MINIMAL FUNCTIONAL MODIFICATION: MINIMAL BEHAVIORAL REALIZATION OF ACTIVATION INTERVENTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Activation interventions provide one internal realization of a desired language-model behavior, but need not provide a minimal one. We study *minimal behavioral realization*: given a frozen reference intervention function and a specified behavioral observation, what is the smallest hidden-state modification required to reproduce the reference behavior across prompting contexts? This problem is context dependent because demonstrations alter both model states and local sensitivities to intervention. We first characterize when directional modification is strictly more economical than optimal scalar contraction. More importantly, we show that *pointwise redundancy need not be shareable across contexts*: the excess cost of a shared finite-width realization is governed by the spectral structure and input predictability of the pointwise optimal intervention field. We further quantify the cost of preserving richer behavioral observations and establish conditions under which local advantages persist at finite intervention amplitudes. These results motivate Minimal Functional Modification (MFM), a shared nonlinear realization module trained directly to balance reference-output fidelity and hidden-state displacement without explicitly constructing Jacobians. Across two language-model families and three multiple-choice reasoning benchmarks, MFM reduces reference intervention displacement cost by 57.6% on average while preserving four-shot answer distributions. Across changed four-shot demonstration sets, MFM yields contextual-response errors of 0.0085–0.0103 across the six model–task settings. When answer-set probability mass is additionally constrained, MFM still achieves 43.4% displacement-cost savings on the evaluated subset.

1 INTRODUCTION

Large language models adapt through in-context demonstrations (Brown et al., 2020; Wei et al., 2022), while activation interventions manipulate behavior through intermediate representations (Todd et al., 2024). Existing work largely asks how to induce or adapt a desired behavior. We instead ask a reference-relative question: *once a useful behavioral effect has been achieved, how much internal modification is actually necessary to realize it?*

A learned intervention is only one internal realization of its behavioral effect. Represented information need not be causally used (Ravfogel et al., 2021; Geiger et al., 2021), and distinct interventions can induce the same observable behavior (Venkatesh & Kurapath, 2026). We therefore study *minimal behavioral realization*: for a frozen model, a frozen reference intervention function, and a specified behavioral observation, we seek the minimum hidden-state displacement, under a fixed metric, that reproduces the refer-

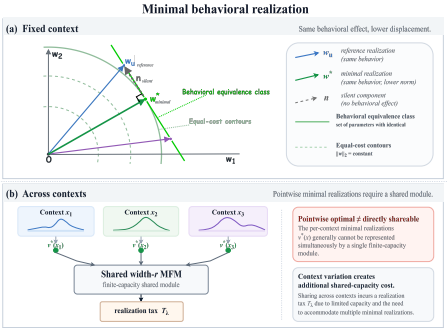

Figure 1: **Minimal behavioral realization.** (a) At a fixed context, the same behavioral observation can be realized with lower displacement than the reference intervention. (b) Across contexts, pointwise minimal realizations need not be directly shareable by a finite-capacity module, inducing a realization tax.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ence behavior. Note here *Frozen reference intervention* means that the reference module is fixed, although its realized correction may vary with context.

Minimality is both reference-relative and observation-relative. In our multiple-choice setting, the primary observation is the answer distribution, and demonstrations remain in the prompt for both reference and candidate. Thus we test whether a realization preserves not only a single reference output but also its response to changing context. This does not imply preservation of unrestricted free-form generations or latent reasoning traces.

The central difficulty, illustrated in Figure 1, is that *pointwise redundancy need not be shareable*. At a fixed context, directional modification can require less displacement than scalar contraction because the reference may contain silent components or active directions with unequal sensitivities. Across contexts, however, the pointwise optima can vary. We show that a shared width-r module incurs an unavoidable cost determined by the spectral tail of the pointwise optimal intervention field, with an additional penalty when its coefficients are difficult to predict from the module input. Hence large pointwise savings can coexist with poor shared realizability.

The attainable savings also depend on what behavior must be preserved. We derive the exact local cost of adding behavioral observations and sufficient conditions for local directional advantages to persist at finite intervention amplitudes. These results motivate *Minimal Functional Modification* (MFM), a shared nonlinear realization module trained to balance reference-output fidelity and hidden-state displacement without explicit Jacobian construction. We evaluate MFM with Qwen2.5-7B and Llama-3.1-8B on ARC-C, CSQA, and LogiQA, measuring answer fidelity, contextual transfer, shared-capacity effects, and the cost of additionally constraining answer-set probability mass.

Our contributions are threefold: **First**, we formulate minimal behavioral realization and characterize when directional modification is strictly more economical than optimal scalar contraction. **Second**, we distinguish *pointwise opportunity* from *shared realizability*, characterizing the latter through the spectrum and input predictability of the pointwise optimal field. **Third**, we quantify the local cost of refining the preserved observation and instantiate the framework in MFM, which substantially reduces reference displacement cost while preserving contextual answer behavior. Throughout, "compression" means reducing intervention displacement cost under a fixed metric, not model size or inference cost.

2 Related Work

In-context learning and internal task representations. Few-shot prompting and chain-of-thought use demonstrations to adapt predictions (Brown et al., 2020; Wei et al., 2022). ICL has been interpreted as latent inference or implicit learning (Xie et al., 2021; Akyürek et al., 2022; Von Oswald et al., 2023), and function/task vectors connect prompts to internal task representations (Todd et al., 2024; Dong et al., 2026). We instead freeze a reference intervention function and ask for the minimum hidden-state displacement needed to reproduce its specified behavior as queries and demonstrations vary.

Activation interventions. Activation engineering and representation-finetuning methods steer behavior through intermediate states (Turner et al., 2023; Rimsky et al., 2024; Li et al., 2023; Zou et al., 2023; Wu et al., 2024; Liang et al., 2026). Recent approaches further adapt intervention strength, direction, or subspace to the input (Lee et al., 2025; Sun et al., 2025; Yang et al., 2026; Feng et al., 2026; Sheng et al., 2026; Heyman & Vandeputte, 2026). These works primarily ask how to construct or adapt an intervention to produce a desired behavior; we instead fix the behavioral effect of a reference intervention and seek its minimum-displacement realization.

Representation geometry and shared capacity. Causal, erasure, and latent-knowledge analyses distinguish represented information from behavioral use (Vig et al., 2020; Ravfogel et al., 2021; Geiger et al., 2021; Burns et al., 2022; Ravfogel et al., 2020; 2022; Belrose et al., 2023). Recent work studies steering non-identifiability, collateral effects, minimum-distortion control, intervention subspaces, and cross-context geometry (Venkatesh & Kurapath, 2026; Nguyen et al., 2026; Skifstad et al., 2026; Wang et al., 2026; Doan et al., 2026; Robertson et al., 2026). Non-identifiability shows that multiple interventions may induce similar behavior, but does not identify the minimum-displacement equivalent realization or determine whether pointwise optima can be amortized by one

shared module. Our shared-capacity analysis addresses both questions by relating excess local cost to the spectrum and predictability of the pointwise optimal realization field.

Behavioral matching and preservation. Knowledge distillation and model editing match or modify output behavior (Hinton et al., 2015; De Cao et al., 2021; Mitchell et al., 2021; Meng et al., 2022). MFM also matches reference outputs, but the backbone stays frozen and the optimized quantity is the hidden-state displacement required to realize the behavior of a specific internal intervention. Using projection and constrained optimization (Bertsekas, 2009; Belrose et al., 2023), we further quantify the cost of refining the preserved observation.

3 Minimal Behavioral Realization

We formulate minimal behavioral realization as a constrained problem: given a reference intervention that induces a desired behavioral observation, we seek the smallest hidden-state modification that reproduces this observation. The analysis proceeds in four stages. First, we characterize the local geometry of the minimum-displacement realization and identify when directional freedom provides gains over scalar contraction. Second, we study whether these pointwise gains can be shared across contexts through a finite-capacity module. Third, we quantify how the minimum displacement changes when the preserved behavioral observation is refined. Fourth, we establish sufficient conditions under which the local advantage survives at finite intervention amplitudes. Finally, we introduce MFM as a practical nonlinear realization of this objective.

3.1 Behavioral matching and displacement cost

Let $x = (q, \mathcal{D}) \sim P_X$ be a query with demonstrations. At a fixed intervention boundary, $z(x) \in \mathbb{R}^d$ is the module input and $c(x)$ contains the remaining downstream context. A differentiable observation $F_c : \mathbb{R}^d \to \mathbb{R}^K$ records the aspect of behavior to preserve. Our primary instance is the renormalized probability vector over fixed single-token answer alternatives. Fix $\Omega \succ 0$ for observation discrepancy and $G \succ 0$ for displacement cost. For a reference correction $u(x)$, define

$$\mathsf{C}_\epsilon(x; u) = \inf_v \{\|v\|_G^2 : \|F_c(z + v) - F_c(z + u)\|_\Omega \le \epsilon\}, \tag{1}$$

$$\mathsf{R}_\epsilon(x; u) = \|u(x)\|_G^2 - \mathsf{C}_\epsilon(x; u) \ge 0. \tag{2}$$

The reference intervention is always feasible. Exact matching defines a behavioral equivalence class, while positive tolerance defines a feasible region around this class. The selected behavioral observation and displacement metric jointly determine the achievable compression and the associated redundancy. The shared penalized objective is

$$\mathcal{L}_\lambda(v) = \mathbb{E}_x \left[\tfrac{1}{2} \|F_c(z + v(x)) - F_c(z + u(x))\|_\Omega^2 + \tfrac{\lambda}{2} \|v(x)\|_G^2 \right], \qquad \lambda > 0. \tag{3}$$

This is the nonlinear objective optimized by MFM. We analyze its first-order geometry and optimize its nonlinear form directly. Operating points are selected by held-out fidelity along the penalty path. Appendix A specifies observation refinement and consistent coordinate transformations.

3.2 Pointwise directional advantage

We first study the local geometry of minimal realization. Around a fixed context, the nonlinear behavioral matching problem can be approximated by the sensitivity of the preserved observation to hidden-state modifications. This local view reveals when directional freedom can reduce displacement beyond uniform scaling of the reference intervention.

Let $J_{\mathrm{raw}} = \partial F_c / \partial z|_z$ and choose L with $L^\top L = \Omega$. We use the observation-whitened Jacobian

$$J = L J_{\mathrm{raw}}, \tag{4}$$

so that $\|J\delta\|_2^2 = \|J_{\mathrm{raw}}\delta\|_\Omega^2$. Linearizing both reference and candidate observations about the same baseline gives the pointwise problem

$$Q_\lambda(v) = \tfrac{1}{2} \|J(v - u)\|_2^2 + \tfrac{\lambda}{2} \|v\|_G^2. \tag{5}$$

The corresponding exact local minimum cost is

$$\mathsf{C}_0(J,u) = \min_v \{\|v\|_G^2 : \ Jv = Ju\}. \tag{6}$$

If $\widetilde{J} = [J; J_R]$ comes from a refined observation, then

$$\ker \widetilde{J} \subseteq \ker J, \qquad \mathsf{C}_0(\widetilde{J},u) \geq \mathsf{C}_0(J,u). \tag{7}$$

Write $B = JG^{-1/2}$ and $H = JG^{-1}J^\top$. Classical projection gives the minimum-cost intervention with exactly the same first-order effect as u:

$$v_0 = G^{-1}J^\top H^\dagger Ju, \qquad u = v_0 + n, \quad Jn = 0. \tag{8}$$

With positive regularization, the unique local minimizer is

$$v_\lambda = G^{-1}J^\top (H+\lambda I)^{-1}Ju. \tag{9}$$

We compare these solutions with optimal scalar contraction of the same reference to identify when directional freedom reduces cost.

Decompose the whitened reference $G^{1/2}u$ into right singular directions of B. An occupied active direction with singular value σ_i is retained by ridge with factor $t_i = \sigma_i^2/(\sigma_i^2+\lambda)$, while the component in $\ker B$ is removed. A scalar contraction must use the same factor for every occupied direction.

Theorem 3.1 (Exact directional advantage over optimal scalar contraction)**.** *Assume $\|u\|_G > 0$ and $\|Ju\|_2 > 0$. Let $\alpha_\lambda u$ be the best scalar contraction for Eq. 5. If occupied active singular directions have singular values σ_i and coefficients a_i, and ν is the norm of the component of $G^{1/2}u$ in $\ker B$, then*

$$\Delta_\lambda = \tfrac{1}{2}\sum_i (\sigma_i^2+\lambda)(\alpha_\lambda - \tfrac{\sigma_i^2}{\sigma_i^2+\lambda})^2 a_i^2 + \tfrac{\lambda}{2}\alpha_\lambda^2\nu^2, \tag{10}$$

where $\alpha_\lambda = \|Ju\|_2^2/(\|Ju\|_2^2 + \lambda\|u\|_G^2)$. The gap is zero if and only if (i) the reference has no silent component and (ii) all occupied active directions have the same positive singular value. Otherwise direction-selective modification is strictly better. The same equality condition holds when comparing minimum displacement at the linear fidelity achieved by v_λ:

$$\|v_\lambda\|_G^2 \leq \min_{\alpha\in[0,1]:\ \|J(\alpha u - u)\|_2 \leq r_\lambda} \|\alpha u\|_G^2, \qquad r_\lambda = \|J(v_\lambda - u)\|_2. \tag{11}$$

Equality holds under exactly the same two conditions.

