ANCHOR RECONSTRUCTION FOR EFFICIENT MODEL-TASK EVALUATION: IDENTIFIABILITY AND STABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

As the numbers of models and tasks grow rapidly, dense evaluation becomes increasingly costly since each new model or task may require many additional evaluations. Anchor reconstruction reduces this burden by using approximately low-rank model-task structure to estimate the full performance vector of a new task or model from evaluations on a few selected models or tasks, called anchors. This setting raises three questions: what determines reconstruction across nonunique factorizations, when do sparse anchor scores uniquely determine the full vector, and how do errors arise, propagate, and become amplified? For a fixed rank-K approximation, we show that the performance subspace rather than a particular basis is the intrinsic object. Anchor reconstruction is identifiable exactly when coordinate restriction is injective on this subspace, while stability is governed by the anchor minimum gain. We derive an end-to-end reconstruction error bound separating target-subspace mismatch, historical-matrix perturbation, and anchor-evaluation noise, with inverse gain amplifying all three. This explains why approximate low rank and enough anchors alone do not ensure stable reconstruction. Experiments on a 132×97 matrix and a $4{,}576 \times 6$ public leaderboard illustrate the rank-budget boundary for identifiability and support the predicted roles of error pathways and inverse-gain amplification in reconstruction stability.

1 INTRODUCTION

Evaluating growing collections of large language models (LLMs) and tasks is increasingly costly, as each new model or task may require many additional model-task evaluations. However, model-task performance matrices are often approximately low rank and therefore exhibit much lower-dimensional structure (Burnell et al., 2023; Schram et al., 2023; Zhang et al., 2024; Zhao et al., 2025). This structure has motivated partial-evaluation methods that reconstruct a full performance vector using evaluations on only a few selected models or tasks, which we refer to collectively as *anchors* (Subramani et al., 2025; Maimon et al., 2026; Zeng & Papailiopoulos, 2026).

Among these approaches, we study *anchor reconstruction*, a selected-coordinate least-squares formulation on a low-dimensional subspace. Classical coordinate and sensor-selection methods study this primitive given a specified or estimated subspace (Everson & Sirovich, 1995; Chaturantabut & Sorensen, 2010; Drmač & Gugercin, 2016; Manohar et al., 2018). In model-task evaluation, this primitive is part of a longer pipeline: a historical performance matrix is used to estimate the performance structure and select anchors, after which only anchor scores for a new task or model are observed and used to reconstruct its full performance vector.

However, nonunique low-rank factorizations and sparse anchor observations raise three questions: (1) what is intrinsic to reconstruction across equivalent factorizations; (2) when do anchor scores uniquely determine a vector in the corresponding subspace; and (3) how do errors arise, propagate, and become amplified, thereby affecting reconstruction stability?

Different factorizations of a fixed rank-K approximation are related by invertible changes of basis from the general linear group. Quotienting out this basis ambiguity identifies their shared performance subspace. We show that, for a given identifiable anchor set, all bases of this subspace yield identical reconstructions, making the subspace the intrinsic object. Anchor reconstruction is selected-coordinate recovery on this subspace and is identifiable exactly when coordinate restriction is injective, equivalently when any basis matrix restricted to the anchors has full column rank.

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.

Identifiability guarantees unique recovery on the subspace but does not ensure stability. We quantify this sensitivity through the anchor minimum gain, which measures the least norm retained by the selected coordinates over unit subspace vectors. We derive an end-to-end error bound separating target-subspace mismatch, historical-matrix perturbation, and anchor-evaluation noise, with inverse gain amplifying all three. Historical perturbation can also change the gain itself as it displaces the estimated subspace. These error pathways and their amplification explain why approximate low rank and enough anchors alone do not ensure stable reconstruction.

Empirically, we evaluate anchor reconstruction on a 132×97 model-task matrix and a $4{,}576 \times 6$ Open LLM Leaderboard matrix (Fourrier et al., 2024). We study reconstruction across rank, anchor budget, and selection rules, examine the three error pathways in the bound, and test inverse gain through observational and controlled experiments. Our contributions are:

- We identify the performance subspace as the intrinsic object of a fixed low-rank approximation and characterize basis-independent reconstruction and anchor identifiability.
- We analyze reconstruction stability through an end-to-end bound separating three error pathways and their common worst-case amplification by inverse anchor gain, showing why approximate low rank and enough anchors alone do not ensure stable reconstruction.
- We empirically examine rank-budget behavior, error pathways, and inverse-gain amplification on a 132×97 model-task matrix and an independent $4{,}576 \times 6$ public leaderboard.

2 Related Work

Efficient model-task evaluation. Historical performance matrices have supported score prediction and analysis of shared structure across AutoML and LLM evaluation (Fusi et al., 2018; Schram et al., 2023; Burnell et al., 2023; Zhao et al., 2025). Zhang et al. (2024) also combine historical scores with model and task descriptors to estimate unobserved performance. OBOE learns from historical algorithm-dataset scores, evaluates selected algorithms on a new dataset, and predicts the remaining scores by latent least squares; its cost-aware selection uses a log-determinant objective (Yang et al., 2019). For LLM benchmarks, SimBA finds representative task subsets and uses their scores to predict held-out model performance (Subramani et al., 2025). Maimon et al. (2026) use a factor-analytic skill space to examine benchmark redundancy and model capabilities, reconstruct new-model performance from selected tasks, and predict new-task performance from selected models. Karuturi et al. (2026) establish rank and conditioning requirements, derive taskwise certificates under factor-estimation error, and select a Minimum Sufficient Evaluation Set using a regularized log-determinant criterion. Nonlinear alternatives include logit-space score completion (Zeng & Papailiopoulos, 2026) and, at the example level, item-response modeling (Polo et al., 2024). We focus on linear selected-coordinate reconstruction.

Selected-coordinate recovery and stability. For a fixed low-dimensional subspace, gappy POD fits latent coefficients by least squares from possibly redundant coordinate observations, whereas DEIM and Q-DEIM study interpolatory coordinate selection (Everson & Sirovich, 1995; Chaturantabut & Sorensen, 2010; Drmač & Gugercin, 2016). Data-driven sensor placement likewise uses subspace structure to choose informative measurements (Manohar et al., 2018). Analyses of noisy and oversampled gappy POD show that additional observations can improve noise stability, while sampled-basis conditioning still governs reconstruction error (Peherstorfer et al., 2020). Optimal-recovery bounds relate approximation error to an inverse inf-sup stability constant, and measurement-selection methods seek to improve that constant (Binev et al., 2017; 2018). For coordinate measurements, this constant corresponds to the minimum gain of the restricted subspace, so pseudoinverse reconstruction and inverse-gain amplification are classical ingredients rather than separate methodological contributions. Singular-subspace perturbation theory bounds how noise in a matrix displaces its estimated subspace (Wedin, 1972). Classical sensor design compares E-optimality for the weakest observed direction with D-optimality for retained volume, while Q-DEIM uses pivoted QR (Nakai et al., 2021; Drmač & Gugercin, 2016). Our analysis tracks how historical-matrix perturbation changes both the estimated subspace and anchor gain, while target mismatch and anchor-evaluation noise enter a common reconstruction bound through distinct pathways.

3 PRELIMINARIES AND PROBLEM FORMULATION

We adopt the term *anchor reconstruction* for a common selected-coordinate formulation that infers a full performance vector from evaluations of a small set of models or tasks (Yang et al., 2019; Maimon et al., 2026; Karuturi et al., 2026). We analyze this formulation in the task-side setting, where a small set of anchor models is used to reconstruct the performance of all models on a new task; the model-side counterpart follows by transposing the performance matrix.

3.1 THE APPROXIMATELY LOW-RANK MODEL-TASK PERFORMANCE MATRIX

Let $\{f_i\}_{i=1}^N$ and $\{t_j\}_{j=1}^M$ denote the model and task collections, respectively. Evaluating all model-task pairs yields the performance matrix $P \in [0,1]^{N\times M}$, where $p_{ij} \in [0,1]$ is the performance score of model f_i on task t_j, and the j-th column $p_{:j} := P_{:j} \in [0,1]^N$ is the performance vector of task t_j. Let $\mathbb{R}_*^{a\times b}$ denote the set of full-column-rank matrices in $\mathbb{R}^{a\times b}$. For a prescribed rank $K \ll \min(N, M)$ and tolerance $0 \leqslant \varepsilon_{\mathrm{lr}} \ll 1$, we use the approximate rank-K representation

$$P = CW + R_{\mathrm{lr}}, \qquad \frac{\|R_{\mathrm{lr}}\|_F}{\|P\|_F} \leqslant \varepsilon_{\mathrm{lr}}, \tag{1}$$

where $C \in \mathbb{R}_*^{N\times K}$, $W^{\mathrm{T}} \in \mathbb{R}_*^{M\times K}$, and $R_{\mathrm{lr}} \in \mathbb{R}^{N\times M}$. The j-th column of equation 1 is $p_{:j} = Cw_{:j} + (R_{\mathrm{lr}})_{:j}$, where $w_{:j} \in \mathbb{R}^K$ is the coefficient vector in the basis C.

3.2 FORMULATION OF ANCHOR RECONSTRUCTION

Consider a new task t_u with unknown target performance vector $p_u \in [0,1]^N$. Using the basis C in equation 1, the formulation approximates the target as $p_u \approx Cw_u$, where $w_u \in \mathbb{R}^K$ is the unknown coefficient vector. Let $\mathcal{A} \subseteq [N]$ be a given anchor-model index set, with $|\mathcal{A}| = L$ and $L \ll N$, and let $\mathcal{R}_\mathcal{A}$ be the anchor-coordinate restriction operator that retains the entries indexed by $\mathcal{A}$,

$$\mathcal{R}_\mathcal{A} : \mathbb{R}^N \to \mathbb{R}^L.$$

With noiseless anchor scores $p_{\mathcal{A}u} := \mathcal{R}_\mathcal{A} p_u \in [0,1]^L$, restricting the target approximation gives

$$p_{\mathcal{A}u} \approx C_\mathcal{A} w_u, \qquad C_\mathcal{A} := \mathcal{R}_\mathcal{A} C \in \mathbb{R}^{L\times K}.$$

If $C_\mathcal{A}$ has full column rank, the least-squares coefficient estimate is unique:

$$\widehat{w}_u = \operatorname*{arg\,min}_{w\in\mathbb{R}^K} \|C_\mathcal{A} w - p_{\mathcal{A}u}\|_2^2 = C_\mathcal{A}^\dagger p_{\mathcal{A}u}, \tag{2}$$

where $C_\mathcal{A}^\dagger = \left(C_\mathcal{A}^{\mathrm{T}} C_\mathcal{A}\right)^{-1} C_\mathcal{A}^{\mathrm{T}}$ is the Moore-Penrose pseudoinverse of $C_\mathcal{A}$. The corresponding anchor reconstruction operator and the resulting reconstruction are therefore

$$D_{C\mathcal{A}} := CC_\mathcal{A}^\dagger \in \mathbb{R}^{N\times L}, \qquad \widehat{p}_u = D_{C\mathcal{A}} p_{\mathcal{A}u}. \tag{3}$$

The formulation above is the reconstruction primitive studied in this paper. We analyze subspace geometry and identifiability in section 4, then stability and error propagation in section 5.

4 SUBSPACE GEOMETRY AND IDENTIFIABILITY

The basis C in section 3 can vary across factorizations. Building on classical selected-coordinate recovery (Everson & Sirovich, 1995; Manohar et al., 2018), we identify the performance subspace as intrinsic, characterize identifiability, and interpret anchor scores as coordinates.

4.1 THE PERFORMANCE SUBSPACE AS AN INTRINSIC OBJECT

The rank-K component CW in equation 1 has column space $\mathcal{S} = \mathrm{col}(CW) = \mathrm{col}(C) \in \mathrm{Gr}(K, N)$, where $\mathrm{Gr}(K, N)$ is the Grassmannian of K-dimensional subspaces of $\mathbb{R}^N$. For a fixed rank-K approximation, changing the latent basis does not alter the represented structure. Let $GL(K)$ be the general linear group of invertible matrices in $\mathbb{R}^{K\times K}$, acting on the factors as

$$(C, W) \cdot G := (CG, G^{-1}W), \qquad G \in GL(K). \tag{4}$$

The action preserves CW and $\mathrm{col}(C)$, and its induced action $C \mapsto CG$ partitions $\mathbb{R}_*^{N\times K}$ into orbits

$$\mathcal{O}(C) := \{CG : G \in GL(K)\}.$$

Each orbit groups basis matrices with the same column space, giving the canonical bijection

$$\mathbb{R}_*^{N\times K}/GL(K) \simeq \mathrm{Gr}(K, N), \qquad \mathcal{O}(C) \mapsto \mathrm{col}(C).$$

Under the full-column-rank condition on $\mathcal{R}_\mathcal{A}C$ used in section 3.2, $D_{C\mathcal{A}}$ is constant on each orbit:

$$D_{(CG)\mathcal{A}} = (CG)(\mathcal{R}_\mathcal{A}CG)^\dagger = D_{C\mathcal{A}}, \qquad G \in GL(K). \tag{5}$$

Q-DEIM establishes this invariance for orthonormal bases and $L = K$ (Drmač & Gugercin, 2016), while model-task factorizations motivate its full $GL(K)$ form for arbitrary bases and $L \geqslant K$. The orbit invariance therefore makes the following operator well defined.

