be a reference distribution and $\mathcal{H} = L^2(P_0)$ with inner product $\langle f, g\rangle = \mathbb{E}_{P_0}[fg]$. Let $R \in \mathcal{H}$ be an observed proxy reward and $U \in \mathcal{H}$ the intended utility. We write

$$U = R - e, \qquad e := R - U, \tag{1}$$

so positive e means that the proxy overstates utility.

Validation consists of a bounded linear operator $A : \mathcal{H} \to \mathbb{R}^m$. The idealized exact-pass uncertainty set is

$$\mathcal{E}(A, \varepsilon) = \{e \in \mathcal{H} : Ae = 0,\ \|e\| \le \varepsilon\}. \tag{2}$$

The set models finite checks, moments, or directions known to be accurate. Section D.5 treats noise and approximate passing. The exact set isolates the structural question.

An optimization rule Ω observes R and returns an endpoint distribution $P_R = \Omega(R)$. Assume $P_R \ll P_0$ and let

$$L_R = \frac{dP_R}{dP_0}, \qquad h_R = L_R - 1 \in \mathcal{H}. \tag{3}$$

Because $\mathbb{E}_{P_0}[L_R] = 1$, h_R is centered. We study the proxy's overstatement of the improvement achieved by this *same endpoint*:

$$B_\Omega(R, U) = \big(\mathbb{E}_{P_R}[R] - \mathbb{E}_{P_0}[R]\big) - \big(\mathbb{E}_{P_R}[U] - \mathbb{E}_{P_0}[U]\big). \tag{4}$$

This is not counterfactual regret between separately reoptimized proxy and true-reward policies.

3 Exact Endpoint Validation Geometry

Theorem 1 (Endpoint identity and sharp invisible bias)**.** *Fix an observed proxy R and its endpoint $P_R = \Omega(R)$. Under Equations (1) and (3),*

$$B_\Omega(R, U) = \langle e, h_R\rangle\,. \tag{5}$$

Let $K = \mathrm{Ker}(A)$. Then

$$\sup_{e \in \mathcal{E}(A,\varepsilon)} |B_\Omega(R, U)| = \varepsilon\,\|\Pi_K h_R\|\,. \tag{6}$$

Whenever $\Pi_K h_R \neq 0$, both signs are attained by $e_\star = \pm\varepsilon \Pi_K h_R / \|\Pi_K h_R\|$.

The upper bound is attained by a residual that passes validation and respects the norm budget, so it is more than a Cauchy-Schwarz estimate. The endpoint identity makes the proof short. Its implication is specific: validation protects an optimizer only against errors in measured directions. Generic proxy accuracy is neither necessary nor sufficient.

Corollary 2 (Optimizer-specific identifiability)**.** *The bias of a fixed endpoint is zero for every $e \in K$ if and only if*

$$h_R \in K^\perp = \overline{\mathrm{Range}(A^*)}. \tag{7}$$

For finite-dimensional validation, $\mathrm{Range}(A^*)$ *is closed.*

Why the endpoint bridge matters. Theorem 1 does not linearize the optimizer. Ω may be nonlinear, iterative, or explicitly dependent on R. After conditioning on the observed proxy and returned endpoint, all optimization details enter through L_R. For a trajectory $P_0, P_1, \ldots, P_K$, the per-step bias telescopes to the same endpoint expression with $L_K = dP_K/dP_0$ whenever common support holds. The theorem therefore covers finite-step optimization without assuming an optimizer independent of the proxy.

Local approximation. For a differentiable policy family π_θ and direction d, a small update $\theta + \eta d$ has

$$L_\eta - 1 = \eta h_{\theta,d} + r_{\eta,d}, \qquad h_{\theta,d}(x,a) = \langle \nabla_\theta \log \pi_\theta(a \mid x), d \rangle\,, \tag{8}$$

and, under a bounded second derivative in $L^2(P_0)$, $\|r_{\eta,d}\| \le M\eta^2/2$. Therefore $B_\eta = \eta \langle e, h_{\theta,d} \rangle + O(\varepsilon\eta^2)$, while the endpoint identity remains exact whenever L_η is available. An action-independent baseline has zero population pairing with the score and changes estimator variance, not this first-order bias geometry.

3.1 Approximate and noisy validation remains sharp

Exact passing isolates the null-space obstruction, but finite validation usually bounds rather than eliminates the visible residual. Let $S = \overline{\mathrm{Range}(A^*)}$ and suppose $\|e\| \le \varepsilon$ and $\|\Pi_S e\| \le \delta$. Write $p = \|\Pi_S h\|$, $q = \|\Pi_{S^\perp} h\|$, $r = \|h\|$, and $d = \min\{\delta, \varepsilon\}$.

Theorem 3 (Sharp approximate-pass envelope)**.** *For $h \neq 0$ and $\varepsilon > 0$,*

$$B_\delta(h) := \sup_{\substack{\|e\| \le \varepsilon \\ \|\Pi_S e\| \le \delta}} |\langle e, h \rangle| = \begin{cases} \varepsilon r, & d \ge \varepsilon p / r, \\ dp + \sqrt{\varepsilon^2 - d^2}\, q, & d < \varepsilon p / r. \end{cases} \tag{9}$$

Both branches are attained. In particular, $B_0(h) = \varepsilon q$, recovering Theorem 1.

The formula exposes a non-vacuity boundary: $\delta < \varepsilon$ alone does not improve on the no-validation radius; validation helps strictly only if $p > 0$ and $\delta < \varepsilon p/r$. With an orthonormal basis of S, n independent prompt blocks, and coordinate-wise σ-sub-Gaussian noise, a simultaneous visible-radius estimate is

$$r_n(\alpha) = \sigma \sqrt{\frac{2m \log(2m/\alpha)}{n}}, \qquad \widehat{\delta} = \min\{\varepsilon, \left\|\widehat{\theta}\right\| + r_n(\alpha)\}. \tag{10}$$

With probability at least $1 - \alpha$, $|\langle e, h \rangle| \le B_{\widehat{\delta}}(h)$ holds simultaneously for all h. A pass threshold controls the joint event of passing and false certification. It does not give conditional coverage among passes, which may be rare. Section D.5 tests this difference.