The theorem identifies two scalar-inaccessible sources of redundancy: occupied silent displacement and unequal sensitivities among occupied active directions. Only directions occupied by the reference contribute to this advantage. Proofs and counterexamples are in Appendices B–C.

3.3 From pointwise opportunity to shared capacity

The pointwise analysis characterizes the best realization available when each context can use its own optimal modification. We next ask whether these context-specific savings can be amortized by a single shared module.

Restore the dependence on x and define

$$A_{\lambda,x} = J_x^\top J_x + \lambda G, \qquad v_\lambda^{\mathrm{pt}}(x) = A_{\lambda,x}^{-1}J_x^\top J_x u(x). \tag{12}$$

Let $\mathcal{Q}_\lambda(v) = \mathbb{E}_x Q_{\lambda,x}(v(x))$ and set

$$T_\lambda(v) = \tfrac{1}{2}\mathbb{E}_x\|v(x) - v_\lambda^{\mathrm{pt}}(x)\|_{A_{\lambda,x}}^2, \qquad \Gamma_\lambda = \mathbb{E}_x\Delta_\lambda(x). \tag{13}$$

Here T_λ measures the cost of failing to realize the pointwise optimal intervention field. For $J_x u(x) = 0$, we set $\Delta_\lambda(x) = 0$. Let $\mathcal{S}$ contain all measurable $s(x) = \alpha(x)u(x)$ with $\alpha(x) \in [0,1]$.

Proposition 3.1 (Opportunity–realization identity)**.** *Assuming the relevant expectations are finite, completing the square gives*

$$\inf_{s\in\mathcal{S}} \mathcal{Q}_\lambda(s) - \mathcal{Q}_\lambda(v) = \Gamma_\lambda - T_\lambda(v). \tag{14}$$

If this margin is $\zeta_\lambda > 0$, every scalar with no larger mean squared linear residual has displacement cost

$$C(s) \geq C(v) + 2\zeta_\lambda/\lambda, \qquad C(v) = \mathbb{E}_x\|v(x)\|_G^2. \tag{15}$$

To characterize when a shared module can approximate the pointwise optimal field, consider a module whose output lies in an affine subspace of dimension at most r. Define

$$w^*(x) = G^{1/2} v_\lambda^{\mathrm{pt}}(x), \quad \bar{w} = \mathbb{E} w^*, \quad \Sigma_* = \mathrm{Cov}(w^*), \quad S_r = \sum_{j>r} \mu_j(\Sigma_*), \tag{16}$$

where eigenvalues are ordered nonincreasingly. The quantity S_r measures the variation in the *required realization field* outside its leading r directions.

Theorem 3.2 (Shared capacity and contextual predictability)**.** *Assume finite second moments. Every affine-output rank-r candidate obeys*

$$T_\lambda(v) \geq \tfrac{\lambda}{2} S_r. \tag{17}$$

Consequently, $\lambda S_r/2 \geq \Gamma_\lambda$ rules out a strict local advantage over the pointwise scalar oracle for every such candidate. For a constructive converse, let U_r contain top eigenvectors of Σ_*, $a_r(x) = U_r^\top(w^*(x) - \bar{w})$, and $v_f(x) = G^{-1/2}(\bar{w} + U_r f(z(x)))$. If $B_x^\top B_x + \lambda I \preceq \beta I$ almost surely, then

$$T_\lambda(v_f) \leq \tfrac{\beta}{2}\{S_r + \varepsilon_r(f)\}, \qquad \varepsilon_r(f) = \mathbb{E}\|f(z) - a_r(x)\|^2. \tag{18}$$

Thus $\beta(S_r + \varepsilon_r(f))/2 < \Gamma_\lambda$ suffices for this candidate to beat the scalar oracle locally.

Proof idea. The whitened Hessian is bounded below by λI. The best affine rank-r mean-square approximation error is S_r. For the constructive candidate, orthogonal residual and coefficient errors add exactly, giving the upper bound. Full proofs are in Appendix E.

The module used below has output $W_2 h(z) + b_2$, hence satisfies the lower bound regardless of its nonlinear input map. The upper bound applies when the input map can represent the required coefficient predictor f. Its prediction error decomposes as

$$\varepsilon_r(f) = \mathbb{E}\|a_r - \mathbb{E}[a_r \mid z]\|^2 + \mathbb{E}\|f(z) - \mathbb{E}[a_r \mid z]\|^2. \tag{19}$$

A concentrated output spectrum does not ensure predictable coefficients. Appendix D gives the corresponding weighted module-visible optimum and separates measurability from architecture.

For contextual prompts, changing demonstrations can alter the local geometry of the minimum realization even when the reference intervention function is frozen. More strongly, there exist rank-one affine-output references with positive directional opportunity on every input for which no rank-one affine-output compressor beats the scalar oracle, even with fully informative inputs (Corollary E.1). The construction perturbs a scalar-equality case whose oracle points are not collinear: its spectral tail remains positive while directional opportunity becomes positive but arbitrarily small. Thus a reference can be easy to express but difficult to share economically across contexts.

3.4 The price of preserving additional behavior

Preserving one behavioral observation does not constrain all aspects of the reference behavior. We therefore study how the minimum realization cost changes when additional observations are required.

Let an added observation have whitened hidden-coordinate Jacobian D, so the refined linear map is $[B; D]$. Write $w_u = G^{1/2}u$, $w_0 = B^\dagger B w_u$, $P_N = I - B^\dagger B$, and $d_R = D(w_u - w_0)$.

Theorem 3.3 (Exact local preservation price)**.** *The increase in minimum squared displacement for exact linear matching is*

$$\Pi_{R|F} := \mathsf{C}_0([J; DG^{1/2}], u) - \mathsf{C}_0(J, u) = d_R^\top (DP_N D^\top)^\dagger d_R, \tag{20}$$

$$w_{F,R} = w_0 + P_N D^\top (DP_N D^\top)^\dagger d_R. \tag{21}$$

The price is zero iff $d_R = 0$, and $0 \leq \Pi_{R|F} \leq \|w_u - w_0\|^2$.

Proof idea. Every solution of the original constraint is $w_0 + n$, $n \in \ker B$. Since $w_0 \perp n$, the added cost is the least squared norm needed to solve $Dn = d_R$ within this kernel. The reference residual makes that problem feasible (Appendix F).

The price depends on both the unpreserved effect and the sensitivity available inside the previous null space. Weak sensitivity is costly for a *fixed* deficit, but cannot produce unbounded price for a fixed reference. Nested observations yield nonnegative cost increments; their sum equals the endpoint cost difference, while individual increments depend on ordering. These increments characterize how the minimum realization cost changes as the behavioral specification expands.

3.5 When the local advantage survives finite amplitude

Let $h_x(w) = LF_{c(x)}(z(x) + G^{-1/2}w)$ have a ρ_x-Lipschitz Jacobian on convex neighborhoods containing the relevant segments, with $Dh_x(0) = B_x$. For $w = G^{1/2}v(x)$, define

$$\begin{aligned} e_x(v) &= \rho_x \max\{\|w\|, \|w_u\|\}\|w - w_u\|, \\ K(v) &= \mathbb{E}\left[\|B_x(w - w_u)\|e_x(v) + \tfrac{1}{2}e_x(v)^2\right], \\ K_{\mathcal{S}} &= \mathbb{E}\left[\rho_x\|B_x w_u\|\|w_u\|^2 + \tfrac{1}{2}\rho_x^2\|w_u\|^4\right]. \end{aligned} \tag{22}$$

Assuming the relevant expectations are finite, Taylor's integral remainder gives $|\mathcal{L}_\lambda(v) - \mathcal{Q}_\lambda(v)| \le K(v)$ and a uniform scalar bound $K_{\mathcal{S}}$. Consequently,

$$\inf_{s\in\mathcal{S}} \mathcal{L}_\lambda(s) - \mathcal{L}_\lambda(v) \ge \Gamma_\lambda - T_\lambda(v) - K(v) - K_{\mathcal{S}}. \tag{23}$$

Combining with Theorem 3.2, $\beta(S_r + \varepsilon_r(f))/2 + K(v_f) + K_{\mathcal{S}} < \Gamma_\lambda$ is a sufficient nonlinear advantage condition. Applying this certificate at the deployed reference amplitude requires a valid curvature bound. We do not estimate a global curvature certificate in the LLM experiments; instead, we directly test whether the local capacity diagnostics remain qualitatively aligned with nonlinear fidelity. Appendix G proves the population bound and its matched-mean-fidelity consequence; Appendix H gives a pointwise fixed-budget sufficient condition.

3.6 MFM: Shared Nonlinear Realization without Explicit Jacobian Construction

The previous analysis characterizes the optimal realization field and the limitations of sharing it across contexts. MFM provides a practical nonlinear amortization of this field by learning a shared realization module that directly minimizes behavioral mismatch and displacement.

With baseline training statistics $(\widehat{\mu}, \widehat{\sigma})$, use

$$\widehat{z}_j = \frac{z_j - \widehat{\mu}_j}{\max(\widehat{\sigma}_j, 10^{-6})}, \qquad m_\theta(z) = W_2\,\mathrm{GELU}(W_1\widehat{z} + b_1) + b_2, \tag{24}$$

where $W_2 \in \mathbb{R}^{d\times r}$. Stage 1 acquires a reference on a frozen backbone via

$$\mathcal{L}_{\mathrm{acquire}}(\psi) = \mathbb{E}_{x,y}[-\log F_c(z + m_\psi(z))_y]. \tag{25}$$

Freeze its selected checkpoint, set $u(x) = m_\psi(z(x))$, and cache $p_u(x) = F_c(z + u)$. Stage 2 initializes $\phi \leftarrow \psi$ and minimizes

$$\mathcal{L}_{\mathrm{MFM}}(\phi) = \mathbb{E}_x\left[\tfrac{1}{2}\|F_c(z + m_\phi(z)) - p_u\|_\Omega^2 + \tfrac{\lambda}{2}\|m_\phi(z)\|_G^2\right]. \tag{26}$$

This is Eq. (3) restricted to a shared function. Both stages use width 64 by default. The standardization statistics are fixed; the primary G is a shrinkage inverse covariance, with Euclidean and diagonal covariance controls. Each metric specifies the displacement cost being minimized.

Diagnostics and model selection. Before fitting the width-sweep candidates, a fixed training Jacobian subset estimates Γ_λ and S_r. After fitting, T_λ and the ratio $\chi_\lambda = T_\lambda/\Gamma_\lambda$ provide accounting diagnostics. These local quantities are used for analysis rather than operating-point selection; nonlinear evaluation directly tests whether their qualitative capacity signal persists at the deployed reference amplitude.

Operating points minimize validation displacement subject to fixed mean-TV, tail-TV, and reference-disagreement constraints. Global/adaptive scalars and coordinate gates separate magnitude calibration from element-wise rescaling, while alternative reference-acquisition procedures test sensitivity to the acquired reference. Augmented observations and changed demonstrations test behavioral scope with fixed module weights. Appendix J gives the training budgets, caching procedure, capacity controls, and fidelity criteria. Section 4 reports the results.

4 Experiments

Our experiments address three empirical questions motivated by the minimal-realization framework. **Q1:** Does the local directional advantage persist at the deployed nonlinear intervention amplitude?

Table 1: **Four-shot realization.** Reference cost is 1.000; ± is SD across three seeds after averaging four prompt draws. Accuracy quantities are percentage points; all cells pass the fidelity budgets.

Task	Ref. Acc./Gain	Global	Adaptive	Coordinate	MFM	ΔAcc.
Qwen2.5-7B (Base)						
ARC-C	68.69 / +3.70	0.874 $\pm$0.004	0.721 $\pm$0.011	0.558 $\pm$0.018	**0.367** $\pm$0.013	-0.15
CSQA	75.90 / +3.00	0.906 $\pm$0.006	0.766 $\pm$0.009	0.613 $\pm$0.015	**0.421** $\pm$0.017	-0.11
LogiQA	44.91 / +3.92	0.928 $\pm$0.009	0.807 $\pm$0.018	0.641 $\pm$0.024	**0.463** $\pm$0.021	-0.20
Llama-3.1-8B (Base)						
ARC-C	66.21 / +3.92	0.891 $\pm$0.005	0.747 $\pm$0.013	0.587 $\pm$0.020	**0.398** $\pm$0.016	-0.10
CSQA	74.69 / +3.19	0.883 $\pm$0.007	0.719 $\pm$0.012	0.566 $\pm$0.017	**0.386** $\pm$0.014	-0.17
LogiQA	42.90 / +4.10	0.946 $\pm$0.011	0.834 $\pm$0.021	0.689 $\pm$0.027	**0.509** $\pm$0.023	-0.29

Q2: Does a shared realization preserve the reference response when the prompting context changes?
Q3: Do limited shared capacity and richer behavioral specifications reduce the attainable savings in the manner suggested by the theory?

We study these questions using two frozen model families and three multiple-choice benchmarks.