Definition 1 (Basis-Independent Anchor Reconstruction Operator)**.** *Let $\mathcal{S} \in \mathrm{Gr}(K, N)$ and $\mathcal{A} \subseteq [N]$. If $\mathcal{R}_\mathcal{A}C$ has full column rank for a basis matrix C of $\mathcal{S}$, define*

$$D_{\mathcal{S}\mathcal{A}} := C(\mathcal{R}_\mathcal{A}C)^\dagger.$$

For a fixed rank-K approximation, the represented performance structure and anchor reconstruction depend on C only through $\mathcal{S} = \mathrm{col}(C)$. Thus, $\mathcal{S}$, rather than its basis, is intrinsic to that approximation. To choose one from data, we use a best-fitting K-dimensional subspace:

$$\widehat{\mathcal{S}} \in \operatorname*{arg\,min}_{\mathcal{S}\in\mathrm{Gr}(K,N)} \|P - \Pi_\mathcal{S}P\|_F^2, \qquad \text{with} \quad \widehat{\mathcal{S}} = \mathrm{col}(U_K), \tag{6}$$

where $\Pi_\mathcal{S}$ projects orthogonally onto $\mathcal{S}$ and $U_K \in \mathbb{R}^{N\times K}$ contains the top-K left singular vectors of P. When $\sigma_K(P) > \sigma_{K+1}(P)$, $\widehat{\mathcal{S}}$ is uniquely determined by P, while U_K is only a basis for it.

4.2 Identifiability on the Performance Subspace

For an anchor set $\mathcal{A} \subseteq [N]$ with $|\mathcal{A}| = L$, consider the restriction of $\mathcal{R}_\mathcal{A}$ to $\mathcal{S}$:

$$\mathcal{R}_\mathcal{A}|_\mathcal{S} : \mathcal{S} \to \mathbb{R}^L. \tag{7}$$

We call $\mathcal{A}$ *identifiable* for $\mathcal{S}$ when anchor scores uniquely determine every vector in $\mathcal{S}$, equivalently when $\mathcal{R}_\mathcal{A}|_\mathcal{S}$ is injective. The following standard equivalences identify this intrinsic condition with the full-column-rank requirement used above.

Proposition 1 (Equivalent Conditions for Anchor Identifiability)**.** *Let $C \in \mathbb{R}_*^{N\times K}$ be any basis matrix of $\mathcal{S} \in \mathrm{Gr}(K, N)$, and let $C_\mathcal{A} := \mathcal{R}_\mathcal{A}C$. Then the following conditions are equivalent:*

$$\mathcal{R}_\mathcal{A}|_\mathcal{S} \text{ is injective} \iff \mathcal{S} \cap \ker(\mathcal{R}_\mathcal{A}) = \{0\} \iff C_\mathcal{A} \text{ has full column rank}. \tag{8}$$

Since $C_\mathcal{A} \in \mathbb{R}^{L\times K}$, full column rank requires $L \geqslant K$. Although this dimension condition does not make an arbitrary anchor set identifiable, it guarantees that every $\mathcal{S} \in \mathrm{Gr}(K, N)$ admits at least one identifiable anchor set of size L, as shown by the existence proof in appendix A.2.

4.3 Anchor Scores as Subspace Coordinates

We can now interpret anchor reconstruction independently of a particular latent basis. For an identifiable anchor set, the restricted map $\mathcal{R}_\mathcal{A}|_\mathcal{S}$ is a linear isomorphism from $\mathcal{S}$ onto its image $\mathcal{R}_\mathcal{A}(\mathcal{S})$, and the intrinsic reconstruction operator is its inverse on this feasible anchor-score space:

$$D_{\mathcal{S}\mathcal{A}}|_{\mathcal{R}_\mathcal{A}(\mathcal{S})} = (\mathcal{R}_\mathcal{A}|_\mathcal{S})^{-1}, \quad D_{\mathcal{S}\mathcal{A}}\mathcal{R}_\mathcal{A}p = p, \quad \forall p \in \mathcal{S}. \tag{9}$$

For $L = K$, identifiability implies $\mathcal{R}_\mathcal{A}(\mathcal{S}) = \mathbb{R}^K$, so the anchor scores serve as coordinates for $\mathcal{S}$, whereas for $L > K$, $\mathcal{R}_\mathcal{A}(\mathcal{S})$ is a K-dimensional subspace of $\mathbb{R}^L$ and the scores provide redundant coordinates. If $L < K$, no linear map from the K-dimensional space $\mathcal{S}$ to $\mathbb{R}^L$ can be injective. Thus, anchors work by preserving every degree of freedom in $\mathcal{S}$. Section 5 next analyzes stability.

5 End-to-End Reconstruction Error and Stability

Building on classical results for subspace approximation, subspace estimation from noisy data, and reconstruction from noisy coordinates (Drmač & Gugercin, 2016; Binev et al., 2017; Peherstorfer et al., 2020; Wedin, 1972), we derive an end-to-end error bound for anchor reconstruction.

5.1 From Identifiability to Stable Reconstruction

Identifiability ensures unique recovery on $\mathcal{S}$ but does not quantify inverse stability. We measure it by the least norm retained by the selected coordinates over unit vectors in $\mathcal{S}$.

Definition 2 (Anchor Minimum Gain). *For $\mathcal{S} \in \mathrm{Gr}(K, N)$ and $\mathcal{A} \subseteq [N]$, define*

$$\alpha_{\mathcal{A}}(\mathcal{S}) := \inf_{s \in \mathcal{S}, \|s\|_2 = 1} \|\mathcal{R}_{\mathcal{A}} s\|_2 .$$

Geometrically, $\alpha_{\mathcal{A}}(\mathcal{S})$ measures the worst-case contraction induced by $\mathcal{R}_{\mathcal{A}}$ across directions in $\mathcal{S}$. The following proposition connects the gain to identifiability and reconstruction sensitivity.

Proposition 2 (Minimum-Gain and Identifiability). *For $|\mathcal{A}| = L \geqslant K$ and any orthonormal basis $U \in \mathbb{R}^{N \times K}$ of $\mathcal{S}$, the minimum gain has the spectral and identifiability characterizations*

$$\alpha_{\mathcal{A}}(\mathcal{S}) = \sigma_K(\mathcal{R}_{\mathcal{A}} U), \qquad \mathcal{A} \text{ is identifiable for } \mathcal{S} \iff \alpha_{\mathcal{A}}(\mathcal{S}) > 0.$$

For an identifiable $\mathcal{A}$, the reconstruction operator also satisfies

$$\|D_{\mathcal{S}\mathcal{A}}\|_2 = \alpha_{\mathcal{A}}(\mathcal{S})^{-1}. \tag{10}$$

Therefore, a small positive gain gives identifiable but sensitive reconstruction, whereas a larger gain improves stability. Proposition 4 further shows that no exact linear reconstruction operator has smaller worst-case amplification than $D_{\mathcal{S}\mathcal{A}}$.

5.2 Reconstruction Error Decomposition and Bound

Let $P^* \in [0,1]^{N \times M}$ and $p_u^* \in [0,1]^N$ be the clean performance matrix and clean target performance vector, respectively. Let $\mathcal{E}_P \in \mathbb{R}^{N \times M}$ and $\epsilon_{\mathcal{A}} \in \mathbb{R}^L$ be historical-matrix perturbation and anchor-evaluation noise, respectively. With the coordinate-restriction operator $\mathcal{R}_{\mathcal{A}}$, we observe

$$P = P^* + \mathcal{E}_P, \qquad p_{\mathcal{A}u} = \mathcal{R}_{\mathcal{A}} p_u^* + \epsilon_{\mathcal{A}}.$$

Let $\widehat{\mathcal{S}}$ and $\mathcal{S}^*$ be the top-K left singular subspaces of P and P^*, with orthogonal projectors $\widehat{\Pi}$ and Π^*, respectively. We define the mismatch of p_u^* relative to $\widehat{\mathcal{S}}$ as $\widehat{r}_u = \left\|(I - \widehat{\Pi}) p_u^*\right\|_2$, which satisfies

$$\widehat{r}_u \leqslant \underbrace{\left\|(I - \Pi^*) p_u^*\right\|_2}_{r_u^*} + \underbrace{\left\|\widehat{\Pi} - \Pi^*\right\|_2}_{d_{\mathrm{Gr}}(\widehat{\mathcal{S}}, \mathcal{S}^*)} \left\|p_u^*\right\|_2, \tag{11}$$

where r_u^* is the mismatch of p_u^* relative to $\mathcal{S}^*$ and $d_{\mathrm{Gr}}(\widehat{\mathcal{S}}, \mathcal{S}^*)$ measures the displacement of $\widehat{\mathcal{S}}$ from $\mathcal{S}^*$ on the Grassmannian. Let $Q := D_{\widehat{\mathcal{S}}\mathcal{A}} \mathcal{R}_{\mathcal{A}}$. Exactness on $\widehat{\mathcal{S}}$ gives $Q\widehat{\Pi} = \widehat{\Pi}$ and the error identity

$$\widehat{p}_u - p_u^* = (Q - I)(I - \widehat{\Pi}) p_u^* + D_{\widehat{\mathcal{S}}\mathcal{A}} \epsilon_{\mathcal{A}}. \tag{12}$$

The first term is the reconstruction error induced by the component of p_u^* outside $\widehat{\mathcal{S}}$, while the second propagates anchor-evaluation noise through the reconstruction operator.

Theorem 1 (Unified Deterministic Reconstruction Bound). *Let $\mathcal{A}$ be identifiable for $\widehat{\mathcal{S}}$ and let $\widehat{p}_u = D_{\widehat{\mathcal{S}}\mathcal{A}} p_{\mathcal{A}u}$ be the unclipped reconstruction. Then*

$$\|\widehat{p}_u - p_u^*\|_2 \leqslant \frac{\widehat{r}_u + \|\epsilon_{\mathcal{A}}\|_2}{\alpha_{\mathcal{A}}(\widehat{\mathcal{S}})} \leqslant \frac{r_u^* + d_{\mathrm{Gr}}(\widehat{\mathcal{S}}, \mathcal{S}^*) \|p_u^*\|_2 + \|\epsilon_{\mathcal{A}}\|_2}{\alpha_{\mathcal{A}}(\widehat{\mathcal{S}})}.$$

We keep r_u^* explicit; appendix B.3 relates it to historical coverage. For anchor-evaluation noise, conditional on P, suppose each anchor score is the average of n_{iu} independent $[0,1]$-valued samples with mean p_{iu}^*. For any $\delta \in (0,1)$, Hoeffding's inequality and a union bound (Hoeffding, 1963) imply that, with probability at least $1 - \delta$,

$$\|\epsilon_{\mathcal{A}}\|_2 \leqslant \left(\sum_{i \in \mathcal{A}} \frac{\log(2L/\delta)}{2 n_{iu}} \right)^{1/2} =: \nu_{\mathcal{A}}(\delta). \tag{13}$$

5.3 Stability under Performance-Subspace Perturbation

Historical-matrix perturbation affects both the mismatch term and the anchor gain by displacing the estimated subspace. To control this displacement, define the rank-K spectral gap of P^* as

$$\gamma_K := \sigma_K(P^*) - \sigma_{K+1}(P^*). \tag{14}$$

If $\gamma_K > 0$ and $\|\mathcal{E}_P\|_2 \leqslant \gamma_K/2$, Wedin's singular-subspace perturbation bound (Wedin, 1972) gives:

$$d_{\mathrm{Gr}}(\widehat{\mathcal{S}}, \mathcal{S}^*) \leqslant 2\,\|\mathcal{E}_P\|_2 / \gamma_K =: \delta_P. \tag{15}$$

This controls the subspace-estimation term in the numerator of theorem 1. The following result controls the corresponding change in the denominator for every fixed anchor set.

Proposition 3 (Gain Stability under Subspace Estimation)**.** *The anchor minimum gain is Lipschitz in the Grassmannian projection distance:*

$$\left|\alpha_{\mathcal{A}}(\widehat{\mathcal{S}}) - \alpha_{\mathcal{A}}(\mathcal{S}^*)\right| \leqslant d_{\mathrm{Gr}}(\widehat{\mathcal{S}}, \mathcal{S}^*) \leqslant \delta_P.$$

Corollary 1 (Clean-Gain Reconstruction Error)**.** *Suppose $\gamma_K > 0$, $\|\mathcal{E}_P\|_2 \leqslant \gamma_K/2$, and the anchor-score sampling model in section 5.2 holds. If $\alpha_{\mathcal{A}}(\mathcal{S}^*) > \delta_P$, then the unclipped reconstruction satisfies, with probability at least $1-\delta$ conditional on P,*

$$\|\widehat{p}_u - p_u^*\|_2 \leqslant \frac{r_u^* + \delta_P\,\|p_u^*\|_2 + \nu_{\mathcal{A}}(\delta)}{\alpha_{\mathcal{A}}(\widehat{\mathcal{S}})} \leqslant \frac{r_u^* + \delta_P\,\|p_u^*\|_2 + \nu_{\mathcal{A}}(\delta)}{\alpha_{\mathcal{A}}(\mathcal{S}^*) - \delta_P}. \tag{16}$$

The first inequality also holds whenever $\mathcal{A}$ is identifiable for $\widehat{\mathcal{S}}$, without the clean-gain condition. For any anchor set selected on $\widehat{\mathcal{S}}$, corollary 2 bounds its gain relative to the clean optimum in terms of subspace-estimation error and selection gap.