4 Validation Design Under a Budget

Suppose a designer can choose an m-dimensional measured subspace $S \subseteq \mathcal{G}$, where $\mathcal{G} \subseteq \mathcal{H}$ is a closed subspace encoding admissible linear measurements. For an optimizer sensitivity h, the squared sharp radius at $\varepsilon = 1$ is $\|\Pi_{S^\perp} h\|^2$. The choice of quantifier over optimizers changes the mathematical problem.

4.1 AVERAGE-RISK DESIGN HAS A SPECTRAL SOLUTION

Let $T \sim \mu$ index a frozen distribution of optimizer sensitivities $h_T \in \mathcal{H}$, and define the uncentered second-moment operator

$$C_\mu = \mathbb{E}_{T\sim\mu}[h_T \otimes h_T]. \tag{11}$$

Theorem 4 (Optimal average-risk validation)**.** *Assume C_μ is trace class and $\dim(\mathcal{G}) \geq m$. Among subspaces $S \subseteq \mathcal{G}$ with $\dim(S) = m$, the minimizers of*

$$\mathcal{R}_\mu(S) = \mathbb{E}_{T\sim\mu} \left\|\Pi_{S^\perp} h_T\right\|^2 \tag{12}$$

are spans of the leading m eigenfunctions of $\Pi_\mathcal{G} C_\mu \Pi_\mathcal{G}$. If their eigenvalues are $\lambda_1 \geq \lambda_2 \geq \cdots$, then

$$\min_S \mathcal{R}_\mu(S) = \mathbb{E} \left\|\Pi_{\mathcal{G}^\perp} h_T\right\|^2 + \sum_{j>m} \lambda_j. \tag{13}$$

The compressed operator is the second moment of *optimizer sensitivity*, rather than a data covariance. The resulting principal-subspace solution has no distribution-free transfer guarantee. If future optimizers rotate into a low-variance or unseen subspace, a training-optimal design can fail sharply.

4.2 WORST-CASE AND OPERATIONAL DESIGN ARE DIFFERENT PROBLEMS

For a bounded optimizer class $\mathcal{K} \subset \mathcal{G}$, the worst-case design value is the constrained Kolmogorov width

$$d_m(\mathcal{K}; \mathcal{G}) = \inf_{\substack{S\subseteq\mathcal{G} \\ \dim(S)=m}} \sup_{h\in\mathcal{K}} \left\|\Pi_{S^\perp} h\right\|. \tag{14}$$

There is no general covariance-eigenspace solution because an average over μ has been replaced by a supremum over $\mathcal{K}$.

Proposition 5 (Two sharp boundaries)**.** *If $\mathcal{K}$ is the unit sphere of an r-dimensional tangent space, then $d_m(\mathcal{K}) = 1$ for $m < r$ and 0 for $m \geq r$. If instead $\mathcal{K}$ is the compact ellipsoid with semiaxes $\sqrt{\lambda_1} \geq \sqrt{\lambda_2} \geq \cdots$, then $d_m(\mathcal{K}) = \sqrt{\lambda_{m+1}}$.*

In practice, validation directions may come from a dictionary, carry different costs, or share prompt-block noise. A validator may also issue a certificate only after a noisy pass. These constraints produce a combinatorial or stochastic design problem. The spectral subspace in Theorem 4 is then an oracle lower bound rather than an operational prescription.

5 ONE GEOMETRY, SEVERAL REWARD OPTIMIZERS

Table 1: Optimizer-specific sensitivities in the common endpoint geometry. For BoN, V is a continuous or randomized-refined rank; discrete scores use the atomwise expression in Section C.

Optimizer	Sensitivity h	Consequence
Best-of-N	$h_N(V) = NV^{N-1} - 1$ (continuous or randomized-refined rank)	Changing N changes the blind directions that validation must cover.
Exponential tilt	$h_\beta(x) = \exp(\beta R(x))/Z_\beta - 1$	KL/tilt strength traces a smooth endpoint path in $\mathcal{H}$.
Policy-gradient step	$h_{\theta,d}(x,a) = \langle \nabla \log \pi_\theta(a \mid x), d\rangle$	Local bias is the residual-score pairing; endpoint $L-1$ is exact.
ReMax baseline	Same population $h_{\theta,d}$ for any action-independent baseline	Baseline centering can reduce variance without changing first-order bias.

For continuous scores, or after randomized refinement within score atoms, the selected sample's likelihood ratio is NV^{N-1}. A genuinely discrete score instead requires the finite difference of F^N across the selected atom; Section C gives the formula. In inference-time scaling analyses (Beirami et al., 2025; Huang et al., 2025; Khalaf et al., 2025; Balashankar et al., 2025; Sriraman & Block, 2026; Paes et al., 2026), validation risk depends on the component of h_N left outside the validation

span, not on optimization strength alone. Exponential tilting covers KL-regularized and reward-guided endpoints, including test-time alignment constructions (Xu et al., 2025; Lin et al., 2025). Policy-gradient geometry describes a local training direction, while the endpoint identity applies beyond the local approximation. ReMax (Li et al., 2024) motivates the baseline distinction: an action-independent baseline can reduce variance but cannot make an invisible systematic residual visible.

6 Empirical Tests of the Theory

The experiments target falsifiable equalities and boundary cases rather than leaderboard gains. We test whether a validation design chosen from observed optimizer sensitivities remains protective for future endpoints and under operational measurement constraints. Section D gives the full protocols, ablations, and negative results.

Exact witnesses and phase transitions. In a 64-dimensional numerical Hilbert space, we construct $e_\star \in \mathrm{Ker}(A)$ analytically. Across BoN, tilt, and policy-gradient families, every nonzero sharp radius is attained to numerical precision, with negligible validation leakage. For six frozen directions per family, identifiability occurs exactly when their span is covered. Before that transition, the residual curves differ: policy-gradient directions retain a complete worst-case blind direction until $m = 6$, whereas BoN and tilt concentrate into fewer dominant modes.

Average design and held-out optimizers. Leading eigenspaces dominate random and ordinary moment heuristics on the frozen average-risk objective, as predicted by Theorem 4. This advantage transfers under interpolation and matched covariance. At low m, it fails under covariance rotation and on a genuinely unseen policy-tangent subspace (Figure 2). The theorem therefore does not imply universal generalization.