4.1 Experimental Setup

We evaluate frozen `Qwen/Qwen2.5-7B` (Team, 2024) and `meta-llama/Llama-3.1-8B` (Grattafiori et al., 2024) on ARC-Challenge (ARC-C; Clark et al., 2018), CommonsenseQA (CSQA; Talmor et al., 2019), and LogiQA (Liu et al., 2020), with 1,172/1,221/651 evaluation queries. A disjoint 128-example pool supplies demonstrations. Reference acquisition and displacement reduction use four-shot prompts, block-14/16 interventions (Qwen/Llama), a width-64 GELU bottleneck, seeds 17/29/43, and AdamW at 10^{-4} for up to five epochs. The default metric is training-state inverse covariance with shrinkage 10^{-3}; Appendix J gives splits, prompts, optimization, and controls.

The primary cost is $C_{\mathrm{rel}} = \mathbb{E}\|v\|_G^2 / \mathbb{E}\|u\|_G^2$. Selection minimizes validation cost subject to mean answer TV $\leq .01$, P95 TV $\leq .05$, and reference-argmax disagreement $\leq 1\%$; accuracy measures utility. Global restricts the candidate to a shared scalar contraction, whereas Adaptive allows a context-dependent scalar. Coordinate allows context-dependent element-wise rescaling; it can change direction, but cannot mix coordinates or introduce components outside the coordinate-scaled reference. MFM permits general directional changes through its learned output subspace. These controls distinguish magnitude calibration, element-wise rescaling, and more general directional realization.

4.2 Directional Realization Beyond Contraction

Q1: displacement-cost savings at matched answer fidelity. Theorem 3.1 compares unrestricted directional modification with optimal pointwise scalar contraction and identifies silent components and unequal active sensitivities as sources of scalar-inaccessible redundancy. Coordinate gating tests whether the observed advantage can already be explained by element-wise rescaling. Table 1 asks how much reference displacement cost can be reduced while preserving its answer observation. MFM saves 57.6% of reference cost and 30.4% relative to coordinate gating in the macro-average; accuracy changes by only -0.29 to -0.10 percentage points while retaining most of the reference's 3.00–4.10-point gain. Table J.2 reports mean TV .0055–.0067 and disagreement .57–.68%. Cost follows Global $>$ Adaptive $>$ Coordinate $>$ MFM in all six settings. The Global-to-Adaptive reduction reflects context-dependent magnitude calibration; the further Coordinate reduction shows the benefit of element-wise rescaling, while MFM's remaining advantage shows that element-wise rescaling alone does not explain the observed savings. On Qwen/LogiQA, MFM also remains below the adaptive scalar baseline under alternative reference acquisition and Euclidean/diagonal displacement metrics (Table J.4).

4.3 Preserving the Reference Response Across Contexts

Q2: cross-shot transfer and contextual response. Minimal realization is context dependent: changing demonstrations alters both the model state and its sensitivity to intervention. We therefore

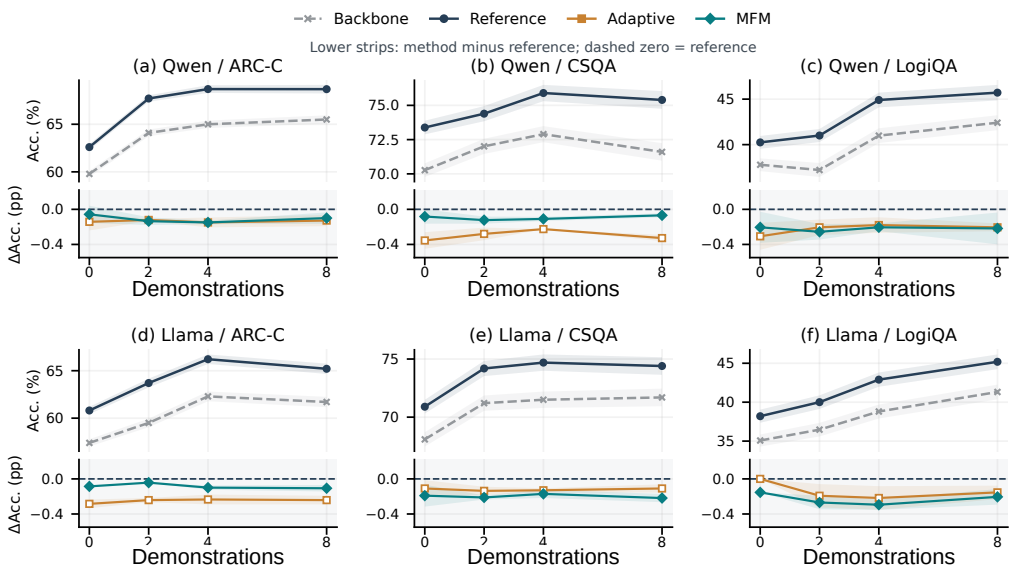

Figure 2: **Behavior across shot counts.** Upper: backbone/reference accuracy. Lower: adaptive/MFM minus reference; bands show paired standard deviations across seeds.

Table 2: **Contextual preservation.** Cost, TV, and disagreement (%) are averaged over the four shot-count conditions; S_{ctx} and response error are averaged over six four-shot demonstration pairs. A/M: adaptive scalar/MFM.

Task	S_{ctx}	C_{A}	C_{M}	TV_{M}	$\mathrm{Dis.}_{\mathrm{M}}$	Resp. error A/M
Qwen2.5-7B (Base)						
ARC-C	0.1163	0.724	0.368	0.0061	0.65	0.0108 / 0.0087
CSQA	0.1204	0.768	0.423	0.0064	0.60	0.0117 / 0.0094
LogiQA	0.1234	0.808	0.466	0.0076	0.66	0.0124 / 0.0100
Llama-3.1-8B (Base)						
ARC-C	0.1197	0.749	0.400	0.0066	0.72	0.0112 / 0.0095
CSQA	0.1177	0.720	0.389	0.0058	0.64	0.0109 / 0.0085
LogiQA	0.1267	0.836	0.512	0.0076	0.79	0.0125 / 0.0103

ask whether a compressor fitted on four-shot prompts preserves the *change in reference behavior induced by changing demonstrations*, rather than merely maintaining aggregate accuracy. We freeze the four-shot reference and compressor and evaluate the same queries with nested 0/2/4/8-shot prompts, without refitting. Figure 2 shows that displacement reduction preserves the reference's task-level trends, but accuracy can hide prompt-wise changes. For fixed query q, define $\delta_m = p_m(q, \mathcal{D}_a) - p_m(q, \mathcal{D}_b)$ and report $S_{\mathrm{ctx}} = \mathbb{E}\|\delta_u\|_1/2$ and $E_{\mathrm{resp}} = \mathbb{E}\|\delta_v - \delta_u\|_1/2$. The reference sensitivity $S_{\mathrm{ctx}} = .1163$–$.1267$ shows substantial demonstration effects; MFM response errors are .0085–.0103 and are below adaptive scaling in all six settings (Table 2), while cross-shot costs remain near their four-shot values. Response error is bounded by the sum of the two prompt-wise TV errors. Four-shot fidelity passes throughout; eight-shot LogiQA fails for one seed per model (11/12 seed–shot conditions pass, versus 12/12 elsewhere), identifying a concrete transfer boundary despite small accuracy changes (Table J.3). These results support contextual preservation without implying unrestricted cross-context invariance.

4.4 Shared Capacity and Behavioral Specification

Q3a: shared capacity and pointwise savings. Theorem 3.2 shows that pointwise savings need not be realizable by a finite-width shared module in the local linearized objective. The cross-context spectrum of the pointwise optimal realization field, rather than reference output rank alone, governs

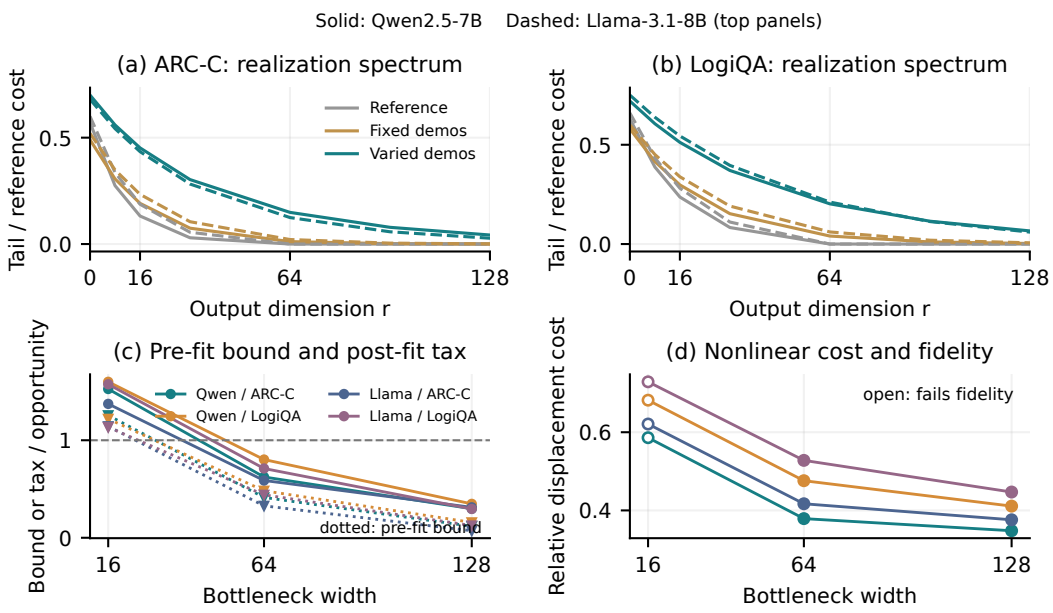

Figure 3: **Contextual-capacity diagnostics.** (a,b) Covariance tails normalized by reference cost. (c) Pre-fit bound $\lambda S_r/(2\Gamma_\lambda)$ (dotted) and post-fit tax T_λ/Γ_λ (solid). (d) Nonlinear cost and fidelity.

complexity. Across both backbones and ARC-C/LogiQA, widths $16/64/128$ share one reference and penalty. Figure 3(a,b) compares fixed and varied demonstrations; variation enlarges the realization spectral tail, including beyond the reference's output rank 64. Thus the pointwise optimal field can contain substantial cross-context variation outside the reference's own output rank. In panel (c), the local necessary ratio $\lambda S_r/(2\Gamma_\lambda)$ and the fitted tax T_λ/Γ_λ both exceed one at width 16, indicating a capacity obstruction to beating the pointwise scalar oracle in the linearized problem. At widths 64 and 128, the fitted local tax falls below the available opportunity. Independently, panel (d) shows that the corresponding width-16 nonlinear candidates fail the chosen fidelity criterion, whereas the wider modules pass at lower displacement. This qualitative alignment supports the capacity interpretation; Theorem 3.2 itself is a local comparison rather than a certificate of nonlinear fidelity.

Q3b: preserving more behavior consumes redundancy. Theorem 3.3 shows that refining an exactly preserved linear observation cannot increase redundancy and can incur an additional minimum-displacement cost. We therefore ask whether the same qualitative tradeoff appears in the trained nonlinear shared setting by additionally preserving answer-set probability mass before renormalization on both backbones and ARC-C/LogiQA. This quantity captures behavior missed by the renormalized answer distribution. Under unchanged answer-distribution budgets, macro-mean mass drift falls from 0.0275 to 0.0038, while cost rises from 0.434 to 0.566, retaining 43.4% reference-cost savings. Answer-only candidates pass the original budgets but fail the mass constraint; strengthened candidates pass both in all three seeds (Table J.5). This experiment does not estimate the exact pointwise price in Theorem 3.3; rather, it tests the theorem's qualitative implication in a finite-tolerance, nonlinear, shared setting. Appendix J.4 reports the tolerance sweep.

5 Conclusion

We studied *minimal behavioral realization*: how to reproduce the specified behavior of a frozen reference intervention with less hidden-state displacement. Our analysis separates pointwise directional redundancy from its cross-context realizability and shows that the attainable savings also depend on the behavioral observation being preserved. Across two model families and three multiple-choice reasoning benchmarks, MFM substantially reduces reference displacement cost while preserving answer fidelity and contextual response; strengthening the preserved observation to include answer-set probability mass reduces, but does not eliminate, these savings.

AI USE STATEMENT

Generative AI assisted with figure preparation and minor language polishing to improve the clarity and readability of the manuscript.

ETHICS STATEMENT

This work studies activation interventions on frozen language models. Its evaluation uses public benchmarks and existing models, whose use is subject to their respective licenses and terms. Behavioral preservation is evaluated using answer distributions and answer-set probability mass.

REPRODUCIBILITY STATEMENT

The intervention, objectives, and validation rule are specified in Section 3.6. Appendix J specifies optimization, data splits, selection procedure, and diagnostic settings; Appendices B–H provide local derivations, shared-capacity and preservation-price proofs, and finite-amplitude transfer conditions.

REFERENCES

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*, 2022.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36:66044–66063, 2023.

Dimitri P Bertsekas. *Convex optimization theory*, volume 1. Athena Scientific Belmont, 2009.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 6491–6506, 2021.

Toan Doan, Thin Nguyen, and Sunil Gupta. A geometric information bottleneck for activation steering. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 933–943, 2026.

Yuxin Dong, Jiachen Jiang, Zhihui Zhu, and Xia Ning. Understanding task vectors in in-context learning: Emergence, functionality, and limitations. In *International Conference on Learning Representations*, volume 2026, pp. 19393–19428, 2026.