5.4 From Reconstruction Risk to Classical Design Criteria

For any orthonormal basis U of $\widehat{\mathcal{S}}$, define the information matrix measuring anchor observability:

$$\mathcal{M}_{\mathcal{A}} := U^{\mathrm{T}}\mathcal{R}_{\mathcal{A}}^{\mathrm{T}}\mathcal{R}_{\mathcal{A}}U = (\mathcal{R}_{\mathcal{A}}U)^{\mathrm{T}}(\mathcal{R}_{\mathcal{A}}U). \tag{17}$$

For an identifiable $\mathcal{A}$, definition 2 and equation 17 give the operator-spectrum correspondence

$$\left\|D_{\widehat{\mathcal{S}}\mathcal{A}}\right\|_2 = \alpha_{\mathcal{A}}(\widehat{\mathcal{S}})^{-1} = \lambda_{\min}(\mathcal{M}_{\mathcal{A}})^{-1/2}. \tag{18}$$

For fixed $L \geqslant K$, minimizing the worst-case amplification in theorem 1 motivates E-optimal anchor selection (Nakai et al., 2021), which by equation 18 maximizes the weakest observed direction:

$$\widehat{\mathcal{A}} \in \underset{\mathcal{A}\subseteq[N],\,|\mathcal{A}|=L}{\arg\max}\ \lambda_{\min}(\mathcal{M}_{\mathcal{A}}). \tag{19}$$

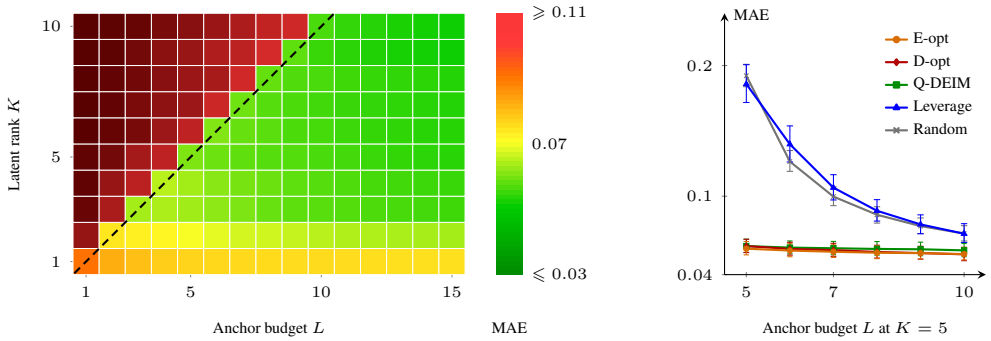
D-optimality maximizes $\log\det(\mathcal{M}_{\mathcal{A}})$ to favor retained volume, while OBOE uses a runtime-constrained fractional log-determinant design (Yang et al., 2019). At $L = K$, Q-DEIM's QR pivots greedily favor volume via $\det(\mathcal{M}_{\mathcal{A}}) = \det(\mathcal{R}_{\mathcal{A}}U)^2$ (Drmač & Gugercin, 2016). MSES greedily selects tasks using a ridge-regularized log determinant in learned latent coordinates (Karuturi et al., 2026), with further criteria in appendix B.5. This analysis traces three error pathways and their common amplification by inverse anchor gain, linking stability to anchor selection.

6 Experiments

For task-side reconstruction, we construct a performance matrix with 132 models and 97 tasks to examine the implications of sections 4 and 5. We assess reconstruction accuracy primarily by MAE, with experimental details and additional experiments, including ranking-preservation diagnostics, in appendix C. Similar patterns appear for the model-side counterpart on P^{T} and on an independent real-world leaderboard; see appendices D and E, respectively.

6.1 Reconstruction across Rank, Budget, and Anchor Selection

We use leave-one-task-out reconstruction with five fixed model folds, selecting anchors outside each evaluation fold and aggregating clipped MAE, with all rules sharing the same estimated subspace, candidate set, and folds; see appendix C.3.



(a) (b)

Figure 1: Reconstruction across rank, budget, and anchor selection. (a) MAE over K and L under E-optimal selection rule; the dashed line marks $L = K$. (b) MAE for five selection rules at $K = 5$.

Reconstruction across rank and budget. We vary $K \in \{1, \cdots, 10\}$ and $L \in \{1, \cdots, 15\}$, selecting anchors for each pair with the E-optimal rule detailed in appendix C.2 and averaging reconstruction MAE over all held-out tasks, with $L < K$ included only as an underdetermined diagnostic. Figure 1a shows consistently high MAE throughout $L < K$, where the pseudoinverse returns a minimum-norm latent completion but the anchor scores do not uniquely identify the target. Reaching $L = K$ makes identifiability possible and sharply reduces MAE, whereas the remaining variation and smaller gains for $L > K$ show that dimensional feasibility alone does not ensure accurate reconstruction. Across ranks, however, increasing K can narrow the spectral gap and heighten reconstruction's worst-case sensitivity to historical-matrix perturbations. Appendix C.10 shows larger Grassmannian displacement at smaller empirical gaps under equal operator-norm perturbations, with anchor-gain changes bounded as in proposition 3. Thus, $L \geqslant K$ removes the dimensional obstruction, but accurate reconstruction still depends on anchor geometry.

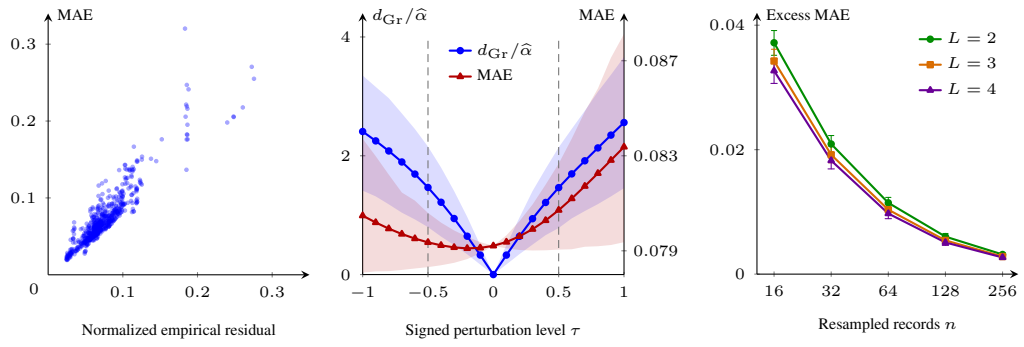
The importance of anchor choice. We fix $K = 5$, vary L from K to $2K$, and compare the E-optimal, D-optimal, and Q-DEIM rules from section 5.4 with two naive baselines, leverage and random, all detailed in appendix C.2. The fixed folds ensure that all methods are tested on identical models and cannot lower their apparent error by selecting difficult models as anchors. Figure 1b shows that the two naive baselines perform substantially worse than the other three rules near $L = K$, with the gaps narrowing as L increases. The same broad pattern persists across all tested ranks, although the ordering among the other three rules varies, as detailed in appendix C.4. Thus, anchor geometry matters most near the minimum feasible budget, where little redundancy is available.

6.2 Reconstruction Error Pathways

We next examine the three error pathways in theorem 1 at a small $K = 2$, where their effects are easier to resolve. Throughout, we select anchors using the E-optimal criterion detailed in appendix C.2.

Target-subspace mismatch. The target performance vector need not lie exactly in a performance subspace estimated from historical tasks. We form 20 splits with approximately 70% historical and 30% held-out tasks, fitting one such subspace and a shared $L = 4$ anchor set per split, and compare each held-out target's normalized projection residual with its reconstruction MAE. Figure 2a shows that the two increase together, with a median within-split Spearman correlation of 0.904, supporting target mismatch as an empirical error source without treating the fitted subspace as an observed clean latent subspace. Experimental details are in appendix C.5.

Historical-matrix perturbation. Using the same split proportions, we form the empirical reference matrix P^{ref} from the historical tasks and obtain its subspace $\mathcal{S}^{\text{ref}}$ with a fixed $L = 2$ anchor set $\mathcal{A}$. We perturb P^{ref} along 8 random directions using signed levels τ with $|\tau| \in \{0.1, 0.2, \cdots, 1.0\}$, then re-estimate the performance subspace and its reconstruction operator while keeping the targets, anchors, and anchor scores fixed. The perturbation construction is given in appendix C.6. Because



(a) (b) (c)

Figure 2: Diagnostics for the three error pathways in theorem 1. (a) MAE against normalized target-subspace residual. (b) Stability-weighted subspace displacement $d_{\mathrm{Gr}}/\widehat{\alpha}$ and MAE versus signed perturbation level; bands are 95% bootstrap intervals, and dashed lines mark $|\tau| = 1/2$. (c) Finite-sample excess MAE relative to empirical full anchor scores; error bars are target-bootstrap intervals.

such perturbations affect both the subspace and the anchor gain, figure 2b reports the stability-weighted displacement $d_{\mathrm{Gr}}/\widehat{\alpha}$, where $\widehat{\alpha} = \alpha_{\mathcal{A}}(\widehat{\mathcal{S}})$, rather than d_{Gr} alone. On both sides, the mean stability-weighted displacement and MAE increase overall with $|\tau|$.

At the target level and with no added anchor noise, we compare the realized unclipped error $E := \|\widehat{p}_u - p_u^{\mathrm{emp}}\|_2$ with the empirical-reference bound $B := \alpha_{\mathcal{A}}(\widehat{\mathcal{S}})^{-1}\left(r_u^{\mathrm{ref}} + d_{\mathrm{Gr}}\,\|p_u^{\mathrm{emp}}\|_2\right)$ from theorem 1. Across all 4,960 target-perturbation pairs, $E \leqslant B$ holds, and $\max(E/B) = 0.215$. The gap reflects two worst-case relaxations: the bound replaces target-specific projector action by $d_{\mathrm{Gr}}\,\|p_u^{\mathrm{emp}}\|_2$ and upper-bounds amplification by $\widehat{\alpha}^{-1}$, discarding direction and target alignment. Although d_{Gr} increases, realized error decreases in 44.9% of successive comparisons. Thus, perturbation magnitude controls the error envelope, while realized error also depends on the perturbation direction and its alignment with the target. Experimental details are in appendix C.6.

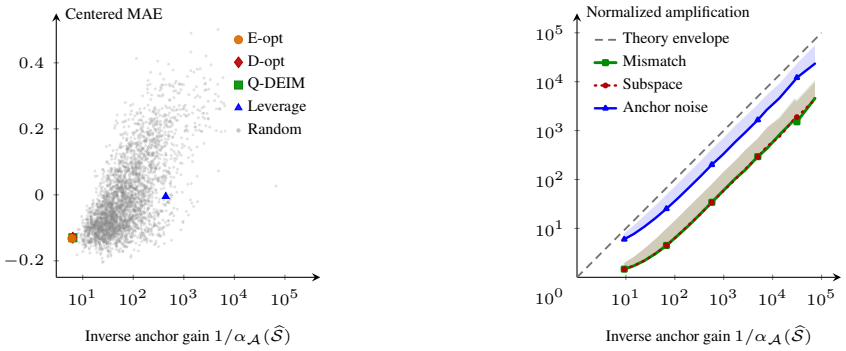
Anchor-evaluation noise. Under the same leave-one-out and five-fold protocol as section 6.1, we use nested anchor sets of sizes $L \in \{2, 3, 4\}$. Using empirical full anchor scores as paired references, we resample $n \in \{2^4, 2^5, 2^6, 2^7, 2^8\}$ correctness records 100 times while holding the performance subspace, anchor ordering, and evaluation folds fixed. Figure 2c shows that finite-sample excess MAE relative to the full-score reconstruction decreases for all L as n increases and is lower for larger anchor sets. Experimental details are in appendix C.7. Together, these diagnostics support empirical target-subspace mismatch, historical-matrix perturbation, and anchor-evaluation noise as distinct pathways to reconstruction error.

6.3 Anchor Gain and Common Amplification

We next test inverse gain as a common amplifier through observational and controlled experiments.

Observed conditioning and error. We revisit the anchor-selection experiment in section 6.1 at $K = L = 5$, comparing each rule's conditioning with its reconstruction error. For each target-fold pair, we sample 100 random anchor sets and center each MAE by the corresponding random-set mean to remove baseline difficulty differences across targets and folds. Figure 3a places E-optimal, D-optimal, and Q-DEIM in the lower-inverse-gain and lower-error region relative to leverage and random selection. Within the random sets, inverse gain and MAE have a median within-pair Spearman correlation of 0.662, meaning that anchor sets with larger inverse gain tend to rank among those with larger errors, which motivates the controlled analysis below. Details are in appendix C.8.

Controlled common amplification. To isolate anchor conditioning, we fix a rank-5 empirical performance subspace and compare 5,000 identifiable $L = 5$ anchor sets spanning a wide inverse-gain range. For each error source, we generate 128 fixed realizations and reuse them across anchor sets,



(a) (b)

Figure 3: Inverse gain and amplification at $K = L = 5$. (a) Inverse gain versus centered MAE for random anchor sets and method centroids. (b) Log-binned controlled amplification, with median curves, 90th-percentile bands, and the theoretical envelope $1/\alpha_{\mathcal{A}}(\widehat{\mathcal{S}})$.

so only anchor geometry varies between reconstruction systems. For target-subspace mismatch, we treat the fixed empirical subspace as the reference clean subspace and use exact coordinate restrictions as noiseless anchor scores, eliminating subspace displacement and anchor-evaluation noise. We then add an orthogonal component of magnitude $\eta = 0.1$ to a vector in the subspace and report

$$g_{\mathrm{mis}}(\mathcal{A}; r) = \left\| D_{\widehat{\mathcal{S}}\mathcal{A}} \mathcal{R}_{\mathcal{A}} p_r^{\mathrm{mis}} - p_r^{\mathrm{mis}} \right\|_2 / \eta.$$

For subspace displacement, directly perturbing the historical matrix would change both the estimated subspace and anchor gain, so we instead fix the estimated subspace and rotate one direction by $\theta = 0.1$ to define a reference clean subspace. We choose the target within it and use noiseless anchor scores, eliminating clean-subspace mismatch and anchor-evaluation noise, and report

$$g_{\mathrm{sub}}(\mathcal{A}; r) = \left\| D_{\widehat{\mathcal{S}}\mathcal{A}} \mathcal{R}_{\mathcal{A}} p_r^{\mathrm{sub}} - p_r^{\mathrm{sub}} \right\|_2 / \sin\theta.$$

For anchor-evaluation noise, we choose the target within the estimated subspace and add noise ϵ_r with $\|\epsilon_r\|_2 = 0.1$ to its exact anchor scores, eliminating both mismatch sources, and report

$$g_{\mathrm{anc}}(\mathcal{A}; r) = \left\| D_{\widehat{\mathcal{S}}\mathcal{A}} \epsilon_r \right\|_2 / \left\| \epsilon_r \right\|_2 .$$

Figure 3b shows that all three normalized amplifications increase with inverse gain while remaining below the common envelope $\alpha_{\mathcal{A}}(\widehat{\mathcal{S}})^{-1}$. Across anchor sets, inverse gain and the per-set 90th-percentile amplification have Spearman correlations of 0.996, 0.996, and 0.998 for the three pathways, respectively, supporting inverse gain as their common amplifier. Details are in appendix C.9.