Operational measurement design. On the frozen 3B-GSM8K pool, a six-atom dictionary leaves a measurable gap between feasible and unrestricted designs (Table 2). At $m = 1$, exhaustive feasible selection reduces held-out average squared radius from 1.2321 to $.7600$, while the unrestricted lower bound is $.2783$. Using all six atoms still leaves $.6268$, compared with $.0031$ for the unrestricted design. Spectral optimality alone is therefore not an implementable prescription.

Table 2: Held-out average squared radius under a six-atom measurement dictionary. The arbitrary eigenspace is an oracle relaxation, not an operational design.

Design	$m = 1$	$m = 2$	$m = 3$	$m = 6$
Ordinary heuristic	1.2321	1.2018	1.1734	.6268
Dictionary exact optimum	.7600	.6326	.6274	.6268
Oracle arbitrary eigenspace	.2783	.1309	.0551	.0031

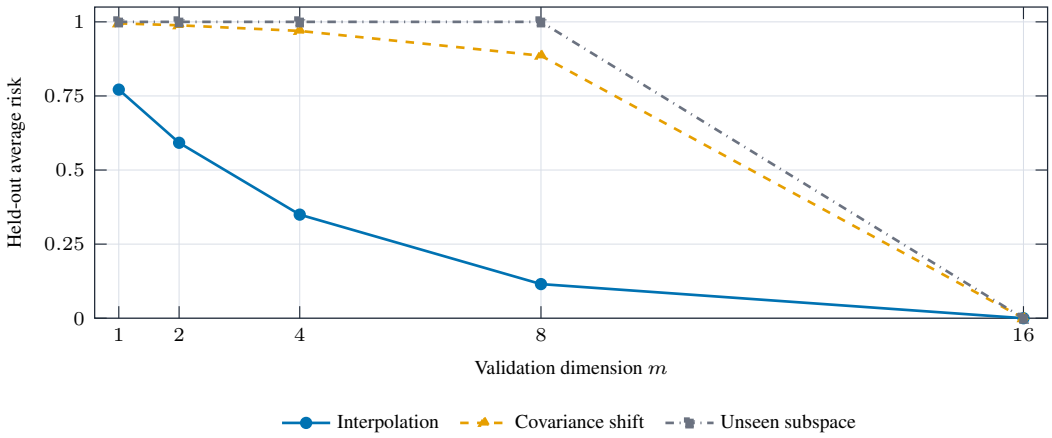

Figure 2: Held-out policy-tangent risk. The training top eigenspace transfers under interpolation, but low-dimensional designs fail under covariance rotation and a new subspace. Full rank $m = 16$ covers all three declared cases.

Real frozen candidate pools. We preregistered four cells: Qwen2.5-3B/7B-Instruct on GSM8K and MATH-500, with 256 held-out prompts and 16 sampled candidates per prompt. Gold labels remained hidden until we froze the candidates, proxies, optimizer grids, tie rules, and hashes. We evaluated two gold-free proxies, self-consistency and mean-logprob rank, for BoN $N \in \{1, 2, 4, 8, 16\}$ and tilt $\beta \in \{0, .5, 1, 2, 4, 8\}$. Signed alignment depends strongly on the proxy. Self-consistency is mostly null or reverse at BoN-16, whereas mean-logprob rank has positive overstatement in all four cells (Section E). Only the 3B-GSM8K mean-logprob cell shows materially negative utility change, so a positive gap does not by itself establish harmful reward hacking. The appendix reports agreement between the direct and geometric forms across all 88 endpoints as an implementation check.

Noise and finite optimization. A preregistered finite-sample suite tests a joint noisy-operator certificate over 1,944 cells. Removing the sampling radius raises the violation rate as high as 0.5093. Conditional-on-pass coverage can also fail even when unconditional joint control holds. In a four-checkpoint Qwen2.5-0.5B trajectory, the direct and geometric endpoint gaps agree below 10^{-16}. The local tangent remainder is non-negligible. An action-independent baseline leaves the mean unchanged and reduces variance to 0.0777 of the uncentered estimator.

Table 3: Decision-relevant theorem obligations. These are diagnostic tests, not performance benchmarks.

Claim under test	Primary falsifier	Result
Sharp radius is attainable	$\lvert \langle e_\star, h \rangle \rvert / (\varepsilon \lVert \Pi_K h \rVert)$	1.000000; leakage at numerical zero.
Average design is not universal	Held-out risk under shifted directions	Passes interpolation; fails covariance/unseen-subspace shift at low m.
Sampling uncertainty is necessary	Violation after deleting sampling radius	Naive control: maximum violation 0.5093.
Baseline affects variance, not mean	Mean change and variance ratio	Mean difference 0; variance ratio 0.0777.

7 Related Work

Reward misspecification and reward hacking. Prior work classifies Goodhart-style failures (Manheim & Garrabrant, 2018), formalizes relationships between proxy and true reward (Skalse et al., 2022; Laidlaw et al., 2025), and measures misspecification as agent capability or optimization pressure increases (Pan et al., 2022; Gao et al., 2023). RewardBench evaluates reward-model discrimination

(Lambert et al., 2024). Studies of inference-time selection show that selection can amplify reward-model errors (Khalaf et al., 2025). We condition on a validation operator and a specified optimizer endpoint, then ask which residual directions can still produce bias.

Optimal recovery, experimental design, and robustness. The null-space perspective is classical in inverse problems and optimal recovery (Engl et al., 1996; Binev et al., 2017). Optimal experimental design chooses measurements to reduce estimation uncertainty (Pukelsheim, 2006), and reward-learning design has been studied in doubly nonparametric settings (Bhatia et al., 2023). Robust optimization studies uncertainty sets and worst-case objectives (Ben-Tal & Nemirovski, 1998; Bertsimas & Sim, 2004). Our use of these tools distinguishes an average over frozen optimizer sensitivities, a supremum over a declared optimizer class, and selection from an operational measurement dictionary.

Widths and active subspaces. Kolmogorov widths quantify the best low-dimensional approximation of a set (Pinkus, 1985), while active subspaces use dominant eigen-directions for dimension reduction (Constantine, 2015). Here, the set or second-moment operator is built from optimizer likelihood-ratio sensitivities rather than generic inputs. A class width gives a minimax quantity; a distribution spectrum gives an average-risk quantity.