Zijian Feng, Tianjiao Li, Zixiao Zhu, Hanzhang Zhou, Junlang Qian, Li Zhang, Chua Deryl, Lee Mak, Gee Ng, and Kezhi Mao. Fine-grained activation steering: Steering less, achieving more. In *International Conference on Learning Representations*, volume 2026, pp. 39421–39443, 2026.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in neural information processing systems*, 34:9574–9586, 2021.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Geert Heyman and Frederik Vandeputte. Steer like the llm: Activation steering that mimics prompting. *arXiv preprint arXiv:2605.03907*, 2026.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Bruce W Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *International conference on learning representations*, volume 2025, pp. 90960–90985, 2025.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in neural information processing systems*, 36:41451–41530, 2023.

Sirui Liang, Pengfei Cao, Jian Zhao, Cong Huang, Jun Zhao, and Kang Liu. Bias-restrained prefix representation finetuning for mathematical reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31916–31924, 2026.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*, 2020.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*, 2022.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. Fast model editing at scale. *arXiv preprint arXiv:2110.11309*, 2021.

Tam Nguyen, Tu Anh Nguyen, Sina Alemohammad, and Richard G Baraniuk. Minimizing collateral damage in activation steering. *arXiv preprint arXiv:2605.01167*, 2026.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 7237–7256, 2020.

Shauli Ravfogel, Grusha Prasad, Tal Linzen, and Yoav Goldberg. Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pp. 194–209, 2021.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D Cotterell. Linear adversarial concept erasure. In *International Conference on Machine Learning*, pp. 18400–18421. PMLR, 2022.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

John T Robertson, Jianing Zhu, Haris Vikalo, and Zhangyang Wang. When is rank-1 steering cheap? geometry, granularity, and budgeted search. *arXiv preprint arXiv:2605.16362*, 2026.

Leheng Sheng, Changshuo Shen, Weixiang Zhao, Junfeng Fang, Xiaohao Liu, Zhenkai Liang, Xiang Wang, An Zhang, and Tat-Seng Chua. Alphasteer: Learning refusal steering with principled null-space constraint. In *International Conference on Learning Representations*, volume 2026, pp. 25861–25904, 2026.

Julian Skifstad, Xinyue Annie Yang, and Glen Chou. Local linearity of llms enables activation steering via model-based linear optimal control. *arXiv preprint arXiv:2604.19018*, 2026.

Jiuding Sun, Sidharth Baskaran, Zhengxuan Wu, Michael Sklar, Christopher Potts, and Atticus Geiger. Hypersteer: Activation steering at scale with hypernetworks. *arXiv preprint arXiv:2506.03292*, 2025.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, 2019.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron Wallace, and David Bau. Function vectors in large language models. In *International conference on learning representations*, volume 2024, pp. 17282–17333, 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Sohan Venkatesh and Ashish Mahendran Kurapath. On the non-identifiability of steering vectors in large language models. *arXiv preprint arXiv:2602.06801*, 2026.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401, 2020.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.

Sihan Wang, Jiayi Zhao, Qingyan Cao, Hongbo Yao, and Lin Shu. Fishback: Pullback fisher geometry for optimal activation steering in transformers. *arXiv preprint arXiv:2605.17231*, 2026.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. Reft: Representation finetuning for language models. *Advances in Neural Information Processing Systems*, 37:63908–63962, 2024.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.

Wannan Yang, Xinchi Qiu, Lei Yu, Yuchen Zhang, Aobo Yang, Narine Kokhlikyan, Nicola Cancedda, and Diego Garcia-Olano. Hallucination reduction with casal: Contrastive activation steering for amortized learning. In *International Conference on Learning Representations*, volume 2026, pp. 10611–10665, 2026.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Behavioral Equivalence, Minimal Realization, and Coordinate Conventions

A.1 Behavioral equivalence and the nonlinear minimal-realization object

Fix an input x, intervention site, baseline state z, contextual downstream state c, observation F_c, observation metric $\Omega \succ 0$, and hidden-state metric $G \succ 0$. For exact preservation define the equivalence class induced by a reference intervention u,

$$\mathcal{E}_0(x;u) = \{v : F_c(z+v) = F_c(z+u)\}.$$

For tolerance $\epsilon > 0$ we use the feasible set

$$\mathcal{E}_\epsilon(x;u) = \{v : \|F_c(z+v) - F_c(z+u)\|_\Omega \le \epsilon\}.$$

The nonlinear minimum cost and redundancy are defined in Eqs. 1–2. Since $u \in \mathcal{E}_\epsilon(x;u)$ for all $\epsilon \ge 0$, redundancy is nonnegative. Equivalence concerns the specified observation at a fixed site and input; collateral drift, transfer, and robustness are separate outcomes.

A.2 Primary and strengthened observations

For a fixed answer-token set $\mathcal{A} = \{a_1, \dots, a_K\}$, the primary observation is the answer-renormalized distribution

$$f_A(x)_k = \frac{\exp \ell_{A,a_k}(x)}{\sum_j \exp \ell_{A,a_j}(x)}.$$

Comparisons share a verbalizer mapping and use $\Omega = I$ by default. Exact equality defines equivalence; evaluation tolerances define the acceptable operating region.

The augmented observation additionally constrains answer-set mass:

$$\bar{f}_A(x) = \left[f_A(x); \sum_{k \in \mathcal{A}} P_A(k \mid x) \right]. \tag{27}$$

Probabilities assigned to the remaining vocabulary, as well as generated traces, remain unconstrained. For answer distributions, $\mathrm{TV}(p,p') \le \sqrt{K}\|p-p'\|_2/2$. A reference argmax of margin $\gamma > 0$ is retained if $\|p-p'\|_2 < \gamma/2$. Mean discrepancy alone does not guarantee retention of every low-margin example.

A.3 Observation refinement cannot create redundancy

Let $\widetilde{F}_c = (F_c, R_c)$ append additional coordinates. At finite amplitude,

$$\widetilde{F}_c(z+v) = \widetilde{F}_c(z+u) \quad \Longrightarrow \quad F_c(z+v) = F_c(z+u),$$

so $\mathcal{E}_0^{\widetilde{F}}(x;u) \subseteq \mathcal{E}_0^{F}(x;u)$: exact minimum cost cannot decrease, and redundancy cannot increase. At nonzero tolerance, this inclusion requires refined feasibility to imply original feasibility; independently rescaled tolerances need not satisfy it.

Locally, if $J_R = \partial R_c/\partial z|_z$, then $\widetilde{J} = [J; J_R]$ and

$$\ker \widetilde{J} = \ker J \cap \ker J_R \subseteq \ker J, \qquad \{v : \widetilde{J}v = \widetilde{J}u\} \subseteq \{v : Jv = Ju\}.$$

This proves Eq. 7. Appendix J.4 evaluates how displacement savings change when answer-set probability mass is added to the observation.

A.4 Coordinate consistency in observation and hidden-state spaces

Write $J_{\mathrm{raw}} = \partial F_c/\partial z$ and let the local discrepancy be $\|J_{\mathrm{raw}}\delta\|_\Omega^2$. Under a hidden-state reparameterization $\widetilde{z} = Az$, the corresponding quantities are $\widetilde{J}_{\mathrm{raw}} = J_{\mathrm{raw}}A^{-1}$ and $\widetilde{G} = A^{-\top}GA^{-1}$. The

bijection $v \mapsto Av$ preserves both behavioral discrepancy and hidden-state cost. Under an invertible reparameterization of the same observation, $\widetilde{F} = CF$, take $\widetilde{\Omega} = C^{-\top}\Omega C^{-1}$; then

$$(CJ_{\mathrm{raw}}\delta)^\top \widetilde{\Omega}(CJ_{\mathrm{raw}}\delta) = (J_{\mathrm{raw}}\delta)^\top \Omega(J_{\mathrm{raw}}\delta).$$

Whitening with $L^\top L = \Omega$ gives $J = LJ_{\mathrm{raw}}$. The generalized sensitivities and scalar gap thus depend on the specified metrics, not arbitrary coordinates. Changing Ω or G defines a different minimal-realization problem.

For $C_z \succ 0$, the inverse covariance transforms consistently. Re-estimating $(\widehat{C}_z + \epsilon s^2 I)^{-1}$ need not, since isotropic shrinkage selects coordinates. A pseudoinverse may lose strict convexity; we use positive shrinkage with $s^2 > 0$. Neither this metric nor input standardization implies invariance of the full learning algorithm.

B Classical Local Geometry

Let $w = G^{1/2}v$, $w_u = G^{1/2}u$, and $B = JG^{-1/2}$. The exact local problem is $\min_w \|w\|_2^2/2$ subject to $Bw = Bw_u$. Its unique solution is

$$w_0 = B^\dagger B w_u = B^\top (BB^\top)^\dagger B w_u.$$

The residual $w_u - w_0 \in \ker B$ is orthogonal to w_0, proving Eq. 8. This identifies first-order redundancy.

Take a thin SVD $B = U\Sigma V^\top$ with positive singular values $\sigma_1, \ldots, \sigma_r$. Writing $w_u = Va + w_\perp$ and $w = Vb + w_{\mathrm{null}}$, the regularized objective is

$$Q_\lambda(G^{-1/2}w) = \tfrac{1}{2}\sum_i \sigma_i^2(b_i - a_i)^2 + \tfrac{\lambda}{2}\left(\sum_i b_i^2 + \|w_{\mathrm{null}}\|_2^2\right).$$

Its unique minimizer has $b_i = t_i a_i$, $w_{\mathrm{null}} = 0$, where $t_i = \sigma_i^2/(\sigma_i^2 + \lambda)$. This proves Eq. 9. The quantities used in the main results are

$$q^2 = \sum_i \sigma_i^2 a_i^2, \qquad a^2 = \sum_i a_i^2 + \nu^2, \tag{28}$$

$$r_\lambda^2 = \sum_i \sigma_i^2(1 - t_i)^2 a_i^2, \qquad d_\lambda^2 = \sum_i t_i^2 a_i^2, \tag{29}$$

$$\ell_\lambda^2 = \sum_i (1 - t_i)^2 a_i^2 + \nu^2. \tag{30}$$

The scalar $a = \|u\|_G$ differs from the spectral coefficients a_i. We have $d_\lambda \le \|v_0\|_G \le a$ and $v_\lambda \to v_0$ as $\lambda \downarrow 0$. If $Ju = 0$, including rank-zero B, the active coefficients vanish and $v_\lambda = v_0 = 0$.

C Optimal Scalar Comparison and Equality Cases

C.1 Proof of Theorem 3.1

For $a > 0$,

$$Q_\lambda(\alpha u) = \tfrac{1}{2}(1 - \alpha)^2 q^2 + \tfrac{\lambda}{2}\alpha^2 a^2.$$

Its minimizer is $\alpha_\lambda = q^2/(q^2 + \lambda a^2)$, in $(0, 1)$ when $q > 0$. Define $A_\lambda = B^\top B + \lambda I \succ 0$. Completing the square about $w_\lambda = G^{1/2}v_\lambda$ gives, for any w,

$$Q_\lambda(G^{-1/2}w) - Q_\lambda(v_\lambda) = \tfrac{1}{2}(w - w_\lambda)^\top A_\lambda (w - w_\lambda). \tag{31}$$

Substitute $w = \alpha_\lambda w_u$. Active eigenvalues of A_λ are $\sigma_i^2 + \lambda$, while its eigenvalue on $\ker B$ is λ. This proves Eq. 10.

For $q > 0$, $\alpha_\lambda > 0$, so zero gap requires $\nu = 0$ and $t_i = \alpha_\lambda$ for every occupied direction $a_i \neq 0$. The map $s \mapsto s/(s + \lambda)$ is strictly increasing for $s > 0$; therefore all occupied σ_i^2 must coincide. Conversely, if they all equal $s > 0$ and $\nu = 0$, then $q^2 = sa^2$ and $\alpha_\lambda = s/(s + \lambda) = t_i$, so the gap is zero. This condition is basis independent: w_u lies entirely in a single positive-eigenvalue eigenspace of $B^\top B$. Unoccupied singular values do not affect it. If $q = 0$, zero is both the unrestricted optimum and the optimum scalar realization; if $u = 0$, every scalar represents the same zero vector.

Matched fidelity. For $q > 0$ and finite $\lambda > 0$, Eq. 30 gives $0 < r_\lambda < q$. Scalar discrepancy is $(1-\alpha)q$, so the cheapest feasible scalar at budget r_λ is $\bar{\alpha}_\lambda = 1 - r_\lambda/q$. The vector v_λ uniquely minimizes squared G-cost subject to $\|J(v-u)\|_2 \le r_\lambda$: otherwise a feasible vector of smaller cost would have smaller Q_λ, or a different vector of equal cost would have no larger Q_λ, both contradicting uniqueness. Hence

$$\|v_\lambda\|_G^2 \le \min_{\alpha\in[0,1]:\ \|J(\alpha u-u)\|_2 \le r_\lambda} \|\alpha u\|_G^2, \qquad r_\lambda = \|J(v_\lambda - u)\|_2. \tag{32}$$

Equality holds if and only if $v_\lambda = \bar{\alpha}_\lambda u$. Since $\bar{\alpha}_\lambda > 0$, this requires $\nu = 0$ and all occupied t_i equal; the preceding argument supplies the converse. Under that condition $\bar{\alpha}_\lambda = \alpha_\lambda$. Outside it, the two scalars need not agree: one minimizes a penalized objective, the other minimizes cost at a fixed discrepancy.