7 Conclusion

We analyze the identifiability and stability of anchor reconstruction for efficient model-task evaluation through subspace geometry. Geometrically, anchor reconstruction is selected-coordinate recovery on a performance subspace. The $GL(K)$ quotient removes basis ambiguity for a fixed rank-K component, identifying this subspace as the intrinsic object, a Grassmannian point. Anchor reconstruction is identifiable exactly when coordinate restriction is injective on this subspace, equivalently when the anchor minimum gain is positive. The same gain governs stability: our end-to-end bound separates target-subspace mismatch, historical-matrix perturbation, and anchor-evaluation noise, all amplified by its inverse. Historical perturbation can also change the gain itself. These three error sources, their common amplification, and the gain's sensitivity to historical perturbation explain why approximate low rank and enough anchors alone do not ensure stable reconstruction. Experiments on a 132×97 performance matrix and a $4{,}576 \times 6$ public leaderboard illustrate the rank-budget boundary and support the predicted error pathways and inverse-gain amplification. Together, these results give a basis-independent account of anchor reconstruction that connects subspace geometry to identifiability, stability, and the origins, propagation, and amplification of error.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and polishing, identifying potentially relevant related work, and limited coding assistance. We did not use generative AI tools to originate research ideas, design the research methodology, generate theoretical results or proofs, or conduct any part of the research autonomously. We reviewed all AI-assisted work: language edits were checked against the authors' intended meaning, suggested references were verified against the original sources, and AI-assisted code changes were inspected and tested by the authors. We take responsibility for the final content of this work, including text, claims, code, or artifacts produced with the aid of generative AI.

REFERENCES

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

Peter Binev, Albert Cohen, Olga Mula, and James Nichols. Greedy algorithms for optimal measurements selection in state estimation using reduced models. *SIAM/ASA Journal on Uncertainty Quantification*, 6:1101–1126, 2018.

Ryan Burnell, Han Hao, Andrew R. A. Conway, and José Hernández-Orallo. Revealing the structure of language model capabilities. arXiv preprint arXiv:2306.10062, 2023.

Saifon Chaturantabut and Danny C. Sorensen. Nonlinear model reduction via discrete empirical interpolation. *SIAM Journal on Scientific Computing*, 32:2737–2764, 2010.

Zlatko Drmač and Serkan Gugercin. A new selection operator for the discrete empirical interpolation method—improved a priori error bound and extensions. *SIAM Journal on Scientific Computing*, 38:A631–A648, 2016.

Richard Everson and Lawrence Sirovich. Karhunen–loeve procedure for gappy data. *Journal of the Optical Society of America A*, 12:1657–1664, 1995.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafra, and Thomas Wolf. Open LLM leaderboard v2. `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`, 2024.

Nicolo Fusi, Rishit Sheth, and Huseyn Melih Elibol. Probabilistic matrix factorization for automated machine learning. In *Advances in Neural Information Processing Systems 31*, pp. 3352–3361, 2018.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

Siddharth Karuturi, Kaustubh S. Bukkapatnam, Laksh Patel, Tanush Ajay Shastry, Akshath Sharma, Mithil Shah, and Matthew Park. Certifiable evaluation: A low-rank framework for foundation model benchmarking with formal performance guarantees. In *ICML 2026 Workshop on Combining Theory and Benchmarks*, 2026.

Michael W. Mahoney and Petros Drineas. CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106:697–702, 2009.

Aviya Maimon, Amir David Nisan Cohen, Gal Vishne, Shauli Ravfogel, and Reut Tsarfaty. From benchmarks to skills: Low-rank factors for LLM evaluation. *Transactions of the Association for Computational Linguistics*, 14:1660–1684, 2026.

Krithika Manohar, Bingni W. Brunton, J. Nathan Kutz, and Steven L. Brunton. Data-driven sparse sensor placement for reconstruction: Demonstrating the benefits of exploiting known patterns. *IEEE Control Systems Magazine*, 38:63–86, 2018.

Kumi Nakai, Keigo Yamada, Takayuki Nagata, Yuji Saito, and Taku Nonomura. Effect of objective function on data-driven greedy sparse sensor optimization. *IEEE Access*, 9:46731–46743, 2021.

Benjamin Peherstorfer, Zlatko Drmač, and Serkan Gugercin. Stability of discrete empirical interpolation and gappy proper orthogonal decomposition with randomized and deterministic sampling points. *SIAM Journal on Scientific Computing*, 42:A2837–A2864, 2020.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 34303–34326, 2024.

Viktoria Schram, Daniel Beck, and Trevor Cohn. Performance prediction via bayesian matrix factorisation for multilingual natural language processing tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1790–1801, 2023.

G. W. Stewart and Ji-guang Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

Nishant Subramani, Alfredo Gomez, and Mona T. Diab. SimBA: Simplifying benchmark analysis using performance matrices alone. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 13220–13233, 2025.

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12:99–111, 1972.

Chengrun Yang, Yuji Akimoto, Dae Won Kim, and Madeleine Udell. OBOE: Collaborative filtering for AutoML model selection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1173–1183, 2019.

Yuchen Zeng and Dimitris Papailiopoulos. You don't need to run every eval. arXiv preprint arXiv:2606.24020, 2026.

Qiyuan Zhang, Fuyuan Lyu, Xue Liu, and Chen Ma. Collaborative performance prediction for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2576–2596, 2024.

Qinyu Zhao, Ming Xu, Kartik Gupta, Akshay Asthana, Liang Zheng, and Stephen Gould. Can we predict performance of large models across vision-language tasks? In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 77781–77806, 2025.

A Proofs for Problem Formulation and Anchor Reconstruction

This appendix proves the claims used in sections 3 and 4.

A.1 Problem Formulation

Least-squares reconstruction. We derive the coefficient estimate in equation 2 and the reconstruction operator in equation 3.

Proof. Let $y = p_{\mathcal{A}u}$ and suppose $C_{\mathcal{A}} = \mathcal{R}_{\mathcal{A}}C$ has full column rank. The least-squares objective $\|C_{\mathcal{A}}w - y\|_2^2$ has normal equations

$$C_{\mathcal{A}}^{\mathrm{T}} C_{\mathcal{A}} w = C_{\mathcal{A}}^{\mathrm{T}} y.$$

Since $C_{\mathcal{A}}^{\mathrm{T}} C_{\mathcal{A}}$ is invertible, the unique minimizer is

$$\widehat{w}_u = (C_{\mathcal{A}}^{\mathrm{T}} C_{\mathcal{A}})^{-1} C_{\mathcal{A}}^{\mathrm{T}} p_{\mathcal{A}u} = C_{\mathcal{A}}^{\dagger} p_{\mathcal{A}u}.$$

Multiplying by C gives $\widehat{p}_u = C C_{\mathcal{A}}^{\dagger} p_{\mathcal{A}u} = D_{C\mathcal{A}} p_{\mathcal{A}u}$. □

A.2 Subspace Geometry and Identifiability

Performance subspace induced by the factorization. We first verify that every column of the represented rank-K approximation lies in the column space of the basis matrix.

Proof. In equation 1, the j-th column of CW is $Cw_{:j} \in \mathrm{col}(C)$. Since $C \in \mathbb{R}_*^{N\times K}$ has full column rank, $\mathrm{col}(C)$ has dimension K and therefore defines a point $\mathcal{S} = \mathrm{col}(C) \in \mathrm{Gr}(K, N)$. □

Factor orbits and quotient identification. We verify the orbit characterization underlying the canonical bijection in section 4.1.

Proof. The identity acts trivially, and for any $G_1, G_2 \in GL(K)$,

$$\big((C, W) \cdot G_1\big) \cdot G_2 = (C, W) \cdot (G_1 G_2),$$

so equation 4 defines a right action. It preserves the represented rank-K approximation because $(CG)(G^{-1}W) = CW$.

Under the induced action on basis matrices, $C' = CG$ implies $\mathrm{col}(C') = \mathrm{col}(C)$. Conversely, full-column-rank matrices C and C' with the same column space are two bases of that space, so $C' = CG$ for some $G \in GL(K)$. Hence,

$$C' \in \mathcal{O}(C) \iff \mathrm{col}(C') = \mathrm{col}(C).$$

The map $\mathcal{O}(C) \mapsto \mathrm{col}(C)$ is therefore well defined and injective. It is surjective because every K-dimensional subspace of $\mathbb{R}^N$ has a basis matrix in $\mathbb{R}_*^{N\times K}$. □

Basis-independent reconstruction. We prove the orbit invariance in equation 5, which makes the reconstruction operator in definition 1 well defined.

Proof. When $C_{\mathcal{A}} = \mathcal{R}_{\mathcal{A}}C$ has full column rank, $\mathcal{R}_{\mathcal{A}}(CG) = C_{\mathcal{A}}G$ is also full column rank for every $G \in GL(K)$. The full-column-rank pseudoinverse formula then gives

$$(CG)(\mathcal{R}_{\mathcal{A}}(CG))^{\dagger} = CG(C_{\mathcal{A}}G)^{\dagger} = CGG^{-1}C_{\mathcal{A}}^{\dagger} = CC_{\mathcal{A}}^{\dagger}.$$

Thus, the operator is constant on each orbit and depends only on $\mathcal{S} = \mathrm{col}(C)$ and $\mathcal{A}$. □

SVD estimation of the performance subspace. We prove that the top-K left singular subspace solves equation 6 and that this solution is unique when $\sigma_K(P) > \sigma_{K+1}(P)$.

Proof. Let $Q \in \mathbb{R}^{N\times K}$ be an orthonormal basis of a subspace $\mathcal{S}'$, so that $\Pi_{\mathcal{S}'} = QQ^{\mathrm{T}}$. Then

$$\|P - \Pi_{\mathcal{S}'}P\|_F^2 = \|P\|_F^2 - \operatorname{tr}\left(Q^{\mathrm{T}}PP^{\mathrm{T}}Q\right).$$

Minimizing the projection residual is therefore equivalent to maximizing $\operatorname{tr}(Q^{\mathrm{T}}PP^{\mathrm{T}}Q)$ over matrices satisfying $Q^{\mathrm{T}}Q = I_K$. By the Ky Fan maximum principle,

$$\max_{Q^{\mathrm{T}}Q=I_K} \operatorname{tr}(Q^{\mathrm{T}}PP^{\mathrm{T}}Q) = \sum_{k=1}^{K} \sigma_k(P)^2,$$

and the maximum is attained when $\operatorname{col}(Q)$ is a top-K left singular subspace of P. Thus, $\widehat{\mathcal{S}} = \operatorname{col}(U_K)$ solves equation 6. If $\sigma_K(P) > \sigma_{K+1}(P)$, this top-$K$ subspace, and hence the minimizing Grassmannian point $\widehat{\mathcal{S}}$, is unique. □

Equivalent conditions for anchor identifiability. We prove proposition 1 by showing equation 8.

Proof. Since $\mathcal{R}_{\mathcal{A}}|_{\mathcal{S}}$ is the restriction of $\mathcal{R}_{\mathcal{A}}$ to $\mathcal{S}$,

$$\ker(\mathcal{R}_{\mathcal{A}}|_{\mathcal{S}}) = \mathcal{S} \cap \ker(\mathcal{R}_{\mathcal{A}}).$$

A linear map is injective if and only if its kernel is zero, so

$$\mathcal{R}_{\mathcal{A}}|_{\mathcal{S}} \text{ is injective } \iff \mathcal{S} \cap \ker(\mathcal{R}_{\mathcal{A}}) = \{0\}.$$

Moreover, because C is a basis matrix of $\mathcal{S}$ and $C_{\mathcal{A}} = \mathcal{R}_{\mathcal{A}}C$, the map $z \mapsto Cz$ is a bijection from $\ker(C_{\mathcal{A}})$ onto $\mathcal{S} \cap \ker(\mathcal{R}_{\mathcal{A}})$. Hence,

$$\mathcal{S} \cap \ker(\mathcal{R}_{\mathcal{A}}) = \{0\} \iff \ker(C_{\mathcal{A}}) = \{0\} \iff C_{\mathcal{A}} \text{ has full column rank.}$$

□

Existence of identifiable anchor sets. We show that every $\mathcal{S} \in \operatorname{Gr}(K, N)$ admits an identifiable anchor set of any size $K \leqslant L \leqslant N$.

Proof. Let $U \in \mathbb{R}^{N\times K}$ be an orthonormal basis matrix of $\mathcal{S}$. Since $\operatorname{rank}(U) = K$, some set $\mathcal{A}_0 \subseteq [N]$ indexes K linearly independent rows of U. For any $L \geqslant K$, extend $\mathcal{A}_0$ to a set $\mathcal{A}$ of size L. Then $\mathcal{R}_{\mathcal{A}}U$ retains these independent rows and has full column rank, so proposition 1 implies that $\mathcal{A}$ is identifiable for $\mathcal{S}$. □

Coordinates on the performance subspace. We prove the intrinsic inverse relation in equation 9.

Proof. Let $p \in \mathcal{S}$ and write $p = Cz$. Since $C_{\mathcal{A}}$ has full column rank, $C_{\mathcal{A}}^{\dagger}C_{\mathcal{A}} = I_K$. Therefore,

$$D_{\mathcal{S}\mathcal{A}}\mathcal{R}_{\mathcal{A}}p = CC_{\mathcal{A}}^{\dagger}\mathcal{R}_{\mathcal{A}}Cz = CC_{\mathcal{A}}^{\dagger}C_{\mathcal{A}}z = Cz = p.$$

Hence $D_{\mathcal{S}\mathcal{A}}|_{\mathcal{R}_{\mathcal{A}}(\mathcal{S})} = (\mathcal{R}_{\mathcal{A}}|_{\mathcal{S}})^{-1}$. □

B Stability Proofs and Selection Criteria

This appendix proves the results in section 5 in their order of appearance.

B.1 From Identifiability to Stable Reconstruction

Minimum-gain characterization of identifiability. We prove proposition 2 from an orthonormal basis of $\mathcal{S}$ and the intrinsic identifiability condition.