8 Limitations and Conclusion

Validation geometry is a conditional certificate. It does not discover the true reward. The exact theorem assumes a common reference measure, an L^2 residual budget, and a fixed observed proxy and endpoint. Other norms produce different dual geometries, and support violations can make endpoint likelihood ratios unusable. The spectral design theorem assumes a frozen optimizer distribution, so it does not promise transfer under optimizer shift. In the real candidate pools, utility is exact mathematical correctness on finite support. These experiments do not establish human-preference alignment, unrestricted online training behavior, or ubiquitous reward hacking. Measurement design with prompt costs and joint noise remains problem-specific.

Within these limits, finite validation protects only the optimizer-sensitive directions that it spans. The invisible projection gives the sharp worst-case bias, and identifiability reduces to a range condition. Validation design must therefore state whether it targets average risk, a worst-case optimizer class, or a feasible measurement set. Under this view, proxy validation is an operator-design problem rather than a scalar accuracy check.

References

Ananth Balashankar, Ziteng Sun, Jonathan Berant, Jacob Eisenstein, Michael Collins, et al. Infalign: Inference-aware language model alignment. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 2025. URL https://proceedings.mlr.press/v267/balashankar25a.html.

Ahmad Beirami, Alekh Agarwal, Jonathan Berant, Alexander D'Amour, Jacob Eisenstein, Chirag Nagpal, and Ananda Theertha Suresh. Theoretical guarantees on the best-of-n alignment policy. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 2025. URL https://proceedings.mlr.press/v267/beirami25a.html.

Aharon Ben-Tal and Arkadi Nemirovski. Robust convex optimization. *Mathematics of Operations Research*, 23(4):769–805, 1998. doi: 10.1287/moor.23.4.769.

Dimitris Bertsimas and Melvyn Sim. The price of robustness. *Operations Research*, 52(1):35–53, 2004. doi: 10.1287/opre.1030.0065.

Kush Bhatia, Wenshuo Guo, and Jacob Steinhardt. Reward learning as doubly nonparametric bandits: Optimal design and scaling laws. In *International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 11149–11171, 2023. URL https://proceedings.mlr.press/v206/bhatia23a.html.

Peter Binev, Albert Cohen, Wolfgang Dahmen, Ronald DeVore, Guergana Petrova, and Przemyslaw Wojtaszczyk. Data assimilation in reduced modeling. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):1–29, 2017. doi: 10.1137/15M1025384.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1706.03741.

Paul G. Constantine. *Active Subspaces: Emerging Ideas for Dimension Reduction in Parameter Studies*. SIAM, 2015. doi: 10.1137/1.9781611973860.

Heinz W. Engl, Martin Hanke, and Andreas Neubauer. *Regularization of Inverse Problems*. Kluwer Academic Publishers, 1996. doi: 10.1007/978-94-009-1740-8.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Audrey Huang, Adam Block, Qinghua Liu, Nan Jiang, Akshay Krishnamurthy, and Dylan J. Foster. Is best-of-n the best of them? coverage, scaling, and optimality in inference-time alignment. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 2025. URL https://proceedings.mlr.press/v267/huang25c.html.

Hadi Khalaf, Claudio Mayrink Verdun, Alex Oesterling, Himabindu Lakkaraju, and Flavio du Pin Calmon. Inference-time reward hacking in large language models. In *Advances in Neural Information Processing Systems*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/590a0cc0306c1c63e2d66a51a407718f-Abstract-Conference.html.

Cassidy Laidlaw, Shivam Singhal, and Anca D. Dragan. Correlated proxies: A new definition and improved mitigation for reward hacking. *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=msEr27EejF.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, et al. Rewardbench: Evaluating reward models for language modeling. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2403.13787.

Ziniu Li, Tian Xu, Yushun Zhang, Zhihang Lin, Yang Yu, Ruoyu Sun, and Zhi-Quan Luo. Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024. URL https://proceedings.mlr.press/v235/li24cd.html.

Baijiong Lin, Weisen Jiang, Yuancheng Xu, Hao Chen, and Ying-Cong Chen. Parm: Multi-objective test-time alignment via preference-aware autoregressive reward model. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 2025. URL https://proceedings.mlr.press/v267/lin25h.html.

David Manheim and Scott Garrabrant. Categorizing variants of goodhart's law. *arXiv preprint arXiv:1803.04585*, 2018. URL https://arxiv.org/abs/1803.04585.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2203.02155.

Lucas Monteiro Paes, Natalie Mackraz, Barry-John Theobald, and Federico Danieli. Theoretical limits of language model alignment. *arXiv preprint arXiv:2605.07105*, 2026. URL https://arxiv.org/abs/2605.07105.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2201.03544.

Allan Pinkus. *n-Widths in Approximation Theory*. Springer, 1985. doi: 10.1007/978-3-642-69894-1.

Friedrich Pukelsheim. *Optimal Design of Experiments*. SIAM, 2006. doi: 10.1137/1.9780898719109.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.18290.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. In *Advances in Neural Information Processing Systems*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/3d719fee332caa23d5038b8a90e81796-Paper-Conference.pdf.

Ved Sriraman and Adam Block. Revisiting the (sub)optimality of best-of-n for inference-time alignment. In *Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, 2026. URL https://proceedings.mlr.press/v336/sriraman26a.html.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2009.01325.

Yuancheng Xu, Udari Madhushani Sehwag, Alec Koppel, Sicheng Zhu, Bang An, Furong Huang, and Sumitra Ganesh. Genarm: Reward guided generation with autoregressive reward model for test-time alignment. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/b7a55b1fa8f2500299e1af3562a1191d-Abstract-Conference.html.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019. URL https://arxiv.org/abs/1909.08593.

A Proofs for Endpoint Validation Geometry

A.1 Proof of Theorem 1

Using $U = R - e$ in Equation (4),

$$\begin{aligned} B_\Omega(R, U) &= \mathbb{E}_{P_R}[e] - \mathbb{E}_{P_0}[e] \\ &= \mathbb{E}_{P_0}[(L_R - 1)e] = \langle e, h_R \rangle , \end{aligned}$$

which proves Equation (5). If $e \in K$, orthogonal decomposition gives $\langle e, h_R \rangle = \langle e, \Pi_K h_R \rangle$. Hence

$$| \langle e, h_R \rangle | \leq \|e\| \, \|\Pi_K h_R\| \leq \varepsilon \, \|\Pi_K h_R\| .$$

When $\Pi_K h_R \neq 0$, substitute $e_\star = \pm \varepsilon \Pi_K h_R / \|\Pi_K h_R\|$. It lies in K, has norm ε, and attains both signs of the upper bound. If the projection is zero, both sides vanish.