C.2 A QUANTITATIVE COST GAP AND ILLUSTRATIVE CASES

For the matched-fidelity cost gap, both $\bar{\alpha}_\lambda u$ and v_λ have residual norm r_λ, hence

$$a^2\bar{\alpha}_\lambda^2 - d_\lambda^2 = \lambda^{-1}(\bar{\alpha}_\lambda w_u - w_\lambda)^\top A_\lambda(\bar{\alpha}_\lambda w_u - w_\lambda) \ge \|\bar{\alpha}_\lambda w_u - w_\lambda\|_2^2. \tag{33}$$

The penalized gap also avoids subtraction of nearly equal objectives:

$$2\Delta_\lambda = \sum_i (\sigma_i^2 + \lambda)(\alpha_\lambda - t_i)^2 a_i^2 + \lambda\alpha_\lambda^2\nu^2.$$

These are pointwise local gains; Appendix D gives the shared and matched-mean-fidelity counterparts.

Three linear cases isolate the conclusion. With $G = I$, $B = [1\ 0]$ and $u = (1,1)$, silent displacement makes the gap strict. With $B = \mathrm{diag}(1,2)$ and $u = (1,1)$, there is no kernel, but unequal occupied sensitivities again make it strict. With $B = [1\ 0]$ and $u = (1,0)$, the gap is zero despite the nontrivial kernel. Scaling and adding an interior baseline embeds these examples locally in a probability simplex. They isolate the two sources of advantage and the equality regime.

D OPPORTUNITY–REALIZATION DECOMPOSITION FOR A SHARED MODULE

For each x, abbreviate

$$q_{\lambda,x}(v) = \tfrac{1}{2}\|J_x(v-u_x)\|_2^2 + \tfrac{\lambda}{2}\|v\|_G^2, \qquad A_{\lambda,x} = J_x^\top J_x + \lambda G.$$

Because $G \succ 0$ and $\lambda > 0$, $A_{\lambda,x} \succ 0$. The first-order condition is $A_{\lambda,x}v = J_x^\top J_x u_x$, so its unique minimizer is $v_\lambda^{\mathrm{pt}}(x)$ from Eq. 12. Completing the square in the original coordinates gives, for every v,

$$q_{\lambda,x}(v) - q_{\lambda,x}(v_\lambda^{\mathrm{pt}}) = \tfrac{1}{2}(v - v_\lambda^{\mathrm{pt}})^\top A_{\lambda,x}(v - v_\lambda^{\mathrm{pt}}) = \tau_\lambda(x; v). \tag{34}$$

Taking expectations proves

$$\mathcal{Q}_\lambda(v) = \mathbb{E}_x q_{\lambda,x}(v_\lambda^{\mathrm{pt}}(x)) + T_\lambda(v). \tag{35}$$

This holds for any measurable field with finite terms, without linearity, convexity, or per-example parameters in the shared class.

D.1 MODULE-VISIBLE AND ARCHITECTURE-RESTRICTED REALIZATION

Let $\mathcal{G}_z$ be the class of square-integrable measurable functions of z with finite weighted tax, and let $\mathcal{M} \subseteq \mathcal{G}_z$ be the architecture. The following definitions isolate idealized measurability, architecture, and candidate errors:

$$\mathcal{I}_\lambda(z) = \inf_{g\in\mathcal{G}_z} T_\lambda(g), \tag{36}$$

$$\mathcal{R}_\lambda(\mathcal{M}) = \inf_{m\in\mathcal{M}} T_\lambda(m) - \mathcal{I}_\lambda(z), \tag{37}$$

$$\mathcal{E}_\lambda(v; \mathcal{M}) = T_\lambda(v) - \inf_{m\in\mathcal{M}} T_\lambda(m). \tag{38}$$

For finite infima, these nonnegative terms sum to $T_\lambda(v)$. Theorem 3.2 bounds affine-output classes from the reference and downstream geometry.

D.2 Proof of Proposition 3.1

For the scalar class $\mathcal{S}$, the optimization is separable over examples. When $q_x = \|J_x u_x\|_2 > 0$, Theorem 3.1 gives the unique pointwise scalar minimizer and its excess $\Delta_\lambda(x)$. When $q_x = 0$, the pointwise ridge and scalar optima are both zero and we set $\Delta_\lambda(x) = 0$. Under the stated integrability and measurability conditions,

$$\inf_{s\in\mathcal{S}} \mathcal{Q}_\lambda(s) = \mathbb{E}_x q_{\lambda,x}(v_\lambda^{\mathrm{pt}}(x)) + \mathbb{E}_x \Delta_\lambda(x) = \mathbb{E}_x q_{\lambda,x}(v_\lambda^{\mathrm{pt}}(x)) + \Gamma_\lambda. \tag{39}$$

Subtracting Eq. 35 proves the first equality in Eq. 14. For the actual z-only architecture $\mathcal{M} \subseteq \mathcal{G}_z$, the definitions in Eqs. 36–38 telescope:

$$T_\lambda(v) = \mathcal{I}_\lambda(z) + \mathcal{R}_\lambda(\mathcal{M}) + \mathcal{E}_\lambda(v;\mathcal{M}). \tag{40}$$

Nested infima make all terms nonnegative. Thus $T_\lambda(v) < \Gamma_\lambda$ is both necessary and sufficient for the candidate to beat the pointwise scalar oracle in the population linearized objective. Here $\mathcal{E}_\lambda$ measures excess local tax relative to the best member of $\mathcal{M}$, whereas Stage 2 minimizes the nonlinear objective.

For the matched-mean-fidelity statement, let $\zeta_\lambda = \Gamma_\lambda - T_\lambda(v) > 0$. Then every $s \in \mathcal{S}$ obeys

$$\mathcal{Q}_\lambda(s) - \mathcal{Q}_\lambda(v) \geq \zeta_\lambda. \tag{41}$$

Since $\mathcal{Q}_\lambda = (R + \lambda C)/2$, if $R(s) \leq R(v)$ then

$$\frac{\lambda}{2}\big(C(s) - C(v)\big) = \mathcal{Q}_\lambda(s) - \mathcal{Q}_\lambda(v) + \frac{1}{2}\big(R(v) - R(s)\big) \tag{42}$$

$$\geq \zeta_\lambda. \tag{43}$$

Equivalently,

$$C(s) \geq C(v) + 2\zeta_\lambda/\lambda. \tag{44}$$

If instead $C(s) \leq C(v)$, the same identity gives $R(s) - R(v) \geq 2\zeta_\lambda$. Thus a positive margin guarantees strict improvement in one coordinate whenever a scalar matches or improves the other.

D.3 Optimal module-visible amortization and conditioning sufficiency

Let $\mathcal{G}_z$ be the class of square-integrable measurable functions of the module input z, and define

$$\bar{A}_\lambda(z) = \mathbb{E}[A_{\lambda,x} \mid z], \qquad b_\lambda(z) = \mathbb{E}[A_{\lambda,x} v_\lambda^{\mathrm{pt}}(x) \mid z].$$

Because $\bar{A}_\lambda(z) \succeq \lambda G$, it is invertible almost surely.

Corollary D.1 (Bayes-optimal z-only amortizer and sufficiency criterion)**.** *Assume $\mathbb{E}\|A_{\lambda,x}\| < \infty$ and $\mathbb{E}[(v_\lambda^{\mathrm{pt}})^\top A_{\lambda,x} v_\lambda^{\mathrm{pt}}] < \infty$. The unique minimizer of T_λ over $\mathcal{G}_z$ is*

$$g_\lambda^{\mathrm{Bayes}}(z) = \bar{A}_\lambda(z)^{-1} b_\lambda(z).$$

Its residual tax is exactly the unrestricted module-input floor from Eq. 36:

$$\mathcal{I}_\lambda(z) = \tfrac{1}{2}\mathbb{E}\|v_\lambda^{\mathrm{pt}}(x) - g_\lambda^{\mathrm{Bayes}}(z(x))\|^2_{A_{\lambda,x}}. \tag{45}$$

Moreover, $\mathcal{I}_\lambda(z) = 0$ if and only if $v_\lambda^{\mathrm{pt}}(x)$ is almost surely measurable with respect to z. Architecture restrictions add $\mathcal{R}_\lambda(\mathcal{M}) \geq 0$; richer conditioning can only decrease $\mathcal{I}_\lambda$.

D.4 Proof of Corollary D.1

Consider the unrestricted class $\mathcal{G}_z$ of square-integrable measurable functions $g(z)$. Conditioning the realization tax on z gives

$$2\,\mathbb{E}[\tau_\lambda(x;g) \mid z] = \mathbb{E}[(g - v_\lambda^{\mathrm{pt}})^\top A_{\lambda,x}(g - v_\lambda^{\mathrm{pt}}) \mid z] \tag{46}$$

$$= g^\top \bar{A}_\lambda(z) g - 2 g^\top b_\lambda(z) + c(z), \tag{47}$$

where $c(z)$ is independent of g. Because $A_{\lambda,x} \succeq \lambda G$, also $\bar{A}_\lambda(z) \succeq \lambda G \succ 0$. The conditional quadratic therefore has the unique minimizer $g_\lambda^{\mathrm{Bayes}}(z) = \bar{A}_\lambda(z)^{-1} b_\lambda(z)$. Substitution and expectation yield Eq. 45. Restricting to $\mathcal{M} \subseteq \mathcal{G}_z$ adds $\mathcal{R}_\lambda(\mathcal{M})$; richer conditioning enlarges the measurable class and cannot increase the infimum. Since $A_{\lambda,x} \succ 0$, the floor is zero exactly when the ridge field is almost surely measurable from z. An irregular measurable target may still be difficult for a continuous, Lipschitz, or finite neural architecture, contributing to $\mathcal{R}_\lambda(\mathcal{M})$. Nearby states alone do not imply information loss.

E Capacity of a Shared Realization Field

Fix the reference u, metric G, observation, distribution, and $\lambda > 0$ across widths. In hidden-whitened coordinates, set

$$H_x = B_x^\top B_x + \lambda I, \qquad w^* = G^{1/2} v_\lambda^{\mathrm{pt}}, \qquad \bar{w} = \mathbb{E} w^*, \qquad \Sigma_* = \mathrm{Cov}(w^*).$$

Then $T_\lambda(v) = \mathbb{E}[(w - w^*)^\top H_x (w - w^*)]/2$. Here H_x is a hidden-space Hessian; it is distinct from the observation-space matrix $H = JG^{-1}J^\top$ used in the local solution.

E.1 Lower bound for every affine-output module

Suppose $w(x)$ lies in $b + \mathrm{range}(U)$, where U has at most r orthonormal columns. Pointwise projection yields

$$\|w(x) - w^*(x)\|^2 \geq \|(I - UU^\top)(w^*(x) - b)\|^2.$$

After expectation, the right-hand side is

$$\mathrm{tr}[(I - UU^\top)\Sigma_*] + \|(I - UU^\top)(\bar{w} - b)\|^2.$$

Minimizing over the offset and subspace gives the sum of discarded covariance eigenvalues, S_r. Since $H_x \succeq \lambda I$, this proves Eq. 17. Invertible multiplication by $G^{1/2}$ preserves affine dimension, so the bound applies directly to $W_2 h(z) + b_2$ with r columns in W_2. The input map h may be nonlinear. Combining with Eq. 14 gives

$$\sup_{v \in \mathcal{M}_r} \left\{ \inf_{s \in \mathcal{S}} \mathcal{Q}_\lambda(s) - \mathcal{Q}_\lambda(v) \right\} \leq \Gamma_\lambda - \tfrac{\lambda}{2} S_r \tag{48}$$

for every subclass $\mathcal{M}_r$ of affine-output rank-r modules. This necessary capacity condition uses only $H_x \succeq \lambda I$; the construction below gives a sufficient condition.

A useful population summary, when $\Gamma_\lambda > 0$, is

$$r_{\mathrm{necessary}} = \min\{r \in \{0, \ldots, d\} : \lambda S_r / 2 < \Gamma_\lambda\}.$$

Every strictly successful local affine-output candidate has rank at least $r_{\mathrm{necessary}}$. Coefficient prediction can require additional width. When $\Gamma_\lambda = 0$, no candidate can strictly beat the local scalar oracle.

E.2 Constructive upper bound and coefficient prediction

Fix any choice of top-r eigenvectors U_r, including a fixed choice within repeated-eigenvalue subspaces. Decompose

$$w^* = \bar{w} + U_r a_r + e_\perp, \qquad U_r^\top e_\perp = 0, \qquad \mathbb{E}\|e_\perp\|^2 = S_r.$$

For $w_f = \bar{w} + U_r f(z)$, pointwise orthogonality gives

$$\mathbb{E}\|w_f - w^*\|^2 = S_r + \mathbb{E}\|f(z) - a_r\|^2. \tag{49}$$

If $H_x \preceq \beta I$ almost surely, multiplying this identity by $\beta/2$ proves Eq. 18. The same error is also a lower bound after multiplication by $\lambda/2$, for this fixed-subspace candidate. The strict sufficient inequality in Theorem 3.2 follows from Proposition 3.1.