Proof. Let $U \in \mathbb{R}^{N\times K}$ have orthonormal columns spanning $\mathcal{S}$. Every unit vector $s \in \mathcal{S}$ can be written as $s = Uz$ with $\|z\|_2 = 1$, and hence

$$\alpha_{\mathcal{A}}(\mathcal{S}) = \inf_{\|z\|_2=1} \|\mathcal{R}_{\mathcal{A}} U z\|_2 = \sigma_K(\mathcal{R}_{\mathcal{A}} U).$$

This value is independent of the chosen orthonormal basis because replacing U by UQ, where $Q \in O(K)$, does not change the singular values of $\mathcal{R}_{\mathcal{A}} U$.

By equation 8, $\mathcal{A}$ is identifiable for $\mathcal{S}$ exactly when $\mathcal{S} \cap \ker(\mathcal{R}_{\mathcal{A}}) = \{0\}$. This condition fails exactly when some unit $s \in \mathcal{S}$ satisfies $\mathcal{R}_{\mathcal{A}} s = 0$, which is equivalent to $\alpha_{\mathcal{A}}(\mathcal{S}) = 0$. Therefore, $\mathcal{A}$ is identifiable if and only if $\alpha_{\mathcal{A}}(\mathcal{S}) > 0$.

For an identifiable anchor set, let $U_{\mathcal{A}} := \mathcal{R}_{\mathcal{A}} U$. Since $U_{\mathcal{A}}$ has full column rank and $D_{\mathcal{S}\mathcal{A}} = U U_{\mathcal{A}}^{\dagger}$, orthonormality of U gives

$$\|D_{\mathcal{S}\mathcal{A}}\|_2 = \left\|U_{\mathcal{A}}^{\dagger}\right\|_2 = \sigma_K(U_{\mathcal{A}})^{-1} = \alpha_{\mathcal{A}}(\mathcal{S})^{-1}.$$

This also proves equation 10. □

Proposition 4 (Optimal Exact Linear Reconstruction)**.** *Let $\mathcal{A}$ be identifiable for $\mathcal{S}$. Every linear operator $T : \mathbb{R}^L \to \mathbb{R}^N$ satisfying $T\mathcal{R}_{\mathcal{A}} s = s$ for all $s \in \mathcal{S}$ obeys the sharp bound*

$$\|T\|_2 \geqslant \|D_{\mathcal{S}\mathcal{A}}\|_2 = \alpha_{\mathcal{A}}(\mathcal{S})^{-1}.$$

Proof. Compactness of the unit sphere in $\mathcal{S}$ gives a unit vector $s_\star \in \mathcal{S}$ satisfying

$$\|\mathcal{R}_{\mathcal{A}} s_\star\|_2 = \alpha_{\mathcal{A}}(\mathcal{S}).$$

If $T\mathcal{R}_{\mathcal{A}} s = s$ for every $s \in \mathcal{S}$, then

$$1 = \|s_\star\|_2 = \|T\mathcal{R}_{\mathcal{A}} s_\star\|_2 \leqslant \|T\|_2 \; \alpha_{\mathcal{A}}(\mathcal{S}),$$

and therefore $\|T\|_2 \geqslant \alpha_{\mathcal{A}}(\mathcal{S})^{-1} = \|D_{\mathcal{S}\mathcal{A}}\|_2$ by equation 10. □

B.2 The Overall Bound of Reconstruction Error

Unified deterministic reconstruction bound. We first derive equation 12, then prove theorem 1.

Proof. Let U_K be any orthonormal basis of $\widehat{\mathcal{S}}$, so $\widehat{\Pi} = U_K U_K^{\mathrm{T}}$, and write

$$U_{\mathcal{A}K} := \mathcal{R}_{\mathcal{A}} U_K, \qquad D := D_{\widehat{\mathcal{S}}\mathcal{A}} = U_K U_{\mathcal{A}K}^{\dagger}, \qquad Q := D\mathcal{R}_{\mathcal{A}}.$$

Identifiability implies that $U_{\mathcal{A}K}$ has full column rank, so $U_{\mathcal{A}K}^{\dagger} U_{\mathcal{A}K} = I_K$ and

$$Q\widehat{\Pi} = U_K U_{\mathcal{A}K}^{\dagger} \mathcal{R}_{\mathcal{A}} U_K U_K^{\mathrm{T}} = \widehat{\Pi}.$$

Moreover, $Q^2 = Q$ and

$$\|D\|_2 = \alpha_{\mathcal{A}}(\widehat{\mathcal{S}})^{-1}.$$

To verify the projection-norm identity, suppose first that Q is nontrivial and let $\mathcal{V} := \mathrm{range}(Q)$. With respect to the orthogonal decomposition $\mathbb{R}^N = \mathcal{V} \oplus \mathcal{V}^{\perp}$, idempotence gives

$$Q = \begin{pmatrix} I_{\mathcal{V}} & A \\ 0 & 0 \end{pmatrix}, \qquad I - Q = \begin{pmatrix} 0 & -A \\ 0 & I_{\mathcal{V}^{\perp}} \end{pmatrix}$$

for some $A : \mathcal{V}^{\perp} \to \mathcal{V}$. Hence

$$QQ^{\mathrm{T}} = \begin{pmatrix} I_{\mathcal{V}} + AA^{\mathrm{T}} & 0 \\ 0 & 0 \end{pmatrix}, \qquad (I-Q)^{\mathrm{T}}(I-Q) = \begin{pmatrix} 0 & 0 \\ 0 & I_{\mathcal{V}^{\perp}} + A^{\mathrm{T}} A \end{pmatrix}.$$

Therefore,

$$\|Q\|_2^2 = 1 + \|A\|_2^2 = \|I - Q\|_2^2 .$$

Thus $\|I - Q\|_2 = \|Q\|_2$; if $Q = I$, the following bound is immediate. Since $\|\mathcal{R}_{\mathcal{A}}\|_2 = 1$, we have

$$\|I - Q\|_2 \leqslant \|D\|_2 = \alpha_{\mathcal{A}}(\widehat{\mathcal{S}})^{-1}.$$

Using $p_{\mathcal{A}u} = \mathcal{R}_{\mathcal{A}} p_u^* + \epsilon_{\mathcal{A}}$ and $(Q - I)\widehat{\Pi} = 0$ gives

$$\widehat{p}_u - p_u^* = (Q - I)p_u^* + D\epsilon_{\mathcal{A}} = (Q - I)(I - \widehat{\Pi})p_u^* + D\epsilon_{\mathcal{A}}.$$

This is equation 12. Consequently,

$$\|\widehat{p}_u - p_u^*\|_2 \leqslant \alpha_{\mathcal{A}}(\widehat{\mathcal{S}})^{-1} \left(\widehat{r}_u + \|\epsilon_{\mathcal{A}}\|_2\right).$$

Finally,

$$\begin{aligned} \left\|(I - \widehat{\Pi})p_u^*\right\|_2 &\leqslant \|(I - \Pi^*)p_u^*\|_2 + \left\|(\Pi^* - \widehat{\Pi})p_u^*\right\|_2 \\ &\leqslant r_u^* + d_{\mathrm{Gr}}(\widehat{\mathcal{S}}, \mathcal{S}^*) \, \|p_u^*\|_2 . \end{aligned}$$

This also proves equation 11, and substitution gives the second inequality in the theorem. □

B.3 Bounds for the Error Sources

Historical coverage of target-subspace mismatch. The main text keeps

$$r_u^* = \|(I - \Pi^*)p_u^*\|_2$$

as the primary quantity because it directly measures the clean target's mismatch with $\mathcal{S}^*$. This mismatch can also be related to how well the clean historical performance matrix covers the target.

For coefficients $\omega \in \mathbb{R}^M$, the vector $P^*\omega$ is a linear combination of the historical task performance vectors and therefore represents the target using the historical matrix. We use $B \geqslant 0$ as a coefficient-norm budget, restricting this representation to $\|\omega\|_2 \leqslant B$. For a fixed B, the inner infimum below is the smallest target-approximation error under this budget, while $B\sigma_{K+1}(P^*)$ bounds the part of the historical representation outside $\mathcal{S}^*$.

Optimizing over B balances these two terms and defines the historical-coverage surrogate

$$\Gamma_{uK} := \inf_{B \geqslant 0} \left\{ \inf_{\omega \in \mathbb{R}^M, \|\omega\|_2 \leqslant B} \|p_u^* - P^*\omega\|_2 + B\sigma_{K+1}(P^*) \right\}.$$

The target-subspace mismatch is then bounded by this surrogate as $r_u^* \leqslant \Gamma_{uK}$.

Proof. Fix $B \geqslant 0$ and any $\omega \in \mathbb{R}^M$ with $\|\omega\|_2 \leqslant B$. Since $I - \Pi^*$ is an orthogonal projector and $\mathcal{S}^*$ is the top-K left singular subspace of P^*,

$$\begin{aligned} r_u^* &= \|(I - \Pi^*)p_u^*\|_2 \\ &\leqslant \|(I - \Pi^*)(p_u^* - P^*\omega)\|_2 + \|(I - \Pi^*)P^*\omega\|_2 \\ &\leqslant \|p_u^* - P^*\omega\|_2 + \|(I - \Pi^*)P^*\|_2 \, \|\omega\|_2 \\ &\leqslant \|p_u^* - P^*\omega\|_2 + B\sigma_{K+1}(P^*). \end{aligned}$$

Taking the infimum first over admissible ω and then over $B \geqslant 0$ gives $r_u^* \leqslant \Gamma_{uK}$. □

Combining this refinement with the main-text noise specialization in equation 16 gives

$$\|\widehat{p}_u - p_u^*\|_2 \leqslant \alpha_{\mathcal{A}}(\widehat{\mathcal{S}})^{-1} \left(\Gamma_{uK} + \delta_P \, \|p_u^*\|_2 + \nu_{\mathcal{A}}(\delta)\right).$$

Historical-matrix perturbation. We justify the bound in equation 15 using Weyl's inequality and Wedin's singular-subspace perturbation theorem (Wedin, 1972; Stewart & Sun, 1990).

Proof. Weyl's inequality gives

$$\sigma_{K+1}(P) \leqslant \sigma_{K+1}(P^*) + \|\mathcal{E}_P\|_2\,,$$
$$\sigma_K(P) \geqslant \sigma_K(P^*) - \|\mathcal{E}_P\|_2\,.$$

Hence the separation entering Wedin's theorem is at least

$$\min\{\sigma_K(P^*) - \sigma_{K+1}(P),\ \sigma_K(P) - \sigma_{K+1}(P^*)\} \geqslant \gamma_K - \|\mathcal{E}_P\|_2\,.$$

Applying the theorem gives

$$d_{\mathrm{Gr}}(\widehat{\mathcal{S}}, \mathcal{S}^*) = \left\|\widehat{\Pi} - \Pi^*\right\|_2 \leqslant \frac{\|\mathcal{E}_P\|_2}{\gamma_K - \|\mathcal{E}_P\|_2}.$$

Under $\|\mathcal{E}_P\|_2 \leqslant \gamma_K/2$, the right-hand side is at most $2\,\|\mathcal{E}_P\|_2\,/\gamma_K = \delta_P$. □

Anchor-evaluation noise. We derive the conditional finite-sample bound in equation 13 using Hoeffding's inequality and a union bound.

Proof. Conditional on P, the estimated subspace, anchor set, and reconstruction operator are fixed. For each $i \in \mathcal{A}$, let ϵ_i be the corresponding coordinate of $\epsilon_{\mathcal{A}}$ and set

$$t_i := \left(\frac{\log(2L/\delta)}{2n_{iu}}\right)^{1/2}.$$

Hoeffding's inequality gives

$$\Pr(|\epsilon_i| > t_i \mid P) \leqslant 2\exp(-2n_{iu}t_i^2) = \frac{\delta}{L}.$$

A union bound gives $|\epsilon_i| \leqslant t_i$ simultaneously for all $i \in \mathcal{A}$ with conditional probability at least $1 - \delta$; no independence across anchors is needed. On this event,

$$\|\epsilon_{\mathcal{A}}\|_2 \leqslant \left(\sum_{i\in\mathcal{A}} t_i^2\right)^{1/2} = \left(\sum_{i\in\mathcal{A}} \frac{\log(2L/\delta)}{2n_{iu}}\right)^{1/2} = \nu_{\mathcal{A}}(\delta).$$

□

Combining the performance-matrix and anchor-noise estimates with theorem 1 gives the first inequality in equation 16.

B.4 Gain Stability under Subspace Estimation

Gain stability. We prove proposition 3 through a basis-independent Lipschitz bound.