A.2 Proof of Corollary 2

The bias vanishes for every $e \in K$ if and only if $h_R \perp K$, i.e., $h_R \in K^\perp$. For any bounded operator between Hilbert spaces,

$$\mathrm{Ker}(A)^\perp = \overline{\mathrm{Range}(A^*)}.$$

Since A maps into $\mathbb{R}^m$, $\mathrm{Range}(A^*)$ is finite-dimensional and therefore closed. Hence

$$\mathrm{Ker}(A)^\perp = \mathrm{Range}(A^*).$$

A.3 Finite trajectories and proxy dependence

Let $P_k \ll P_0$ and $L_k = dP_k/dP_0$. Applying the endpoint identity to each consecutive pair and summing gives

$$\sum_{k=1}^{K} \left(\mathbb{E}_{P_k}[e] - \mathbb{E}_{P_{k-1}}[e]\right) = \mathbb{E}_{P_K}[e] - \mathbb{E}_{P_0}[e] = \langle e, L_K - 1\rangle .$$

This algebra does not require P_k to be independent of R. Conditioning on the observed R fixes the realized trajectory and endpoint. What is not covered is counterfactual regret $J(\Omega(U), U) - J(\Omega(R), U)$, which depends on how the optimizer itself changes when its objective changes.

A.4 Local policy remainder

Let $\ell_\eta = dP_{\theta+\eta d}/dP_\theta$ and assume $\eta \mapsto \ell_\eta$ is twice differentiable in $L^2(P_\theta)$ with $\sup_{t\in[0,\eta]} \|\ell_t''\| \le M$. Taylor's theorem in a Banach space yields

$$\ell_\eta - 1 = \eta \ell_0' + \int_0^\eta (\eta - t)\ell_t'' dt, \qquad \|r_{\eta,d}\| \le \frac{M\eta^2}{2}.$$

For a parametric policy, $\ell_0' = \langle \nabla \log \pi_\theta, d\rangle$. Pairing with e gives

$$|B_\eta - \eta \langle e, h_{\theta,d}\rangle| \le \|e\| \, \|r_{\eta,d}\| \le \varepsilon M \eta^2/2.$$

A.5 Proof of Theorem 3 and the finite-sample radius

Decompose $e = e_S + e_\perp$ and set $a = \|e_S\|$. Feasibility implies $0 \le a \le d$ and $\|e_\perp\| \le \sqrt{\varepsilon^2 - a^2}$, so

$$|\langle e, h\rangle| \le f(a) := ap + \sqrt{\varepsilon^2 - a^2}\, q.$$

The concave function f has unconstrained maximizer $a_\star = \varepsilon p/r$. If $d \ge a_\star$, its value is εr and the witness is $e_\star = \pm\varepsilon h/r$. Otherwise the maximum occurs at $a = d$ and is attained by aligning e_S and $e_\perp$ with the corresponding components of h. This proves Equation (9), including degenerate cases by continuity.

For the sampling statement, let $\widehat{\theta}$ be the mean of n prompt-block coordinate vectors. A coordinate-wise sub-Gaussian tail bound at $t = \sigma\sqrt{2\log(2m/\alpha)/n}$ and a union bound give $\max_j |\widehat{\theta}_j - \theta_j| \le t$ with probability at least $1 - \alpha$. Hence $\left\|\widehat{\theta} - \theta\right\| \le \sqrt{m}, t = r_n(\alpha)$ and $\|\Pi_S e\| = \|\theta\| \le \widehat{\delta}$. The same event implies the envelope for every optimizer sensitivity h; no additional union bound over optimizers is required.

B Proofs and Boundaries for Validation Design

B.1 Proof of Theorem 4

For any orthogonal projector Π_S,

$$\mathcal{R}_\mu(S) = \mathbb{E}\|h_T\|^2 - \mathbb{E}\|\Pi_S h_T\|^2 = \operatorname{tr}(C_\mu) - \operatorname{tr}(\Pi_S C_\mu).$$

If $S \subseteq \mathcal{G}$, then $\operatorname{tr}(\Pi_S C_\mu) = \operatorname{tr}(\Pi_S \Pi_\mathcal{G} C_\mu \Pi_\mathcal{G})$. The Ky Fan maximum principle states that the rank-m projector maximizing this trace spans leading eigenfunctions. Decomposing h_T into $\mathcal{G}$ and $\mathcal{G}^\perp$ yields Equation (13). The second moment is uncentered because the risk concerns h_T itself, not deviations from the mean sensitivity.

B.2 Proof of Proposition 5

For the unit sphere of an r-dimensional space, any $m < r$ subspace has a nonzero orthogonal complement inside that space. A unit vector in this complement has residual norm one, while no projection residual exceeds one. If $m \geq r$, choose the full tangent span. For a compact ellipsoid with orthogonal semiaxes $\sqrt{\lambda_j}$, choosing the first m axes leaves worst residual $\sqrt{\lambda_{m+1}}$; the standard width lower bound shows no other m-space can reduce it (Pinkus, 1985).

B.3 No unrestricted universal design

The full-sphere case is an impossibility result. If the admissible optimizer class can use any unit direction in a tangent space of rank exceeding m, every m-measurement design leaves a unit blind direction. A covariance spectrum can assign that direction low average mass, but it cannot remove it. We claim spectral optimality only for Equation (12) and width optimality only after specifying $\mathcal{K}$.