Let $\bar{a}_r(z) = \mathbb{E}[a_r \mid z]$. Expanding $f - a_r = (f - \bar{a}_r) + (\bar{a}_r - a_r)$ and taking conditional expectation kills the cross term, proving Eq. 19. Thus even $S_r = 0$ need not ensure low tax. The GELU input map must represent the coefficient predictor, so output rank alone is insufficient. Exact weighted tax can be evaluated without a deterministic β bound; finite-sample PCA and coefficient fitting estimate the population construction.

E.3 WHY IDENTICAL POINTWISE OPPORTUNITIES NEED NOT SHARE A REALIZATION

A constant reference can require a dispersed oracle field. Let $G = I$, $u = e_0 \in \mathbb{R}^{n+1}$, and take uniformly sampled indices $i \in \{1, \dots, n\}$ and signs $s \in \{-1, 1\}$ with

$$B_{i,s} = (e_0 + se_i)^\top/\sqrt{2}, \qquad t = (1+\lambda)^{-1}.$$

Every input has the same singular value, reference norm, occupied silent energy, and positive scalar gap. Nevertheless,

$$w^*_{i,s} = \tfrac{t}{2}(e_0 + se_i), \quad \bar{w} = \tfrac{t}{2}e_0, \quad \Sigma_*|_{\mathrm{span}\{e_1,\dots,e_n\}} = \tfrac{t^2}{4n} I_n.$$

Thus $S_r = t^2(n-r)/(4n)$ for $0 \le r \le n$. The scalar gap is $\Gamma_\lambda = \lambda/[4(1+\lambda)(1+2\lambda)]$, independent of n. Pointwise opportunity therefore does not determine the width needed to approximate the oracle. The next construction establishes impossibility at the reference rank.

Corollary E.1 (Positive pointwise opportunity can be inaccessible at the reference rank)**.** *There exist finite input distributions, a rank-one affine-output reference, and positive directional opportunity on every input, for which no rank-one affine-output compressor beats the scalar oracle in the local objective, even when the module input identifies each example.*

Proof. Let $\lambda = 1$, $G = I$, and take three equally likely inputs with references $u_i = (1, s_i)$, $s_i \in \{0, 1, 2\}$. These lie on an affine line. Initially set $B_i = \sigma_i u_i^\top/\|u_i\|$, with $\sigma_i^2 \in \{1/4, 4, 1/4\}$ in that order. Then $w_i^* = t_i u_i$ with $t_i \in \{1/5, 4/5, 1/5\}$. The three oracle points $(1/5, 0), (4/5, 4/5), (1/5, 2/5)$ are not collinear. Therefore $S_1 > 0$, whereas $\Gamma_1 = 0$ by the scalar equality condition. Rotate each row direction by the same sufficiently small nonzero angle δ in the plane, keeping its norm. Each reference now has both a nonzero active and a nonzero silent component, hence a strictly positive scalar gap. The oracle covariance and Γ_1 vary continuously with δ. Therefore for sufficiently small $|\delta| > 0$, $S_1(\delta)/2 > \Gamma_1(\delta) > 0$. Theorem 3.2 rules out every rank-one affine-output compressor. The three inputs may be distinct and fully observable; the obstruction is output capacity, not hidden input information. □

The reference can be realized by a width-one GELU module on a suitable finite set of scalar inputs: choose nonnegative inputs whose GELU values are $0, 1, 2$, output column $(0, 1)^\top$, and bias $(1, 0)^\top$. These are local linear observation examples; scaled zero-sum coordinates around an interior simplex point embed them in differentiable probability observations, with consistent rescaling of the local parameters.

F THE COST OF ADDITIONAL BEHAVIORAL CONSTRAINTS

F.1 PROOF OF THEOREM 3.3

All quantities are pointwise and use the same fixed hidden metric. Let $P_N = I - B^\dagger B$ and $w_0 = B^\dagger B w_u$. The original feasible set is $w_0 + \ker B$, and $w_0 \perp \ker B$. The refined minimum therefore equals

$$\|w_0\|^2 + \min_{n \in \ker B} \{\|n\|^2 : Dn = d_R\}, \qquad d_R = D(w_u - w_0).$$

Since $w_u - w_0 \in \ker B$, the added constraint is feasible even if $DP_N D^\top$ is singular. Choose an orthonormal basis N of $\ker B$; then $n = N\xi$, $P_N = NN^\top$, and the minimum-norm solution of $DN\xi = d_R$ is

$$n_* = P_N D^\top (DP_N D^\top)^\dagger d_R.$$

Its squared norm is $d_R^\top (DP_N D^\top)^\dagger d_R$. Because $d_R \in \mathrm{range}(DP_N D^\top)$, this is zero iff $d_R = 0$. The feasible reference residual has squared norm $\|w_u - w_0\|^2$, giving both the upper bound and the refined solution in the theorem.

If $DN = U\,\mathrm{diag}(\gamma_j)V^\top$ is a thin SVD, the price is

$$\Pi_{R|F} = \sum_{\gamma_j > 0} \frac{(U_j^\top d_R)^2}{\gamma_j^2}. \tag{50}$$

Deficit and sensitivity both depend on the reference, bounding $\Pi_{R|F}$ by its original exact local redundancy.

F.2 Nested specifications and tolerance

For nested linear observations $B_0, B_1, \dots, B_m$, each formed by stacking additional rows, let c_j be the exact minimum squared cost. Then $c_j - c_{j-1} \geq 0$ and

$$\sum_{j=1}^{m}(c_j - c_{j-1}) = c_m - c_0.$$

The endpoint is independent of the order of adding a fixed set of constraints; the individual prices generally are not. For example, with $B_0 = [1\ 0]$, $w_u = (a, b)$, and two identical added rows $[0\ 1]$, the first costs b^2 and the second costs zero. Prices do not generally add if each is computed relative to B_0 alone. The refined oracle field has its own covariance, so its capacity spectrum must also be recomputed.

For tolerance effects, take $B = [1\ 0]$, $w_u = (a, b)$, and $D = (c, \eta)$. For every $\eta \neq 0$, exact refinement costs b^2, even if $|\eta|$ is tiny. If the primary constraint remains exact but the additional discrepancy may be at most ϵ, its cost is instead

$$\Pi_\epsilon = (|b| - \epsilon/|\eta|)_+^2, \qquad \eta \neq 0. \tag{51}$$

Indeed, the new variable is $n = (0, t)$ and the constraint is $|\eta(t - b)| \leq \epsilon$. The nearest feasible t to zero gives the formula. At $\eta = 0$ both exact and tolerant prices are zero. Thus finite tolerance can substantially change the price when the added observation has low sensitivity.

F.3 What the price measures

Transporting G under invertible hidden-coordinate changes preserves feasible costs and the exact price. Invertible transformations of the added observation also preserve its exact constraint. Changing the metric changes the price. The nonlinear exact minimum remains monotone under refinement, but the closed-form price applies only to the linear problem.

A finite-amplitude analytic example. In two dimensions, let $u_i = (x_i, b_i)$, $G = I$, $F(x, y) = x$, and $R_i(x, y) = \eta_i(y + a_i y^2)$ with positive coefficients. At baseline zero, the local extra-tolerance solution is $y_i = (b_i - \epsilon_R/\eta_i)_+$. The nonlinear solution is the nonnegative root of $y_i + a_i y_i^2 = (b_i + a_i b_i^2 - \epsilon_R/\eta_i)_+$. Restricting y to a shared constant further intersects the feasible intervals. This distinguishes local, nonlinear, and shared prices.

G Population Transfer to the Nonlinear Objective

G.1 A uniform scalar comparison

Let $w_x = G^{1/2}v(x)$, $w_{u,x} = G^{1/2}u(x)$, and $h_x(w) = LF_{c(x)}(z(x) + G^{-1/2}w)$. Suppose Dh_x is ρ_x-Lipschitz on an open convex neighborhood containing the origin, $w_{u,x}$, w_x, and their relevant segments. Then

$$\begin{aligned} h_x(w_x) - h_x(w_{u,x}) &= B_x(w_x - w_{u,x}) + r_x, \\ \|r_x\| &\leq \rho_x \max\{\|w_x\|, \|w_{u,x}\|\}\|w_x - w_{u,x}\| = e_x(v). \end{aligned}$$

Integrating the Jacobian along the segment and using $Dh_x(0) = B_x$ and convexity of the norm gives this bound. Define

$$\begin{aligned} K(v) &= \mathbb{E}\left[\|B_x(w_x - w_{u,x})\|e_x(v) + \tfrac{1}{2}e_x(v)^2\right], \\ K_{\mathcal{S}} &= \mathbb{E}\left[\rho_x\|B_x w_{u,x}\|\|w_{u,x}\|^2 + \tfrac{1}{2}\rho_x^2\|w_{u,x}\|^4\right]. \end{aligned} \tag{52}$$

For vectors a, r,

$$\left|\tfrac{1}{2}\|a + r\|^2 - \tfrac{1}{2}\|a\|^2\right| \leq \|a\|\|r\| + \tfrac{1}{2}\|r\|^2.$$

The displacement penalty is identical in $\mathcal{L}_\lambda$ and $\mathcal{Q}_\lambda$. Taking expectations proves $|\mathcal{L}_\lambda(v) - \mathcal{Q}_\lambda(v)| \leq K(v)$. For every scalar field $\alpha(x) \in [0, 1]$,

$$\|B_x(\alpha w_u - w_u)\| \leq \|B_x w_u\|, \qquad e_x(\alpha u) \leq \rho_x\|w_u\|^2.$$

Thus $K_{\mathcal{S}}$ bounds the entire measurable scalar class, giving

$$\inf_{s \in \mathcal{S}} \mathcal{L}_\lambda(s) - \mathcal{L}_\lambda(v) \geq \Gamma_\lambda - T_\lambda(v) - K(v) - K_{\mathcal{S}}. \tag{53}$$

G.2 Matched nonlinear fidelity and a capacity sufficient condition

Write $D(v) = \mathbb{E}\|F_c(z+v) - F_c(z+u)\|_\Omega^2$ and $C(v) = \mathbb{E}\|v\|_G^2$. If the right-hand side of Eq. 53 is $\zeta_{\mathrm{nl}} > 0$, then for every $s \in \mathcal{S}$ with $D(s) \le D(v)$,

$$C(s) \ge C(v) + 2\zeta_{\mathrm{nl}}/\lambda.$$

Rearranging $\mathcal{L}_\lambda = (D + \lambda C)/2$ proves this as in the local case. It compares mean squared discrepancy; experiments additionally constrain TV, quantiles, and disagreement. For the constructive rank-r candidate from Theorem 3.2,

$$\Gamma_\lambda - \frac{\beta}{2}(S_r + \varepsilon_r(f)) - K(v_f) - K_{\mathcal{S}} > 0 \tag{54}$$

is sufficient. The spectral lower bound, by contrast, rules out local advantage only: nonlinear curvature can invalidate a local impossibility prediction. Applying the nonlinear certificate requires a valid curvature bound over the specified neighborhoods.

For linear observations $\rho_x = 0$, the curvature budgets vanish and the population identity is recovered. With bounded Jacobians and curvature, scaling a fixed reference and candidate by t makes the local margin $O(t^2)$ and the remainder budgets $O(t^3) + O(t^4)$. Therefore a strictly positive local margin persists for sufficiently small positive t. Evaluation at the deployed amplitude $t = 1$ tests whether this regime extends to the trained reference.

H Finite-Amplitude Transfer from Realization Tax

For a fixed input, let $w = G^{1/2}v$, $w_u = G^{1/2}u$, $B = JG^{-1/2}$, $A_\lambda = B^\top B + \lambda I$, and $w_\lambda = G^{1/2}v_\lambda$. Define

$$\tau_\lambda(v) = \tfrac{1}{2}(w - w_\lambda)^\top A_\lambda (w - w_\lambda), \qquad e_\tau = \sqrt{2\tau_\lambda(v)/\lambda}.$$

Let $d_\lambda = \|v_\lambda\|_G$, $a = \|u\|_G$, $q = \|Ju\|_2$, $r_\lambda = \|J(v_\lambda - u)\|_2$, and $\kappa = q - La^2$.

Theorem H.1 (Finite-amplitude transfer from realization tax)**.** *Choose the same observation factor L_Ω used in Eq. 4, with $L_\Omega^\top L_\Omega = \Omega$, and suppose $h(w) = L_\Omega F_c(z + G^{-1/2}w)$ has an L-Lipschitz Jacobian on an open convex neighborhood containing the relevant line segments. Then*

$$\|F_c(z+v) - F_c(z+u)\|_\Omega \le r_\lambda + \sqrt{2\tau_\lambda(v)} + \frac{L}{2}\left[a^2 + (d_\lambda + e_\tau)^2\right] =: \eta_\lambda(v), \tag{55}$$
$$\|v\|_G \le d_\lambda + e_\tau.$$

If $\kappa > 0$ and a nonlinear fidelity budget δ satisfies

$$\eta_\lambda(v) \le \delta < \kappa, \qquad d_\lambda + e_\tau < a(1 - \delta/\kappa), \tag{56}$$

then v meets that budget and has strictly smaller squared G-cost than every scalar contraction αu, $\alpha \in [0,1]$, meeting the same budget.