Proof. If $L < K$, rank-nullity gives $\alpha_{\mathcal{A}}(\mathcal{S}) = 0$ for every $\mathcal{S} \in \mathrm{Gr}(K, N)$, so the claim is immediate. Assume henceforth that $L \geqslant K$. For any $\mathcal{S} \in \mathrm{Gr}(K, N)$, let $\Pi_{\mathcal{S}}$ be its orthogonal projector and let U be an orthonormal basis matrix. Since $\Pi_{\mathcal{S}} = UU^{\mathrm{T}}$,

$$(\mathcal{R}_{\mathcal{A}}\Pi_{\mathcal{S}})(\mathcal{R}_{\mathcal{A}}\Pi_{\mathcal{S}})^{\mathrm{T}} = (\mathcal{R}_{\mathcal{A}}U)(\mathcal{R}_{\mathcal{A}}U)^{\mathrm{T}}.$$

Thus, $\mathcal{R}_{\mathcal{A}}\Pi_{\mathcal{S}}$ and $\mathcal{R}_{\mathcal{A}}U$ have the same singular values up to zero padding, so $\alpha_{\mathcal{A}}(\mathcal{S}) = \sigma_K(\mathcal{R}_{\mathcal{A}}\Pi_{\mathcal{S}})$. For $\mathcal{S}_1, \mathcal{S}_2 \in \mathrm{Gr}(K, N)$, Weyl's inequality (Stewart & Sun, 1990) and $\|\mathcal{R}_{\mathcal{A}}\|_2 = 1$ give

$$|\alpha_{\mathcal{A}}(\mathcal{S}_1) - \alpha_{\mathcal{A}}(\mathcal{S}_2)| \leqslant \|\mathcal{R}_{\mathcal{A}}(\Pi_{\mathcal{S}_1} - \Pi_{\mathcal{S}_2})\|_2 \leqslant \|\Pi_{\mathcal{S}_1} - \Pi_{\mathcal{S}_2}\|_2 = d_{\mathrm{Gr}}(\mathcal{S}_1, \mathcal{S}_2).$$

Setting $\mathcal{S}_1 = \widehat{\mathcal{S}}$ and $\mathcal{S}_2 = \mathcal{S}^*$ in this Lipschitz bound gives

$$\left|\alpha_{\mathcal{A}}(\widehat{\mathcal{S}}) - \alpha_{\mathcal{A}}(\mathcal{S}^*)\right| \leqslant d_{\mathrm{Gr}}(\widehat{\mathcal{S}}, \mathcal{S}^*).$$

The perturbation bound in equation 15 further gives $d_{\mathrm{Gr}}(\widehat{\mathcal{S}}, \mathcal{S}^*) \leqslant \delta_P$. □

Proof of corollary 1. The first inequality follows from theorem 1, equation 15, and equation 13, as shown above. By proposition 3, $\alpha_{\mathcal{A}}(\widehat{\mathcal{S}}) \geqslant \alpha_{\mathcal{A}}(\mathcal{S}^*) - \delta_P > 0$. Substituting this margin proves the second inequality. □

Oracle comparison for selected anchors. The following result separates subspace-estimation error from the optimization error of an anchor set selected on $\widehat{\mathcal{S}}$.

Definition 3 (Optimal Anchor Gain and Selection Gap)**.** *For $\mathcal{S} \in \mathrm{Gr}(K, N)$ and $L \geqslant K$, define*

$$\phi_L(\mathcal{S}) := \max_{\mathcal{A}\subseteq[N],\, |\mathcal{A}|=L} \alpha_{\mathcal{A}}(\mathcal{S}).$$

For a size-L anchor set $\widehat{\mathcal{A}}$ selected based on $\widehat{\mathcal{S}}$, define

$$\varepsilon_{\mathrm{sel}} := \phi_L(\widehat{\mathcal{S}}) - \alpha_{\widehat{\mathcal{A}}}(\widehat{\mathcal{S}}) \geqslant 0.$$

Corollary 2 (Oracle Comparison for Data-Dependent Anchor Selection)**.** *The selected set satisfies*

$$\alpha_{\widehat{\mathcal{A}}}(\widehat{\mathcal{S}}) \geqslant \phi_L(\mathcal{S}^*) - d_{\mathrm{Gr}}(\widehat{\mathcal{S}}, \mathcal{S}^*) - \varepsilon_{\mathrm{sel}},$$
$$\alpha_{\widehat{\mathcal{A}}}(\mathcal{S}^*) \geqslant \phi_L(\mathcal{S}^*) - 2d_{\mathrm{Gr}}(\widehat{\mathcal{S}}, \mathcal{S}^*) - \varepsilon_{\mathrm{sel}}.$$

Proof. The basis-independent Lipschitz bound in proposition 3 holds uniformly over anchor sets:

$$|\alpha_{\mathcal{A}}(\mathcal{S}_1) - \alpha_{\mathcal{A}}(\mathcal{S}_2)| \leqslant d_{\mathrm{Gr}}(\mathcal{S}_1, \mathcal{S}_2).$$

Maximizing over all size-L sets therefore gives

$$|\phi_L(\mathcal{S}_1) - \phi_L(\mathcal{S}_2)| \leqslant d_{\mathrm{Gr}}(\mathcal{S}_1, \mathcal{S}_2).$$

By the definition of $\varepsilon_{\mathrm{sel}}$,

$$\alpha_{\widehat{\mathcal{A}}}(\widehat{\mathcal{S}}) = \phi_L(\widehat{\mathcal{S}}) - \varepsilon_{\mathrm{sel}} \geqslant \phi_L(\mathcal{S}^*) - d_{\mathrm{Gr}}(\widehat{\mathcal{S}}, \mathcal{S}^*) - \varepsilon_{\mathrm{sel}},$$

which proves the first inequality. Applying the Lipschitz bound again to $\widehat{\mathcal{A}}$ proves the second. □

B.5 Selection Criteria through the Information Matrix

Let $u_i := U^{\mathrm{T}} e_i \in \mathbb{R}^K$ be the subspace coordinates of f_i. Then $\mathcal{M}_{\mathcal{A}} = \sum_{i\in\mathcal{A}} u_i u_i^{\mathrm{T}}$, so selection criteria can be compared through the directions and magnitudes retained by the same information matrix. Inverse and unregularized log-determinant criteria below require $\mathcal{M}_{\mathcal{A}}$ to be positive definite.

Spectral and noise-based criteria. E-optimality maximizes $\lambda_{\min}(\mathcal{M}_{\mathcal{A}})$ and therefore directly minimizes the worst-case operator amplification in equation 18. A-optimality instead minimizes

$$\mathrm{tr}(\mathcal{M}_{\mathcal{A}}^{-1}) = \left\| D_{\widehat{\mathcal{S}}\mathcal{A}} \right\|_F^2.$$

If anchor noise has mean zero and covariance $\sigma^2 I_L$, the expected squared reconstruction error due solely to this noise is $\sigma^2 \mathrm{tr}(\mathcal{M}_{\mathcal{A}}^{-1})$. G-optimality minimizes $\max_{i\in[N]} u_i^{\mathrm{T}} \mathcal{M}_{\mathcal{A}}^{-1} u_i$, the largest coordinatewise noise variance divided by σ^2 under the same model. These average and coordinatewise criteria need not minimize the worst-case bound in theorem 1.

Determinant criteria and their implementations. D-optimality maximizes $\log\det(\mathcal{M}_{\mathcal{A}}) = 2\sum_{j=1}^{K} \log \sigma_j(\mathcal{R}_{\mathcal{A}} U)$, rewarding retained volume rather than only the weakest direction. MSES uses a greedy regularized log-determinant criterion in its learned latent coordinates (Karuturi et al., 2026). The corresponding comparator in our coordinates chooses $i \notin \mathcal{A}_{\ell-1}$ to maximize $\log\det(\mathcal{M}_{\mathcal{A}_{\ell-1}\cup\{i\}} + \lambda I_K)$ at step ℓ, as in appendix C.2. This is a regularized greedy approximation, not an exact optimizer of unregularized D-optimality.

OBOE uses fractional weights $v \in [0,1]^N$ under a runtime constraint (Yang et al., 2019). Writing $\mathcal{M}(v) := \sum_i v_i u_i u_i^{\mathrm{T}} = U^{\mathrm{T}} \mathrm{diag}(v) U$, its nonsingular weighted design maximizes $\log\det \mathcal{M}(v)$ subject to $\sum_i v_i \widehat{t}_i \leqslant \tau$. If OBOE's factor matrix spans the same subspace and equals UG for $G \in GL(K)$, its original log-determinant differs by the constant $2\log|\det G|$; the runtime constraint and fractional relaxation remain specific to OBOE.

Q-DEIM and leverage selection. Q-DEIM applies column-pivoted QR to U^{T} (Drmač & Gugercin, 2016). At each of its first K pivots, the squared residual row norm is the multiplicative increase in $\det((\mathcal{R}_{\mathcal{A}}U)(\mathcal{R}_{\mathcal{A}}U)^{\mathrm{T}})$; at $L = K$ this Gram determinant equals $\det(\mathcal{M}_{\mathcal{A}})$. Thus Q-DEIM greedily favors retained volume without globally maximizing D-optimality. Our $L > K$ implementation adds the highest-leverage remaining coordinates after the K pivots, as detailed in appendix C.2. Selecting the top-L leverage scores $\|u_i\|_2^2$ exactly maximizes $\mathrm{tr}(\mathcal{M}_{\mathcal{A}}) = \sum_{i\in\mathcal{A}} \|u_i\|_2^2$, but this total can be large even when the selected directions are nearly redundant.

Random and skill-space coverage. If $\mathcal{A}$ is sampled uniformly without replacement from a candidate set $\mathcal{C}$, then

$$\mathbb{E}\,\mathcal{M}_{\mathcal{A}} = \frac{L}{|\mathcal{C}|}\sum_{i\in\mathcal{C}} u_i u_i^{\mathrm{T}}.$$

For $\mathcal{C} = [N]$ this equals $(L/N)I_K$, but isotropy in expectation does not guarantee a well-conditioned realized set. For new-task reconstruction, Maimon et al. (2026) instead greedily select anchor models to maximize their minimum distance in a learned skill space. This coverage objective is not generally determined by $\mathcal{M}_{\mathcal{A}}$, so it is complementary to the information-matrix criteria rather than another exact spectral design.

B.6 Additional Consequences

Non-expansiveness of clipping. We prove that clipping cannot increase reconstruction error.

Proof. Coordinate-wise clipping is the Euclidean projection onto $[0,1]^N$. Since $p_u^* \in [0,1]^N$, non-expansiveness of this projection gives

$$\left\|\mathrm{clip}_{[0,1]}(\widehat{p}_u) - p_u^*\right\|_2 \leqslant \|\widehat{p}_u - p_u^*\|_2 .$$

□

C Experiment Details

This appendix provides the data catalog, protocols, and supplementary diagnostics for the task-side experiments in section 6 on the 132×97 score matrix. It follows the same sequence of rank-budget feasibility and anchor geometry, the three error pathways, and inverse-gain amplification.

C.1 Models and Tasks

Table 1 lists the 52 non-adapter checkpoints by series, version, and functional family, table 2 lists the 80 LoRA adapters with their base models and fine-tuning configurations, and table 3 lists the 97 tasks by semantic family.

Table 1: Non-adapter checkpoints.

Series	Version	Family	Models
Qwen	Qwen-3.5	General	9B, 4B, 4B-Base, 2B, 2B-Base, 0.8B, 0.8B-Base
	Qwen-3	General	4B-Thinking-2507, 4B-Instruct-2507
	Qwen-2.5	General	7B-Instruct, 3B-Instruct, 1.5B, 1.5B-Instruct, 0.5B, 0.5B-Instruct
		Coder	7B, 3B, 1.5B, 0.5B
		Math	7B, 1.5B
	Qwen2	General	1.5B, 0.5B
	Qwen1.5	General	4B, 1.8B, 0.5B
LLaMa	Llama-3.2	General	3B, 3B-Instruct, 1B, 1B-Instruct
	Meta-Llama-3.1	General	8B-Instruct
	Meta-Llama-3	General	8B
	llama-2	General	13b
	codellama	Coder	7b
MediPhi	MediPhi	Medical	MediPhi, Clinical, Guidelines, Instruct, MedCode, MedWiki, PubMed
NextCoder	NextCoder	Coder	14B, 7B
Phi	Phi-3.5-mini	General	instruct
	Phi-3-mini	General	128k-instruct, 4k-instruct
Gemma	Gemma-2	General	9b-it, 2b, 2b-it
	Gemma	General	7b-it, 2b, 2b-it

Table 2: LoRA adapters.

Base Model	Fine-Tune Task	Configurations (learning rate, epochs)
Qwen3.5-4B	CEval	(3e-4, 32), (2e-4, 16), (1e-4, 8)
Qwen3.5-2B	anli_r2	(3e-4, 32), (2e-4, 16), (1e-4, 8)
	arc_easy	(1e-4, 8)
	openbookqa	(1e-4, 8)
	qasc	(1e-4, 8)
Qwen3.5-0.8B	arc_easy	(1e-4, 8)
	qasc	(1e-4, 8)
	xnli_zh	(3e-4, 32), (2e-4, 16), (1e-4, 8)
Qwen2.5-0.5B	openbookqa	(3e-4, 16), (2e-4, 16), (1e-4, 8)
Qwen2.5-Coder-0.5B	CEval	(2e-4, 16)
	HumanEval	(5e-4, 32), (2e-4, 16), (1e-4, 8)
	MBPP	(5e-4, 32), (1e-4, 32)
	algebra_synthetic	(1e-4, 32)
	gsm8k	(5e-4, 32), (2e-4, 32), (1e-4, 32)
Qwen2.5-1.5B	HumanEval	(2e-4, 16), (1e-4, 8)
	MBPP	(5e-4, 32), (2e-4, 32)
	algebra_synthetic	(2e-4, 32), (1e-4, 32)
	gsm8k	(5e-4, 32), (2e-4, 32), (1e-4, 32)
Qwen2.5-Math-1.5B	HumanEval	(5e-4, 32), (2e-4, 16), (1e-4, 8)
	MBPP	(5e-4, 32)
	algebra_synthetic	(5e-4, 32), (2e-4, 32), (1e-4, 32)
	gsm8k	(5e-4, 32), (2e-4, 32), (1e-4, 32)
Qwen2.5-Coder-1.5B	HumanEval	(5e-4, 32), (2e-4, 32), (1e-4, 32)
	MBPP	(5e-4, 32), (1e-4, 32)
	algebra_synthetic	(2e-4, 32), (1e-4, 32)
	gsm8k	(5e-4, 32), (2e-4, 32), (1e-4, 32)
Llama-3.2-1B	HumanEval	(5e-4, 32), (2e-4, 16), (1e-4, 8)
	MBPP	(5e-4, 32), (2e-4, 32), (1e-4, 32)
	algebra_synthetic	(2e-4, 32), (1e-4, 32)
	gsm8k	(5e-4, 32), (2e-4, 32), (1e-4, 32)
	anli_r3	(2e-4, 16), (1e-4, 8)
MediPhi-Clinical	qasc	(3e-4, 32), (2e-4, 16), (1e-4, 8)
MediPhi-Instruct	truthfulqa_mc	(3e-4, 32), (2e-4, 16), (1e-4, 8)
gemma-2-2b	commonsenseqa	(2e-4, 16), (1e-4, 8)
gemma-2-2b-it	commonsenseqa	(1e-4, 8)
	superglue_rte	(2e-4, 16), (1e-4, 8)

Table 3: Task families.