C Optimizer-Specific Derivations

C.1 Best-of-N

Let $S = R(X)$, $F(s) = P_0(S \leq s)$, $F(s^-) = P_0(S < s)$, and $p_s = P_0(S = s)$. If S is nonatomic, then $V = F(S)$ is uniform on $[0, 1]$. Equivalently, for arbitrary S, augment each candidate with an independent $U \sim \mathrm{Unif}[0, 1]$ and define the randomized refined rank

$$V = F(S^-) + Up_S.$$

Then V is uniform, BoN selects the largest refined rank, and its likelihood ratio on the augmented space is

$$L_N(V) = NV^{N-1}, \qquad h_N(V) = NV^{N-1} - 1.$$

For a discrete score without randomized refinement, $F(S)$ is not uniform. Suppose ties are resolved by candidate index (or any rule independent of candidate identity within a score atom), so selection preserves the conditional law $P_0(X \mid S = s)$. The selected score equals s with probability $F(s)^N - F(s^-)^N$, hence the exact likelihood ratio for $p_s > 0$ is

$$L_N(x) = \frac{F(s)^N - F(s^-)^N}{p_s}, \qquad h_N(x) = \frac{F(s)^N - F(s^-)^N}{p_s} - 1, \quad s = R(x). \tag{15}$$

The continuous expression is the nonatomic special case, not a consequence of deterministic tie-breaking. Both forms show why changing N changes the sensitivity family that validation must cover.

C.2 Exponential tilting

For $P_\beta(dx) = \exp(\beta R(x))P_0(dx)/Z_\beta$,

$$h_\beta(x) = \frac{\exp(\beta R(x))}{Z_\beta} - 1, \qquad B_\beta = \langle e, h_\beta \rangle .$$

No small-β approximation is required. Expanding at $\beta = 0$ recovers the centered reward direction as the first-order sensitivity.

C.3 Policy gradients and baselines

For any action-independent baseline $b(x)$,

$$\mathbb{E}_{a \sim \pi_\theta(\cdot|x)}[b(x)\nabla_\theta \log \pi_\theta(a \mid x)] = 0.$$

The population direction, hence its first-order pairing with e, is unchanged. The covariance of its Monte Carlo estimator can change substantially, which is the role exploited by ReMax-style baselines (Li et al., 2024).

D Complete Experimental and Ablation Evidence

D.1 Experimental philosophy and common metrics

Every experiment is paired with a falsifiable proposition. We report the signed direct gap B, geometric contraction $\langle e, h\rangle$, projection radius $\|\Pi_{\mathrm{Ker}\, A} h\|$, attainment ratio, alignment cosine $\langle e, h\rangle / (\|e\|\, \|h\|)$, and risks across optimizer directions. Hyperparameters are not selected for predictive accuracy. Selector, local conformal, and AUROC branches are outside the contribution because they do not test the geometric quantities in the theorems.

D.2 Exact synthetic witnesses and identifiability

The ambient numerical space has dimension 64. Each family contains six frozen sensitivities. Validation subspaces are chosen either from the sensitivity span or from preregistered generic directions. For every nonzero radius, the analytic witness $e_\star$ is inserted directly; success requires norm feasibility, $Ae_\star = 0$, and attainment.

Table 4: Maximum relative residual over six frozen directions as validation dimension changes. Values near zero indicate family identifiability.

Family	$m = 0$	1	2	3	4	5	6
BoN	1.000000	.911327	.672554	.368758	.136020	.026550	$< 10^{-12}$
Exponential tilt	1.000000	.700797	.290365	.066534	.006081	.000139	$< 10^{-12}$
Policy gradient	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	$< 10^{-12}$

At $\varepsilon = 0.1$, a representative top-8 BoN witness has achieved bias 0.0038208631162165124 versus predicted 0.003820863116216513, attainment ratio one, and zero reported validation leakage. All families attain their nonzero radii numerically.

D.3 Average-risk design ablation

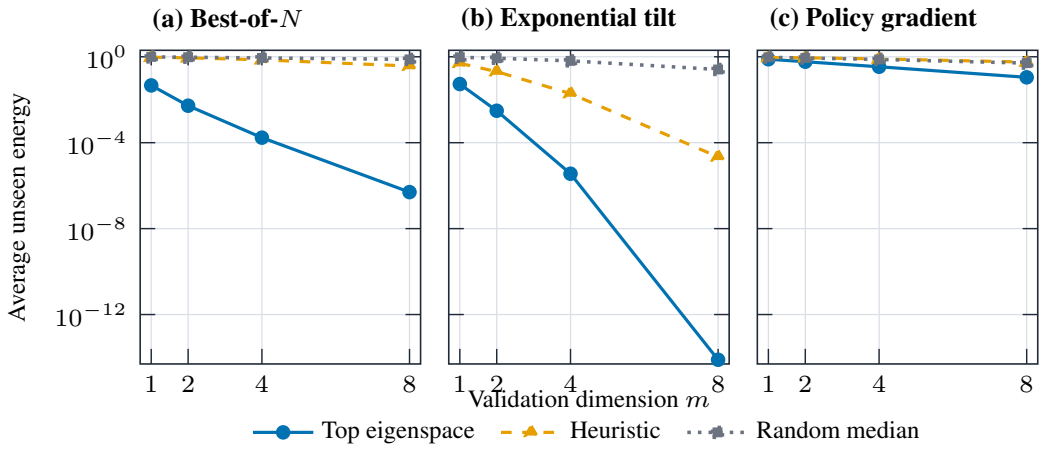

Figure 3: Frozen-family average-risk design (log scale). Leading eigenspaces solve the declared average objective. The policy-gradient family is less compressible than BoN or tilt. Color, marker shape, and line style redundantly encode each design.

This comparison makes no IID prediction claim. On the same interaction distribution, a sufficiently flexible direct regression can match an operator implementation. The experiment tests only the spectral design theorem for a frozen second moment.

Table 5: Average relative unseen energy for raw sensitivities: top eigenspace / heuristic / random median.

Family	Design	$m = 1$	$m = 2$	$m = 4$	$m = 8$
BoN	Top	.046426	.005256	.000171	5.00×10^{-7}
	Heuristic	.956071	.890155	.719313	.370153
	Random	.984849	.953095	.879097	.756082
Tilt	Top	.053685	.003068	3.58×10^{-6}	7.83×10^{-15}
	Heuristic	.497126	.203031	.019336	2.18×10^{-5}
	Random	.946800	.853626	.645054	.255836
Policy gradient	Top	.766841	.587208	.342187	.110425
	Heuristic	.925015	.894276	.794867	.565989
	Random	.943124	.876213	.749668	.506576

D.4 Held-out optimizer directions

Training and test directions are separated before design. Interpolation and matched-covariance tests stay within the learned geometry. Covariance shift and the unseen-subspace test depart from it by construction.