H.1 Proof of Theorem H.1

Using the theorem's notation and $A_\lambda = B^\top B + \lambda I$, we have

$$\tau_\lambda(v) = \tfrac{1}{2}(w - w_\lambda)^\top A_\lambda (w - w_\lambda)$$

and the positive-semidefinite inequalities $A_\lambda \succeq \lambda I$ and $A_\lambda \succeq B^\top B$,

$$\|w - w_\lambda\|_2 \le \sqrt{2\tau_\lambda(v)/\lambda} = e_\tau, \qquad \|B(w - w_\lambda)\|_2 \le \sqrt{2\tau_\lambda(v)}. \tag{57}$$

Hence $\|w\|_2 \le d_\lambda + e_\tau$ and

$$\|B(w - w_u)\|_2 \le \|B(w_\lambda - w_u)\|_2 + \|B(w - w_\lambda)\|_2 \le r_\lambda + \sqrt{2\tau_\lambda(v)}. \tag{58}$$

For a map with L-Lipschitz Jacobian, Taylor's integral remainder at the origin satisfies

$$h(w) = h(0) + Dh(0)w + R(w), \qquad \|R(w)\|_2 \le \tfrac{L}{2}\|w\|_2^2. \tag{59}$$

Here $Dh(0) = B$. Applying Eq. 59 to w and w_u, then using Eq. 58, gives

$$\|h(w) - h(w_u)\|_2 \le \|B(w - w_u)\|_2 + \|R(w)\|_2 + \|R(w_u)\|_2 \tag{60}$$

$$\le r_\lambda + \sqrt{2\tau_\lambda(v)} + \frac{L}{2}\left[(d_\lambda + e_\tau)^2 + a^2\right], \tag{61}$$

which is Eq. 55. The stated cost bound follows from $\|w\|_2 \le d_\lambda + e_\tau$.

For $\alpha \in [0, 1]$, integrating along the scalar path gives

$$h(w_u) - h(\alpha w_u) = (1 - \alpha)Dh(0)w_u + \int_\alpha^1 [Dh(tw_u) - Dh(0)]w_u \, dt, \tag{62}$$

$$\|h(w_u) - h(\alpha w_u)\|_2 \ge (1 - \alpha)q - \tfrac{L}{2}(1 - \alpha^2)a^2 \tag{63}$$

$$\ge (1 - \alpha)(q - La^2) = (1 - \alpha)\kappa. \tag{64}$$

If $\kappa > 0$ and a scalar has nonlinear discrepancy at most $\delta < \kappa$, then $\alpha \ge 1 - \delta/\kappa$ and therefore $\|\alpha u\|_G \ge a(1 - \delta/\kappa)$. Under Eq. 56, the candidate both meets the budget and has strictly smaller norm than this lower bound, proving the theorem.

Sharper optional remainder. The segment from w to w_u yields the alternative bound

$$\|h(w) - h(w_u)\|_2 \le r_\lambda + \sqrt{2\tau_\lambda(v)} + L \max\{a, d_\lambda + e_\tau\}(\ell_\lambda + e_\tau). \tag{65}$$

Under the common neighborhood assumptions, use the smaller curvature term in Eq. 55. Both bounds are sufficient, not necessary.

Exact pointwise-ridge limit. When the candidate equals the pointwise ridge solution, $\tau_\lambda(v) = 0$ and the new bound reduces to $r_\lambda + L(a^2 + d_\lambda^2)/2$, or the sharper segment version $r_\lambda + La\ell_\lambda$. As $\lambda \downarrow 0$, $r_\lambda \to 0$, $d_\lambda \to b = \|v_0\|_G$, and $\ell_\lambda \to \nu$. The nonempty regime below therefore needs no separate module-error constant. For shared candidates, realization tax accounts for the additional linearized residual and displacement.

H.2 A NONEMPTY NONLINEAR REGIME FOR POSITIVE REGULARIZATION

On a sufficiently small convex neighborhood of the origin, consider

$$h(w_1, w_2) = \big(1/2 + cw_1 + \beta w_2^2/2, 1/2 - cw_1 - \beta w_2^2/2\big), \quad c > 0, \quad \beta \ne 0.$$

Choose the neighborhood so its probabilities lie in $(0, 1)$; logits can be their logarithms. The Jacobian Lipschitz constant is $L = \sqrt{2}|\beta|$. For $w_u = (t, t)$, $t > 0$, and $G = I$, $w_0 = (t, 0)$, $a = \sqrt{2}t$, $b = \nu = t$, and $q = \sqrt{2}ct$. At $\lambda = 0$ and zero realization tax, the sharper segment remainder is $La\nu = 2|\beta|t^2$, while $\kappa(1 - b/a) = (\sqrt{2}ct - 2\sqrt{2}|\beta|t^2)(1 - 1/\sqrt{2})$. For sufficiently small positive t, the first quantity is strictly smaller than the second. Fix such a t. Continuity of $r_\lambda, d_\lambda, \ell_\lambda$ at zero implies that a nonempty budget interval persists for sufficiently small positive λ. By continuity, it also tolerates sufficiently small positive realization tax. This provides an analytic regime with nonzero curvature, regularization, and shared approximation error.

I FROM THE NONLINEAR MINIMAL-REALIZATION PROBLEM TO MFM

I.1 NONLINEAR OBJECTIVE AND LOCAL APPROXIMATION

Restricting Eq. 3 to $v_\phi(x) = m_\phi(z(x))$ and caching $p_u(x) = F_{c(x)}(z(x) + u(x))$ gives exactly the Stage-2 objective, Eq. 26.

The local theory is obtained by replacing both observations with first-order expansions about the same baseline:

$$F_c(z + v) - F_c(z + u) \approx J_{\mathrm{raw},x}(v - u).$$

Whitening by Ω and averaging over x gives $\mathcal{Q}_\lambda(v_\phi)$ in Eq. 13. Its pointwise ridge oracle defines the local opportunity; Proposition 3.1 measures the shared candidate's shortfall under the local Hessian geometry.

I.2 Penalty form versus constrained minimum realization

Write

$$D_{\mathrm{fit}}(\phi) = \frac{1}{2}\mathbb{E}\|F_c(z + m_\phi(z)) - p_u\|_\Omega^2, \qquad C(\phi) = \frac{1}{2}\mathbb{E}\|m_\phi(z)\|_G^2.$$

The Lagrangian of $\min C(\phi)$ subject to $D_{\mathrm{fit}}(\phi) \le \tau$, divided by a positive fidelity multiplier μ, yields $D_{\mathrm{fit}} + \lambda C$ with $\lambda = 1/\mu$, up to a constant. The neural objective is generally nonconvex, so the penalty path can cover only part of the constrained tradeoff. We therefore select operating points using explicit validation fidelity constraints.

I.3 What each realization-tax component means

Identical reference and student architectures make the reference reachable at Stage-2 initialization. The ridge field may nevertheless depend on context not measurable from z, giving $\mathcal{I}_\lambda(z) > 0$. When z identifies the field almost surely, this floor is zero even for irregular or hard-to-learn maps. Architecture restrictions contribute $\mathcal{R}_\lambda(\mathcal{M})$; fitting and nonlinear-objective mismatch contribute to $\mathcal{E}_\lambda(v_\phi; \mathcal{M})$.

I.4 Why scalar and coordinate-selective comparisons are both needed

The class $\mathcal{S}$ in Proposition 3.1 contains every measurable contraction $\alpha(x)u(x)$, including learned scalar gates. A positive margin $\Gamma_\lambda - T_\lambda(v)$ therefore rules out magnitude calibration alone, but does not distinguish reorientation from coordinate shrinkage. The gate $a_\theta(z) \odot u(x)$ tests this distinction: matching MFM supports shrinkage alone; a remaining MFM advantage at matched fidelity supports reorientation beyond element-wise rescaling.

I.5 Scope of the compression claim

Minimality is relative to the intervention site, observation, geometry, distribution, and candidate class. Behavioral matching retains the reference's predictions, including its mistakes. Task utility, collateral drift, generated reasoning, and inference cost therefore remain separate evaluation criteria.

J Experimental Details and Supplementary Results

Experiments used one server with eight NVIDIA GeForce RTX 4090 GPUs.

J.1 Data, prompts, hardware, and training

Table J.1: Data allocation used in the experiments. Training queries, demonstration pool, and tuning queries are disjoint. CSQA uses labeled development data only for final evaluation.

Task	Train available	Demo pool	Fit	Tune	Evaluate	Eval. split
ARC-C	1,119	128	991	299	1,172	test
CSQA	9,741	128	4,096	1,000	1,221	development
LogiQA	7,376	128	4,096	651	651	test

We used `allenai/ai2_arc` (`ARC-Challenge`), `tau/commonsense_qa`, and English `Train.txt`, `Eval.txt`, and `Test.txt` from LogiQA. We sorted splits by stable ID (file row for LogiQA) and permuted them with NumPy PCG64 seed 20260920. The first 128 training IDs formed the demonstration pool. For CSQA, the next 1,000 examples were used for tuning and the following 4,096 for fitting; for LogiQA, the next 4,096 were used for fitting; all remaining ARC-C training items were used. Dataset and model revisions were frozen before training, and ID manifests were exported. We checked duplicate IDs and normalized-text overlaps before running the experiments.

For seed $s \in \{17, 29, 43\}$, we derived draws from SHA-256 of the UTF-8 string `task|split|query-id|s|draw`, taking the first eight bytes as a little-endian PCG64 seed.

We sampled eight distinct pool IDs; their prefixes defined the 2/4/8-shot prompts. Training used draw 0, while validation and evaluation used draws 0–3. The query itself was excluded from its demonstrations. The prompt was:

```
Choose the correct option.  Reply with its letter.
{demonstrations, each ending with Answer:  X}
Context:  {passage, if present}
Question:  {question}
A. {option} ...
Answer:
```

Examples were separated by blank lines. We verified that the leading-space letter verbalizers were single-token continuations without changing prefix tokenization, scored only those logits, and normalized over each item's actual choices. We applied a 16,384-token limit with deterministic redraws when necessary and logged overflow handling. ARC-C inference retained each item's native choice count, and all reported probability vectors were computed from the actual evaluation item manifests.

We used BF16 frozen backbones, FP32 losses, FP64 covariance factorization, no weight quantization, and downstream autograd. Training used microbatch 1 with gradient accumulation to an effective batch size of 32. We hooked the final unpadded prompt token after block 14/16 for Qwen/Llama.

We used standardized-input width-64 GELU modules with Xavier-uniform W_1, $b_1 = 0$, and $W_2 = b_2 = 0$. AdamW used 10^{-4}, betas $(.9, .999)$, zero weight decay, clipping 1.0, and five epochs per stage. Reference acquisition was selected by validation answer cross-entropy; the selected reference targets were then frozen and cached for the exact compression prompts. We fitted $G = (\widehat{C}_z + 10^{-3}s^2 I)^{-1}$, $s^2 = \mathrm{tr}(\widehat{C}_z)/d$, using training baseline states only and computed costs by Cholesky solves. We swept $\lambda = \alpha A_{\mathrm{ref}}/C_{\mathrm{ref}}$ with $\alpha \in \{0, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$, $A_{\mathrm{ref}} = \mathbb{E}\|p_u - p_0\|_2^2$, and $C_{\mathrm{ref}} = \mathbb{E}\|u\|_G^2$ on training prompts.

J.2 Comparators, selection, and full fidelity

Global scaling searched $\alpha \in \{0, .01, \ldots, 1\}$ with refinement to .001 near a boundary. Adaptive and coordinate gates used width-64 GELU networks with one or d sigmoid outputs, respectively, initialized by zero output weights and $\log 99$ biases. Learned controls shared the reference target, optimizer, epoch budget, and penalty grid, with the unchanged reference as fallback.

We selected operating points separately per seed using validation data only and then froze them across all evaluation shot counts. Each validation draw was required to pass mean TV .01, P95 TV .05, and disagreement 1%; cost ties were broken by TV, smaller penalty, then earlier epoch. The selected disagreement range is 0.60–0.92%, and every next-cheaper candidate fails disagreement (minimum 1.08%). Validation records contain candidate IDs, all three constraints, the nearest threshold, and the next rejected point. The test sweep used the exact selected test records summarized in Table 1.

We used the empirical higher-order statistic for quantiles. Draws were averaged within each seed and then seeds were weighted equally; SD across the three seeds used the sample standard deviation ($n-1=2$). Accuracy and disagreement were computed from integer counts. Cross-shot summaries were first averaged within each shot and then weighted equally across 0/2/4/8 shots. Maxima and pass counts used unrounded values. Shared means were rendered consistently to three cost decimals, four TV decimals, and two percentage decimals.

J.3 ICL aggregation and contextual capacity

Four four-shot draws yielded six unordered context pairs per query. We computed S_{ctx} and response error from the full answer distributions by averaging pairs, queries, and then seeds. The reference remained fixed while its target was recomputed for each prompt. We checked the bound $E_{\mathrm{resp}} \le \mathbb{E}[\mathrm{TV}(p_v^a, p_u^a) + \mathrm{TV}(p_v^b, p_u^b)]$ on the same pairs. Table J.3 separates typical fidelity from the worst individual cell. The two failing seed–shot conditions were Qwen/LogiQA seed 43 at eight shots and Llama/LogiQA seed 29 at eight shots; both exceeded mean-TV and disagreement thresholds and were retained in all summaries.