Family	Tasks
Code Generation	HumanEval, MBPP
CS ML Security	college_computer_science, high_school_computer_science, computer_security, machine_learning
Math Numeric	algebra_synthetic, asdiv, gsm8k, multi_arith, svamp
Math	abstract_algebra, college_mathematics, high_school_statistics, elementary_mathematics, high_school_mathematics
Science Engineering	astronomy, college_chemistry, college_physics, conceptual_physics, electrical_engineering, high_school_chemistry, high_school_physics
Medical Biology Health	anatomy, clinical_knowledge, college_biology, college_medicine, high_school_biology, human_aging, human_sexuality, medical_genetics, nutrition, professional_medicine, virology
Law Politics	high_school_government_and_politics, jurisprudence, professional_law, security_studies, us_foreign_policy, international_law
Business Economics	management, econometrics, marketing, business_ethics, public_relations, high_school_macroeconomics, high_school_microeconomics, professional_accounting
Humanities Ethics	moral_disputes, high_school_european_history, high_school_us_history, high_school_world_history, moral_scenarios, philosophy, prehistory, world_religions
Social Science Psychology	high_school_geography, high_school_psychology, professional_psychology, sociology
Logic	formal_logic, logical_fallacies
Broad Exam	CEval, CMMLU, global_facts, miscellaneous
Nli Classification	anli_r1, anli_r2, anli_r3, superglue_cb, superglue_rte, xnli_de, xnli_en, xnli_es, xnli_fr, xnli_zh
Semantic Yes/No	superglue_wic
Commonsense	arc_easy, arc_challenge, commonsenseqa, hellaswag, openbookqa, piqa, qasc, superglue_copa, truthfulqa_mc, winogrande
Multilingual Commonsense	xcopa_et, xcopa_id, xcopa_it, xcopa_sw, xcopa_ta, xcopa_th, xcopa_tr, xcopa_vi, xcopa_zh
Instruction Generation	ifeval_small

Performance-matrix construction. We evaluate each registered model on every retained task and place the resulting task-level score in the corresponding model-task entry. The score is pass@1 for code-generation tasks and empirical accuracy otherwise. We retain the 132 registered model rows and the 97 task columns observed for every retained model, producing the complete matrix $P \in [0,1]^{132\times 97}$ used throughout the experiments without centering or rescaling.

Let $\sigma_1(P) \geqslant \sigma_2(P) \geqslant \cdots$ be the singular values of P. Table 4 reports each singular-value share, its cumulative top-k share, the relative rank-k SVD residual, and the adjacent spectral gap:

$$s_k := \frac{\sigma_k(P)}{\sum_j \sigma_j(P)}, \qquad c_k := \frac{\sum_{\ell=1}^{k} \sigma_\ell(P)}{\sum_j \sigma_j(P)},$$

$$\rho_k^{\mathrm{svd}} := \frac{\|P - P_k\|_F}{\|P\|_F} = \sqrt{\frac{\sum_{j>k} \sigma_j(P)^2}{\sum_j \sigma_j(P)^2}}, \quad \Delta_k := \sigma_k(P) - \sigma_{k+1}(P),$$

where P_k is the rank-k SVD truncation of P. The first singular value accounts for 46.4% of the total, while the leading five, ten, and twenty collectively account for 58.4%, 66.5%, and 76.5%, respectively. Because P is not centered, the dominant component largely reflects overall performance level; the remaining spectrum captures variation beyond this global effect.

Table 4: Singular-spectrum diagnostics for the complete 132×97 performance matrix. The s_k and c_k entries are percentages; ρ_k^{svd} is a relative Frobenius residual, and Δ_k is reported on the original singular-value scale.

k	s_k	c_k	ρ_k^{svd}	Δ_k	k	s_k	c_k	ρ_k^{svd}	Δ_k	k	s_k	c_k	ρ_k^{svd}	Δ_k	k	s_k	c_k	ρ_k^{svd}	Δ_k
1	46.411	46.41	0.1807	47.4481	26	0.684	80.88	0.0583	0.0232	51	0.339	93.03	0.0252	0.0071	76	0.127	98.52	0.0074	0.0104
2	3.976	50.39	0.1599	0.4536	27	0.663	81.54	0.0566	0.0267	52	0.332	93.36	0.0242	0.0177	77	0.118	98.64	0.0069	0.0035
3	3.571	53.96	0.1408	1.4320	28	0.640	82.18	0.0550	0.0155	53	0.317	93.68	0.0233	0.0055	78	0.115	98.75	0.0065	0.0077
4	2.290	56.25	0.1322	0.1742	29	0.626	82.81	0.0533	0.0076	54	0.312	93.99	0.0223	0.0208	79	0.108	98.86	0.0060	0.0067
5	2.134	58.38	0.1243	0.2455	30	0.619	83.42	0.0517	0.0281	55	0.293	94.28	0.0214	0.0092	80	0.102	98.96	0.0056	0.0095
6	1.914	60.30	0.1174	0.0898	31	0.594	84.02	0.0501	0.0088	56	0.285	94.57	0.0206	0.0032	81	0.094	99.05	0.0053	0.0028
7	1.834	62.13	0.1108	0.3483	32	0.586	84.60	0.0486	0.0050	57	0.282	94.85	0.0197	0.0060	82	0.091	99.14	0.0049	0.0047
8	1.523	63.65	0.1060	0.0719	33	0.581	85.19	0.0470	0.0156	58	0.277	95.13	0.0188	0.0176	83	0.087	99.23	0.0046	0.0025
9	1.458	65.11	0.1014	0.1059	34	0.567	85.75	0.0454	0.0368	59	0.261	95.39	0.0180	0.0161	84	0.085	99.32	0.0042	0.0041
10	1.364	66.47	0.0972	0.1273	35	0.534	86.29	0.0440	0.0190	60	0.246	95.64	0.0172	0.0055	85	0.081	99.40	0.0038	0.0119
11	1.250	67.72	0.0935	0.0492	36	0.518	86.80	0.0426	0.0239	61	0.241	95.88	0.0164	0.0135	86	0.070	99.47	0.0035	0.0021
12	1.206	68.93	0.0900	0.0968	37	0.496	87.30	0.0413	0.0048	62	0.229	96.11	0.0157	0.0110	87	0.068	99.54	0.0032	0.0037
13	1.119	70.05	0.0868	0.0614	38	0.492	87.79	0.0399	0.0150	63	0.220	96.33	0.0150	0.0097	88	0.065	99.60	0.0029	0.0026
14	1.064	71.11	0.0838	0.0576	39	0.479	88.27	0.0386	0.0298	64	0.211	96.54	0.0143	0.0094	89	0.063	99.66	0.0026	0.0078
15	1.013	72.13	0.0810	0.0421	40	0.452	88.72	0.0374	0.0064	65	0.203	96.74	0.0136	0.0121	90	0.056	99.72	0.0023	0.0021
16	0.975	73.10	0.0783	0.1007	41	0.446	89.17	0.0362	0.0051	66	0.192	96.93	0.0130	0.0090	91	0.054	99.77	0.0020	0.0071
17	0.885	73.99	0.0761	0.0369	42	0.442	89.61	0.0350	0.0084	67	0.184	97.12	0.0124	0.0007	92	0.048	99.82	0.0017	0.0042
18	0.852	74.84	0.0739	0.0218	43	0.434	90.05	0.0338	0.0244	68	0.183	97.30	0.0118	0.0036	93	0.044	99.87	0.0014	0.0027
19	0.833	75.67	0.0717	0.0090	44	0.412	90.46	0.0326	0.0034	69	0.180	97.48	0.0111	0.0117	94	0.041	99.91	0.0011	0.0051
20	0.824	76.50	0.0696	0.0346	45	0.409	90.87	0.0314	0.0306	70	0.169	97.65	0.0105	0.0099	95	0.037	99.94	0.0008	0.0078
21	0.794	77.29	0.0675	0.0449	46	0.382	91.25	0.0304	0.0017	71	0.160	97.81	0.0100	0.0032	96	0.030	99.97	0.0006	0.0042
22	0.753	78.04	0.0656	0.0049	47	0.380	91.63	0.0293	0.0146	72	0.158	97.97	0.0094	0.0056	97	0.026	100.00	0.0000	0.0293
23	0.749	78.79	0.0637	0.0496	48	0.367	92.00	0.0282	0.0124	73	0.153	98.12	0.0088	0.0143					
24	0.705	79.50	0.0619	0.0091	49	0.356	92.35	0.0272	0.0187	74	0.140	98.26	0.0083	0.0088					
25	0.696	80.19	0.0601	0.0138	50	0.340	92.69	0.0262	0.0008	75	0.132	98.39	0.0078	0.0051					

C.2 Anchor-Selection Rules

All experiments use one of five selection rules applied to an orthonormal basis $U \in \mathbb{R}^{N\times K}$ and a candidate set $\mathcal{C} \subseteq [N]$.

E-optimal. We use a greedy approximation to the E-optimal criterion in equation 19 (Nakai et al., 2021). Starting from $\mathcal{A}_0 = \varnothing$, at step ℓ we set $r_\ell := \min\{\ell, K\}$ and select

$$i_\ell^\star \in \operatorname*{arg\,max}_{i\in\mathcal{C}\setminus\mathcal{A}_{\ell-1}} \sigma_{r_\ell}\big(\mathcal{R}_{\mathcal{A}_{\ell-1}\cup\{i\}} U\big), \qquad \mathcal{A}_\ell := \mathcal{A}_{\ell-1} \cup \{i_\ell^\star\},$$

with ties broken by the smaller coordinate index. Before reaching K anchors, this rule maximizes the smallest available singular value. If $\mathcal{R}_{\mathcal{C}} U$ has full column rank, each step through $\ell = K$ adds an independent row, so the selected set is identifiable once $L \geqslant K$.

D-optimal. Following Karuturi et al. (2026), at step ℓ we select coordinate $i_\ell^\star$ by a greedy D-optimal approximation with ridge $\lambda = 10^{-6}$:

$$i_\ell^\star \in \operatorname*{arg\,max}_{i\in\mathcal{C}\setminus\mathcal{A}_{\ell-1}} \log\det\Big(U^{\mathrm{T}} \mathcal{R}_{\mathcal{A}_{\ell-1}\cup\{i\}}^{\mathrm{T}} \mathcal{R}_{\mathcal{A}_{\ell-1}\cup\{i\}} U + \lambda I_K\Big).$$

Q-DEIM. Q-DEIM computes the column-pivoted factorization

$$(\mathcal{R}_{\mathcal{C}}U)^{\mathrm{T}}\,\Pi = QR$$

and takes the first K pivot coordinates (Drmač & Gugercin, 2016). For $L > K$, the Q-DEIM label denotes a heuristic extension of the original algorithm that adds the $L - K$ remaining candidates with the largest leverage scores $\|u_{i:}\|_2^2$.

Leverage and random. The leverage rule selects the L candidates with largest $\|u_{i:}\|_2^2$ (Mahoney & Drineas, 2009). Across 100 independent draws, random selection takes uniform size-L subsets of $\mathcal{C}$ without replacement. Reported random baselines average over these sets unless stated otherwise.

C.3 Leave-One-Out Reconstruction

For fixed K and L, we hold out each task column $p_{:j}$ in turn, leaving $P^{(-j)}$ as the historical matrix. Using only $P^{(-j)}$, we select anchor models $\mathcal{A}$ using the rules in appendix C.2. We then apply the anchor-reconstruction formulation in section 3 to the held-out task's anchor scores $p_{\mathcal{A}j}$, obtaining the unclipped reconstruction $\widehat{p}_{:j}(\mathcal{A})$, which is clipped to $[0, 1]$ to give $\widetilde{p}_{:j}(\mathcal{A})$.

Thus, $p_{:j}$ is used only for anchor scores and final evaluation, not for subspace estimation or anchor selection. For a single anchor set, we evaluate the clipped reconstruction on its non-anchor models:

$$\mathrm{MAE}_j(\mathcal{A}) := \frac{1}{N-L}\sum_{i\notin\mathcal{A}} |\widetilde{p}_{ij}(\mathcal{A}) - p_{ij}|\,.$$

The following result connects this metric to the Euclidean reconstruction error in theorem 1.

Corollary 3 (Non-Anchor Mean Absolute Error)**.** *Let $\overline{\mathcal{A}} := [N] \setminus \mathcal{A}$, and let $q_{:j}$ be either the raw reconstruction $\widehat{p}_{:j}$ or its coordinate-wise clipping $\widetilde{p}_{:j}$. Then*

$$\frac{1}{N-L}\,\|\mathcal{R}_{\overline{\mathcal{A}}}(q_{:j} - p_{:j})\|_1 \leqslant \frac{1}{\sqrt{N-L}}\,\|\widehat{p}_{:j} - p_{:j}\|_2\,.$$

Proof. The ℓ_1-ℓ_2 inequality and contractivity of coordinate restriction give

$$\frac{1}{N-L}\,\|\mathcal{R}_{\overline{\mathcal{A}}}(q_{:j} - p_{:j})\|_1 \leqslant \frac{1}{\sqrt{N-L}}\,\|q_{:j} - p_{:j}\|_2\,.$$

For the raw reconstruction, the remaining comparison is equality; for the clipped reconstruction, it follows from non-expansiveness of clipping. □

Thus, taking $p_{:j} = p^*_{:j}$, the Euclidean bound in theorem 1 also controls the non-anchor MAE.

However, this metric omits the anchor models and makes the evaluation set depend on the selected set. This can distort comparisons across budgets or selection rules when difficult models are chosen as anchors and removed from evaluation.