At $m = 4$, unit-normalized policy-tangent test risks for the training top eigenspace are 0.3495 under interpolation, 0.9694 under covariance shift, and 1.0000 for an unseen subspace. Corresponding test-oracle risks are 0.3332, 0.3333, and 0.2064. These shifted cases are the negative ablation for universal spectral-design claims.

D.5 Noisy and finite-sample validation operator

The noisy suite separates a deterministic null-space radius from uncertainty in estimating visible validation coordinates. The joint certificate controls the event that a residual passes a noisy threshold and still violates the reported radius; it is not a conditional coverage statement among issued certificates.

Table 6: Finite-sample operator checks and negative controls.

Test	Cells/repetitions	Main value	Interpretation
Exact noisy-constraint witness	1,728 cells	max objective error 8.88×10^{-16}	Closed-form constrained radius is attained.
Finite-sample main grid	1,944 cells	largest violation minus α: -0.0091	Frozen fixed-S joint certificate passes.
Delete sampling radius	1,944 cells	max violation 0.5093	Naive certificate falsified in 934 cells.
Conditional-pass trap	200,000 reps	joint false rate 0.006115	Every rare pass is false conditionally; unconditional control still holds.

Deleting simultaneous calibration or the $\sqrt{m}$ term did not cause failures on this frozen grid, so these experiments do not establish either factor as necessary in general. Deleting the sampling radius does falsify the certificate.

E Real Oracle Geometry on Frozen Candidate Pools

E.1 Protocol

The four formal cells use Qwen2.5-3B-Instruct or Qwen2.5-7B-Instruct and GSM8K or MATH-500. Each cell contains 256 test prompts and 4,096 sampled responses ($K = 16$); a separately generated greedy response is an audit only and is excluded from the base pool. Sampling uses temperature 0.8 and seed 20260730. Utilities are exact parsed-answer correctness; MATH-500 uses frozen symbolic verification. Candidate generation is gold-blind. Candidate hashes, model revisions, decoding, proxy definitions, optimizer registries, splits, bootstrap seed 20260801, and stopping rules are frozen before oracle joining.

Self-consistency groups equivalent parsed answers and scores by group frequency. Mean-logprob rank is frozen from candidate sequence log probabilities and does not use gold. BoN uses $N \in \{1, 2, 4, 8, 16\}$; tilt uses $\beta \in \{0, .5, 1, 2, 4, 8\}$. Prompt bootstrap intervals use 10,000 resamples.

Because these are finite candidate pools with discrete proxy values, BoN endpoint weights are computed from the exact frozen selection rule on each pool, not from the continuous formula NV^{N-1}. The generic endpoint theorem then uses the resulting empirical likelihood ratio directly.

Table 7: BoN-16 real-pool results. Gap is proxy shift minus utility shift; intervals are two-sided prompt-bootstrap 95% intervals.

Cell	Proxy	Gap	95% interval	Utility shift	Cosine
3B-GSM8K	Self-consistency	-.02352	[-.05001, .00312]	.11009	-.09556
7B-GSM8K	Self-consistency	-.01907	[-.03600, -.00251]	.06065	-.19233
3B-MATH-500	Self-consistency	.00526	[-.01813, .02803]	.07611	.01373
7B-MATH-500	Self-consistency	-.00890	[-.03230, .01271]	.08081	-.02639
3B-GSM8K	Mean-logprob rank	.49894	[.47284, .52534]	-.03384	.32545
7B-GSM8K	Mean-logprob rank	.46264	[.44533, .48054]	.00236	.30624
3B-MATH-500	Mean-logprob rank	.46471	[.44116, .48830]	.00040	.30877
7B-MATH-500	Mean-logprob rank	.46408	[.44432, .48384]	.00101	.30711

Table 8: Exponential tilt at $\beta = 4$ on the same pools.

Cell	Proxy	Gap	95% interval	Utility shift	Cosine
3B-GSM8K	Self-consistency	-.01195	[-.02311, -.00056]	.06948	-.10953
7B-GSM8K	Self-consistency	-.01042	[-.01819, -.00267]	.04141	-.19482
3B-MATH-500	Self-consistency	-.00303	[-.01269, .00650]	.04608	-.02562
7B-MATH-500	Self-consistency	-.00040	[-.00892, .00812]	.04077	-.00354
3B-GSM8K	Mean-logprob rank	.31289	[.30071, .32533]	-.01586	.48463
7B-GSM8K	Mean-logprob rank	.29592	[.28689, .30527]	.00107	.46497
3B-MATH-500	Mean-logprob rank	.29234	[.28251, .30217]	.00469	.46125
7B-MATH-500	Mean-logprob rank	.29506	[.28579, .30420]	.00196	.46364

Across all 88 registered endpoints, calibration of direct gap on geometric contraction has slope 1, intercept 0, RMSE 3.65×10^{-17}, and maximum absolute error 1.06×10^{-16}. We do not collapse the signed results into AUROC because the theorem predicts a calibrated continuous contraction rather than a binary harmfulness label.

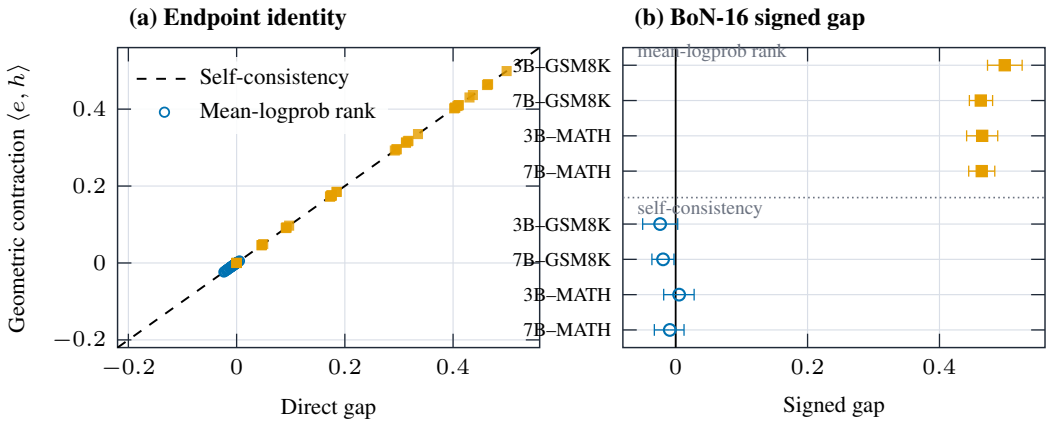

Figure 4: Implementation check and proxy-specific signed effects. (a) The 88 preregistered endpoints satisfy the algebraic endpoint identity to numerical precision. (b) BoN-16 signed gaps (points) with 95% bootstrap intervals show that alignment differs sharply by proxy.