Table J.2: **Complete four-shot fidelity results.** Accuracy and disagreement are percentages. A seed passes only if all four prompt draws pass every budget. Results correspond to Table 1.

Task	Method	Acc.	Cost	Mean TV	P95 TV	Disagree.	Seeds passed
Qwen2.5-7B							
ARC-C	Global	68.49	0.874	0.0078	0.0214	0.80	3/3
ARC-C	Adaptive	68.54	0.721	0.0070	0.0190	0.72	3/3
ARC-C	Coordinate	68.65	0.558	0.0064	0.0172	0.77	3/3
ARC-C	MFM	68.54	0.367	0.0057	0.0153	0.60	3/3
CSQA	Global	75.75	0.906	0.0083	0.0225	0.74	3/3
CSQA	Adaptive	75.68	0.766	0.0076	0.0202	0.80	3/3
CSQA	Coordinate	75.79	0.613	0.0070	0.0188	0.66	3/3
CSQA	MFM	75.79	0.421	0.0061	0.0162	0.57	3/3
LogiQA	Global	44.74	0.928	0.0087	0.0241	0.86	3/3
LogiQA	Adaptive	44.73	0.807	0.0081	0.0215	0.77	3/3
LogiQA	Coordinate	44.78	0.641	0.0077	0.0211	0.61	3/3
LogiQA	MFM	44.70	0.463	0.0065	0.0176	0.61	3/3
Llama-3.1-8B							
ARC-C	Global	66.01	0.891	0.0080	0.0213	0.77	3/3
ARC-C	Adaptive	65.98	0.747	0.0073	0.0200	0.82	3/3
ARC-C	Coordinate	66.11	0.587	0.0068	0.0183	0.68	3/3
ARC-C	MFM	66.11	0.398	0.0062	0.0170	0.68	3/3
CSQA	Global	74.34	0.883	0.0079	0.0216	0.84	3/3
CSQA	Adaptive	74.56	0.719	0.0070	0.0191	0.72	3/3
CSQA	Coordinate	74.53	0.566	0.0060	0.0163	0.66	3/3
CSQA	MFM	74.52	0.386	0.0055	0.0148	0.59	3/3
LogiQA	Global	42.72	0.946	0.0085	0.0229	0.77	3/3
LogiQA	Adaptive	42.68	0.834	0.0081	0.0221	0.83	3/3
LogiQA	Coordinate	42.61	0.689	0.0071	0.0188	0.77	3/3
LogiQA	MFM	42.60	0.509	0.0067	0.0183	0.68	3/3

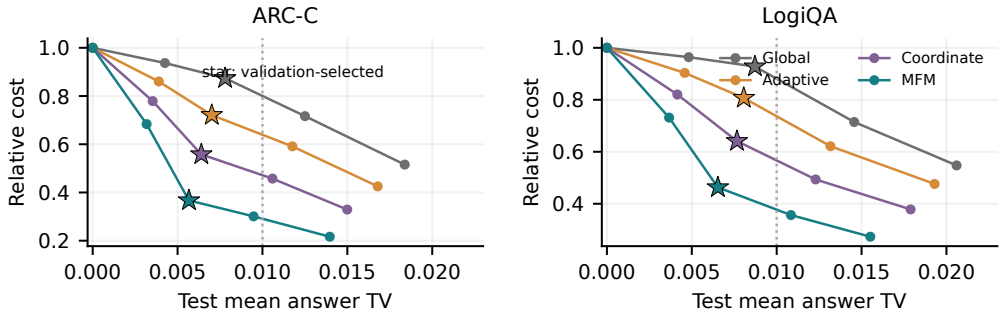

Figure J.1: **Discrete candidate results for Qwen2.5-7B.** Each marker averages three seeds and four test draws; stars are validation-selected points, as reported in Tables 1 and J.2. Lines connect discrete candidates; selection uses all three validation constraints.

For the capacity diagnostics, we used 512 fitting, 128 tuning, and 256 evaluation queries from the corresponding disjoint splits. We compared a common demonstration quartet with query-specific quartets while holding query IDs, reference, G, and total prompt counts fixed. With $CC^\top = \widehat{C}_z + 10^{-3}s^2 I$, we used $w = C^{-1}v$ and $B = JC$, solved ridge systems in observation space, and centered the oracle field before SVD. We used the same penalty $0.1A_{\mathrm{ref}}/C_{\mathrm{ref}}$ for widths 16/64/128 while keeping one reference fixed across widths. Width 64 copied its initialization, width 16 retained the largest 16 output-column norms, and width 128 added zero-output units with Xavier input weights. We selected by the validation objective and evaluated tax on held-out prompts; unstable ratios with very small Γ were excluded according to the same rule across settings.

We computed spectral tails by summing empirical eigenvalues and report $\lambda S_r/2$, T, and $\Gamma - T$. Reference tails vanish at rank 64. Figure 3(c) includes the pre-fit bound to distinguish an architecture-level obstruction from optimization failure. These are empirical local diagnostics. Nonlinear transfer additionally requires $\Gamma - T - K(v) - K_{\mathcal{S}} > 0$.

Table J.3: **Cross-context fidelity checks.** Maxima cover all seed–shot–draw cells. A seed–shot passes only when all its draws pass. Response error uses four-shot pairs.

Task	Max mean TV	Max P95 TV	Max dis.	Seed–shots passed	Response error
Qwen2.5-7B					
ARC-C	0.0067	0.0187	0.68	12/12	0.0087
CSQA	0.0072	0.0209	0.66	12/12	0.0094
LogiQA	0.0135	0.0370	1.23	11/12	0.0100
Llama-3.1-8B					
ARC-C	0.0074	0.0206	0.77	12/12	0.0095
CSQA	0.0063	0.0173	0.74	12/12	0.0085
LogiQA	0.0144	0.0418	1.38	11/12	0.0103

J.4 Reference, metric, and observation controls

Table J.4 reports Qwen/LogiQA controls averaged over the same three seeds and four four-shot draws. The Standard/Shrinkage row reuses the main experimental records exactly. Early stopping selected a checkpoint within one validation-accuracy percentage point of the acquired reference. Norm regularization added $\eta\|u\|_G^2/2$ with $\eta \in \{10^{-5}, 10^{-4}, 10^{-3}\}$ under the same utility tolerance. We refit compression for each reference. The Euclidean and diagonal controls retained their respective references while refitting compression and recalibrating penalties, with $G = I$ or $G = \mathrm{diag}(\widehat{C}_z + 10^{-3}s^2 I)^{-1}$. Each row is normalized to its own reference, so costs across geometries are not absolute comparisons. MFM retains a cost advantage under all three geometries and both alternative reference-acquisition procedures, with the magnitude varying across settings.

Table J.4: **Reference and metric controls.** Same seed/draw aggregation as the main four-shot tables. Accuracy and disagreement are percentages; each reference has unit cost in its own geometry.

Reference	Metric	Ref. Acc.	Adaptive	MFM	Mean TV	Disagree.
Standard	Shrinkage	44.91	0.807	0.463	0.0065	0.61
Standard	Euclidean	44.91	0.787	0.493	0.0073	0.61
Standard	Diagonal	44.91	0.825	0.480	0.0061	0.77
Early stopped	Shrinkage	44.14	0.784	0.482	0.0071	0.46
Early stopped	Euclidean	44.14	0.814	0.516	0.0059	0.61
Norm regularized	Shrinkage	44.44	0.843	0.553	0.0057	0.61
Norm regularized	Euclidean	44.44	0.820	0.544	0.0068	0.61

Strengthening the observation on language-model tasks. We ran a four-shot experiment on Qwen2.5-7B and Llama-3.1-8B with ARC-C and LogiQA, using the same reference, hook, width 64, G, three seeds, four demonstration draws, and evaluation queries as Table 1. Answer-only rows reuse the four-shot results; answer-set mass is computed from full-vocabulary logits on the same prompts.

For full-vocabulary logits ℓ, we computed $F_k = \exp \ell_{a_k} / \sum_{j\in\mathcal{A}} \exp \ell_j$ and $R = \sum_{j\in\mathcal{A}} \exp \ell_j / \sum_{v\in\mathcal{V}} \exp \ell_v$. We used the same single-token verbalizers and score position in both cases and implemented R with the difference of the two log-sum-exp normalizers. Here R retains the probability mass removed by answer-set renormalization. Drift was $|R_v - R_u|$, in probability units, with the reference recomputed for each prompt. Together (F, R) fixes the individual answer-token probabilities RF_k, but not the distribution among non-answer tokens or generated reasoning traces.

The strengthened objective was

$$\mathcal{L}_{\lambda,\eta}(v) = \mathbb{E}\big[\tfrac{1}{2}\|F_v - F_u\|_2^2 + \tfrac{\eta}{2}(R_v - R_u)^2 + \tfrac{\lambda}{2}\|v\|_G^2\big] .$$

We kept the original F coordinates and discrepancy unchanged. We reused the optimizer and five-epoch budget, crossed the existing λ grid with $\eta \in \{0, 1, 10, 100\}$, and initialized each fit from the fixed reference. We pooled all checkpoints with the original answer-only solution and the reference

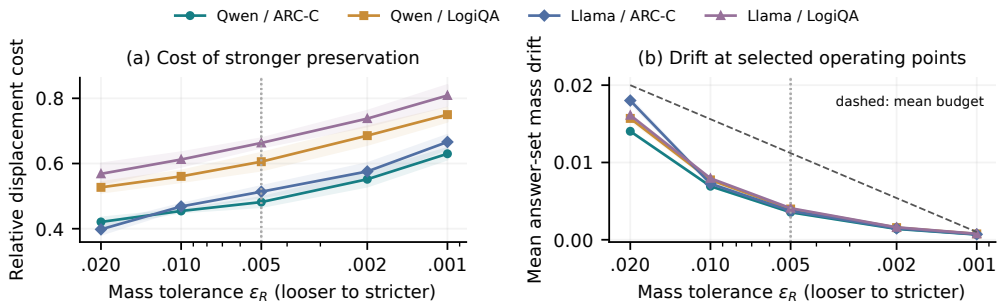

Figure J.2: **Observation-strengthening results.** Validation-selected candidates evaluated on fixed four-shot test sets. Tightening mass tolerance reduces measured drift and increases measured cost; bands are seed SD. The dotted vertical line marks Table J.5's tolerance. The dashed line in (b) is the mean-drift budget; all individual P95 budgets are checked separately.

Table J.5: **Enhanced-observation results.** Default $\epsilon_R = .005$, P95 budget .025. Cost uses the same reference denominator as Table 1; $\pm$ is seed SD. TV and mass drift are probabilities; disagreement is percent. Pass counts require all five constraints on all four test draws within each seed.

Task	Observation	C_{rel}	Mean TV	Dis.	Mean $\lvert\Delta R\rvert$	P95 $\lvert\Delta R\rvert$	Pass
Qwen2.5-7B							
ARC-C	Answer-only	$0.367_{\pm 0.013}$	0.0057	0.60	0.0211	0.0559	0/3
ARC-C	Strengthened	$0.482_{\pm 0.015}$	0.0048	0.57	0.0036	0.0094	3/3
LogiQA	Answer-only	$0.463_{\pm 0.021}$	0.0065	0.61	0.0322	0.0839	0/3
LogiQA	Strengthened	$0.606_{\pm 0.027}$	0.0051	0.55	0.0040	0.0104	3/3
Llama-3.1-8B							
ARC-C	Answer-only	$0.398_{\pm 0.016}$	0.0062	0.68	0.0180	0.0471	0/3
ARC-C	Strengthened	$0.513_{\pm 0.017}$	0.0054	0.65	0.0037	0.0097	3/3
LogiQA	Answer-only	$0.509_{\pm 0.023}$	0.0067	0.68	0.0388	0.1014	0/3
LogiQA	Strengthened	$0.663_{\pm 0.014}$	0.0049	0.60	0.0041	0.0107	3/3

fallback. For each seed and tolerance $\epsilon_R \in \{.020, .010, .005, .002, .001\}$, we selected the minimum validation cost among candidates satisfying, on *every* validation draw, the original answer budgets (mean TV .01, P95 TV .05, disagreement 1%), plus mean mass drift $\leq \epsilon_R$ and P95 mass drift $\leq 5\epsilon_R$. We froze the selected candidate before testing. The same evaluated candidate pool was used for every tolerance, giving nested validation feasible sets. Held-out costs are evaluated separately at each selected point.

All rows in this experiment satisfy the original answer constraints, including answer P95 TV; the maximum individual strengthened P95 TV is 0.0161. Answer-only rows pass the original constraints but fail the added mass budget. Metrics average draws within seed, then seeds equally. P95 values are averages of per-draw quantiles; pass decisions use the individual unrounded quantiles. Each draw reuses the same 1,172/651 evaluation queries.

At the default tolerance, macro-mean mass drift changes from 0.0275 to 0.0038, while cost changes from 0.434 to 0.566, retaining 43.4% savings relative to the same reference. These are finite-tolerance costs of a trained shared nonlinear module; Theorem 3.3 gives the exact pointwise linear optimum. Appendix F illustrates their relation.