F Operational Measurement-Budget Design

The operational study reuses the frozen 3B-GSM8K pool. We fix the training and held-out optimizer registries before design. Measurements come from a finite dictionary of gold-label-equivalent prompt statistics. The comparison includes an ordinary heuristic, the exact optimum for small dictionaries, a greedy feasible design, random dictionary subsets, and the unrestricted top-eigenspace oracle relaxation.

For random one-atom selection, median risk is 1.2531 with range [.7600, 1.3090]. In the primary finite-sample cell ($n = 128$, joint noise 0, issuance threshold $\tau = 0.1$, $m = 1$), self-consistency issuance is 100%. Average certificate radii are .8589 for the heuristic, .7371 for exact or greedy selection, and .5722 for the oracle relaxation, all with 100% signed-bias coverage. Mean-logprob measurements issue no certificate at that threshold. The gap measures the cost of restricting measurements to a feasible dictionary. It does not make the oracle relaxation implementable.

G Finite Optimizer Trajectory

One Qwen2.5-0.5B LoRA direction is learned from 32 gold-free self-consistency-selected GSM8K responses. Evaluation conditions the reference and checkpoint policies on each frozen eight-candidate support, making endpoint likelihood ratios exactly computable. The implementation gate requires reliable importance weights and common support; otherwise the experiment would stop rather than substitute a proxy density ratio.

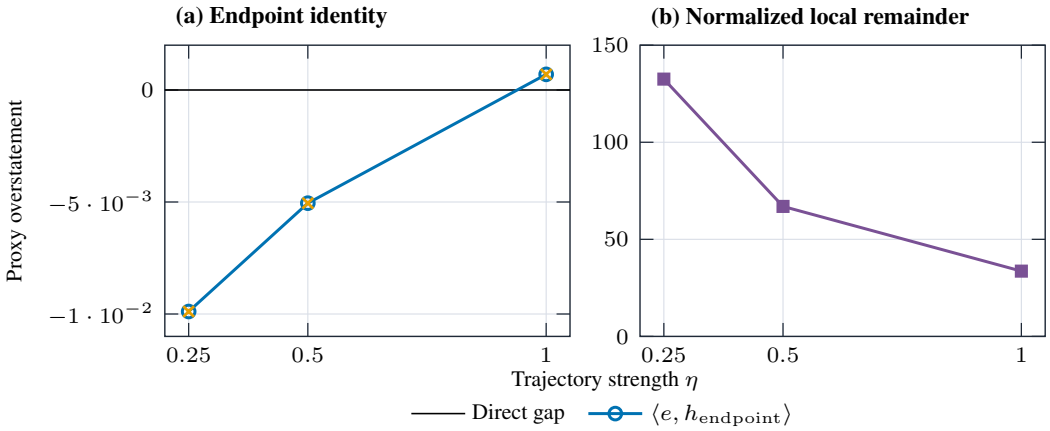

Figure 5: Endpoint and local trajectory diagnostics. (a) Direct and geometric endpoint gaps overlap at every checkpoint. (b) The normalized local remainder remains non-negligible, so the tangent is only an approximation.

Table 9: Trajectory endpoint identity and local approximation.

η	Direct gap	Geometry	Max prompt error	ESS	$\|r_\eta\| / \eta^2$
.25	-.009889	-.009889	9.02×10^{-17}	.8732	132.533
.50	-.005052	-.005052	5.55×10^{-17}	.8764	66.956
1.00	.000698	.000698	6.25×10^{-17}	.8620	33.622

The action-independent baseline leaves every prompt-level population mean unchanged (maximum absolute difference 0). It reduces mean variance from 214.686 to 16.685, a ratio of .0777. The variance reduction matches the stated boundary but does not establish an endpoint bias correction.

H Artifact Map, Hashing, and Stopping Rules

The supplementary data directory contains the following source-derived tables:

- `oracle_endpoints.csv`: all 88 real proxy-optimizer endpoints;
- `phase_transition.csv`: all 21 family-dimension cells;
- `exact_design_comparison.csv`: 72 normalization-method-dimension cells;
- `heldout_optimizer_shift.csv`: 90 held-out transfer summaries;
- `measurement_design.csv`: 78 operational/oracle design summaries;
- `trajectory.csv` and `noisy_validation_summary.csv`.

Every row includes its source path and source SHA-256. The oracle matrix aggregate also stores result hashes for each formal cell. Infrastructure-only amendments record GPU type, count, parallelism, batch scheduler, and expected cost without changing scientific rules. Failures before gold access are retained in the reader-facing result logs. Primary stopping rules are identity-gate failure, hash mismatch, support/importance-weight failure, or a preregistered timeout; no stopping rule depends on the sign of reward hacking.

I Additional Limitations and Scope Exclusions

The following are explicitly outside the present claim set:

- universal validation design over unrestricted optimizer directions;
- counterfactual regret after reoptimizing the true reward;
- guarantees under support mismatch without a valid density ratio;
- learning the true residual e from unlabeled data;
- superiority of an operator estimator over same-interaction direct regression;
- AUROC, selector gains, or local conformal certificates as primary contributions.

The real experiments are plug-in diagnostics on frozen candidate pools. Extending the theory to adaptively chosen, costly, nonlinear measurements and to jointly estimated $(\widehat{e}, \widehat{h})$ calibration is future work.

J Use of Large Language Models

The authors used large language models to help organize the manuscript, inspect machine-readable experiment outputs, and draft plotting and table-extraction code. The authors checked theorem statements, numerical claims, citations, source hashes, and manuscript text against frozen project artifacts. Language models did not judge the mathematical correctness utilities reported in the experiments.