Five-fold evaluation. We partition the N models into five fixed folds $\{\mathcal{I}_q\}_{q=1}^5$. For each target j and fold q, anchors are selected only from $[N] \setminus \mathcal{I}_q$, reconstruction uses their target scores, and clipped MAE is evaluated on $\mathcal{I}_q$:

$$\mathrm{MAE}_{j,q}(\mathcal{A}) := \frac{1}{|\mathcal{I}_q|}\sum_{i\in\mathcal{I}_q} |\widetilde{p}_{ij}(\mathcal{A}) - p_{ij}|\,.$$

The fold errors are aggregated as

$$\mathrm{MAE}_j := \frac{1}{N}\sum_{q=1}^{5} |\mathcal{I}_q|\,\mathrm{MAE}_{j,q}(\mathcal{A}_{j,q}).$$

Because the folds partition $[N]$, every model is evaluated as a non-anchor once for each target and setting. The resulting MAE therefore covers all models, and different selection rules can be compared on identical evaluation sets.

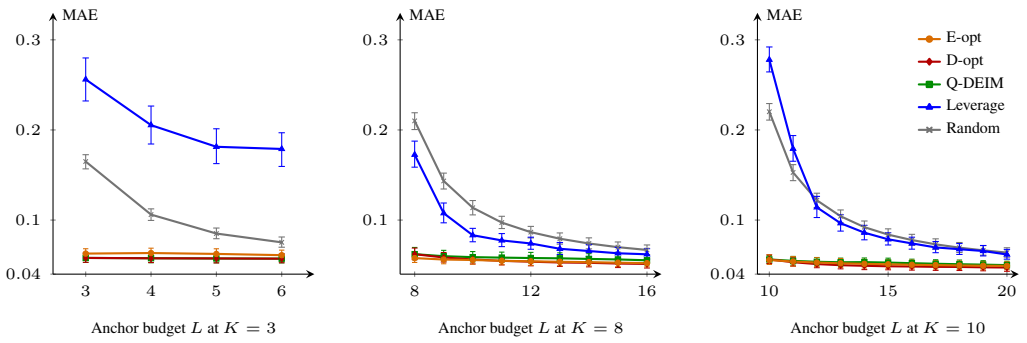
Clipping and ranking preservation. We supplement the anchor-selection comparison in section 6.1 at $K = 5$ and $L \in \{5, 10\}$ using empirical anchor scores without added perturbations. We report the percentage of held-out predictions outside $[0, 1]$, their mean exceedance, and unclipped and clipped MAE. Mean exceedance is the average distance to $[0, 1]$ among out-of-bounds predictions; it and both MAEs are in score units. All statistics retain the evaluation coordinates and averaging weights of the corresponding comparison, including its random draws. Under these weights, the MAE reduction equals the out-of-bounds fraction multiplied by the mean exceedance. Table 5 shows little change in MAE for E-optimal, D-optimal, and Q-DEIM, but substantial clipping effects for leverage and random selection at $L = K$. These effects are much smaller at $L = 10$. Ranking Spearman compares true and clipped scores within each target–fold pair, using average ranks for ties and averaging correlations separately computed for each random draw with the same evaluation weights. Undefined correlations from constant vectors are excluded, with weights renormalized within each target before averaging targets. Only random selection at $L = 5$ has undefined correlations, affecting 3 of 48,500 draws (0.006%); all other settings have none. The three geometric rules also preserve rankings better than the naive baselines, especially at $L = 5$.

Table 5: Clipping diagnostics and ranking preservation at $K = 5$.

L	Rule	Out-of-bounds (%)	Mean exceedance	Unclipped MAE	Clipped MAE	Ranking Spearman
5	E-optimal	0.12	0.0421	0.05991	0.05986	0.559
5	D-optimal	0.20	0.0516	0.06195	0.06185	0.554
5	Q-DEIM	0.20	0.0516	0.06195	0.06185	0.554
5	Leverage	17.17	1.9693	0.52390	0.18584	0.196
5	Random	19.04	5.9247	1.32021	0.19223	0.225
10	E-optimal	0.20	0.0539	0.05582	0.05571	0.579
10	D-optimal	0.08	0.0312	0.05555	0.05552	0.590
10	Q-DEIM	0.15	0.0411	0.05863	0.05857	0.583
10	Leverage	0.59	0.2671	0.07288	0.07130	0.549
10	Random	0.81	0.1465	0.07264	0.07144	0.484

C.4 Anchor Selection at Additional Ranks

To test whether the pattern depends on $K = 5$, we repeat the experiment for $K \in \{3, 8, 10\}$ and every integer budget $K \leqslant L \leqslant 2K$. The targets and coordinate folds are shared across ranks, with 100 random draws per rank; only the estimated rank-K basis and resulting anchor selections change.



(a) (b) (c)

Figure 4: Anchor-selection comparisons at additional fixed ranks. Mean clipped MAE under five-fold evaluation for (a) $K = 3$, (b) $K = 8$, and (c) $K = 10$; error bars are bootstrap 95% intervals.

Across the additional ranks, E-optimal, D-optimal, and Q-DEIM outperform leverage and random selection throughout the reported budgets, although their ordering varies and the gaps generally narrow as L grows. At $K = 3$, leverage is consistently less accurate than random selection, showing that large individual row norms need not produce complementary coordinates.

C.5 Target-Subspace Mismatch Details

We construct 20 task-family-stratified splits by randomly selecting held-out targets within each family in appendix C.1 while retaining at least one historical task in families with multiple members, yielding roughly 70% historical tasks and 30% held-out targets per split. That is, each split contains 66 historical tasks and 31 held-out targets. For each split s, we estimate a K-dimensional subspace $\mathcal{S}_K^{(s)}$ with $K = 2$ from its historical tasks as in equation 6 and select a shared anchor set $\mathcal{A}^{(s)}$ with $L = 4$ using the E-optimal rule in appendix C.2. Neither the subspace nor the anchors use held-out targets. For each held-out target t_u, let p_u^{emp} denote its full empirical performance vector and measure its normalized projection residual

$$r_{u,\text{emp}}^{(s)} := \frac{1}{\sqrt{N}} \left\| \left(I - U_K^{(s)} (U_K^{(s)})^{\mathrm{T}} \right) p_u^{\text{emp}} \right\|_2,$$

where $U_K^{(s)}$ is an orthonormal basis of $\mathcal{S}_K^{(s)}$. This fitted-subspace residual serves as an empirical counterpart to the unobserved r_u^* in section 5.2. Using $\mathcal{A}^{(s)}$, we reconstruct each target from its empirical anchor scores as in section 3, clip to $[0, 1]$, and compute MAE over non-anchor models.

Within each split, we compute the Spearman correlation ρ_s between the residual and MAE across its 31 targets. Table 6 reports the complete split-level results.

Table 6: Within-split Spearman correlations between empirical target-subspace residual and reconstruction MAE.

s	ρ_s	s	ρ_s	s	ρ_s	s	ρ_s	s	ρ_s
0	0.946	4	0.904	8	0.946	12	0.879	16	0.904
1	0.895	5	0.929	9	0.929	13	0.883	17	0.834
2	0.910	6	0.899	10	0.927	14	0.848	18	0.912
3	0.883	7	0.881	11	0.951	15	0.909	19	0.904

All 20 correlations are positive, ranging from 0.834 to 0.951. Their median is 0.904, with a 95% interval of $[0.889, 0.919]$ from 2,000 bootstrap resamples of the split-level correlations. Figure 2a pools the 620 raw pairs $(r_{u,\text{emp}}^{(s)}, \text{MAE}_u(\mathcal{A}^{(s)}))$, while the correlations remain computed within splits to keep the fitted subspace and anchor system fixed. The consistent positive association supports empirical target-subspace mismatch as an error source without treating the fitted SVD subspace as the unobserved clean subspace $\mathcal{S}^*$.

C.6 Historical-Matrix Perturbation Details

For one task-family-stratified split constructed as in appendix C.5, we retain 66 historical tasks and 31 held-out targets. Let $P^{\text{ref}} \in \mathbb{R}^{N \times 66}$ be the historical block and p_u^{emp} the full empirical vector of each target. These empirical references are not population-level clean quantities. At $K = L = 2$, write a full SVD $P^{\text{ref}} = U \Sigma V^{\mathrm{T}}$, set $\mathcal{S}^{\text{ref}} = \text{col}(U_K)$, and denote the fixed E-optimal anchors selected from U_K as in appendix C.2 by $\mathcal{A}^{(0)}$. The empirical spectral gap is

$$\widehat{\gamma}_K := \sigma_K(P^{\text{ref}}) - \sigma_{K+1}(P^{\text{ref}}).$$

Let u_j and v_j denote the columns of U and V. For $q = 1, \cdots, 8$, we independently draw $G_q \in \mathbb{R}^{N \times 66}$ with i.i.d. $\mathcal{N}(0, 1)$ entries and define

$$\widetilde{H}_q = \left(u_{K+1}^{\mathrm{T}} G_q v_K \right) u_{K+1} v_K^{\mathrm{T}} + \left(u_K^{\mathrm{T}} G_q v_{K+1} \right) u_K v_{K+1}^{\mathrm{T}}, \quad H_q = \frac{\widetilde{H}_q}{\left\| \widetilde{H}_q \right\|_2}.$$

These 8 directions couple the K-th and $(K+1)$-th singular directions across the truncation boundary without using target scores. For $\tau \in \{-1.0, -0.9, \cdots, -0.1, 0.1, \cdots, 0.9, 1.0\}$, we apply

$$E_{P,q,\tau} = \tau \widehat{\gamma}_K H_q, \qquad P'_{q,\tau} = P^{\text{ref}} + E_{P,q,\tau},$$

so $\|E_{P,q,\tau}\|_2 = |\tau| \widehat{\gamma}_K$ for every direction. Let $\widehat{\mathcal{S}}_{q,\tau}$ be the top-K left singular subspace of $P'_{q,\tau}$, with $\mathcal{A}^{(0)}$ and the target anchor scores held fixed. Hence $|\tau| = 1/2$ gives $\|E_{P,q,\tau}\|_2 = \widehat{\gamma}_K/2$,

the empirical boundary of the sufficient perturbation condition in equation 15. The dashed lines in figure 2b mark this boundary, not a necessary condition or empirical threshold.

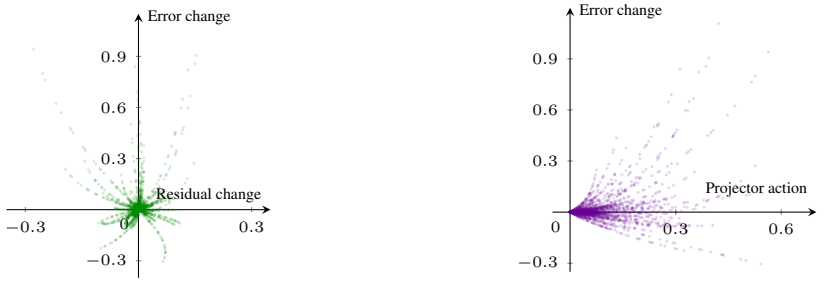
For the empirical-reference bound comparison in section 6.2, define the target residual

$$r_u^{\mathrm{ref}} := \|(I - \Pi_{\mathcal{S}^{\mathrm{ref}}})p_u^{\mathrm{emp}}\|_2 .$$

For each target, we compute MAE over non-anchor models after clipping the reconstruction to $[0, 1]$. In figure 2b, MAE is averaged over the 31 targets within each direction. At each signed τ, means and 95% bootstrap intervals are computed across the 8 independent directions using 10,000 resamples.

All $8 \times 20 = 160$ perturbed systems remain identifiable. All 4,960 checks of $E \leqslant B$ pass with numerical tolerance $10^{-10} \max\{1, B\}$, with maximum, mean, and median E/B of 0.215, 0.093, and 0.087. The bound replaces the projector difference acting on each target by $d_{\mathrm{Gr}}(\widehat{\mathcal{S}}_{q,\tau}, \mathcal{S}^{\mathrm{ref}})\, \|p_u^{\mathrm{emp}}\|_2$ and bounds amplification by $\alpha_{\mathcal{A}^{(0)}}(\widehat{\mathcal{S}}_{q,\tau})^{-1}$, discarding direction and target alignment.

We use figure 5 to examine whether error reductions reflect a better subspace fit and whether subspace changes acting on a target necessarily worsen reconstruction, comparing each perturbed outcome with its unperturbed reference. The error E is lower in 2,370 of the 4,960 pairs (47.8%). Among these cases, the estimated-subspace residual changes by 1.42×10^{-4} on average and increases in 957 cases, while the target-specific projector action has mean 0.0745. Figure 5a shows that error can decrease even when the target residual increases, so the decrease does not necessarily come from a better subspace fit. Figure 5b shows that nonzero target-specific projector action can accompany lower error, so its magnitude alone does not determine whether reconstruction improves or worsens. Both panels display every second reference-relative outcome without filtering by error change. Thus, historical perturbation affects the error envelope through both subspace displacement and anchor gain, while realized error also depends on its direction and alignment with the target. The worst-case bound does not imply monotonic growth of realized error.



(a) (b)

Figure 5: Directional diagnostics for the signed perturbations in figure 2b. (a) Changes in the target residual and reconstruction error relative to the reference system. (b) Target-specific projector action against reference-relative error change. Both panels show every second pair for legibility.

C.7 Anchor-Evaluation Noise Details

At fixed $K = 2$ and L, for each held-out task and evaluation fold, let $\mathcal{S}$ be the fitted historical subspace and $\mathcal{A}$ the selected anchor set. Each anchor's empirical full score p_{iu}^{full} averages all available binary correctness records. We draw n record indices with replacement from each anchor's fixed pool to obtain $\overline{p}_{iu}^{(n)}$. Here n counts resampled records from fixed pools, not new online evaluations. The empirical anchor noise and full-reference reconstruction are

$$\epsilon_{\mathcal{A}}^{(n)} := \overline{p}_{\mathcal{A}u}^{(n)} - p_{\mathcal{A}u}^{\mathrm{full}}, \qquad \widehat{p}_u^{\mathrm{full}} := D_{\mathcal{S}\mathcal{A}} p_{\mathcal{A}u}^{\mathrm{full}}.$$

Linearity and the inverse-gain bound give

$$\widehat{p}_u^{(n)} - \widehat{p}_u^{\mathrm{full}} = D_{\mathcal{S}\mathcal{A}} \epsilon_{\mathcal{A}}^{(n)}, \qquad \left\|\widehat{p}_u^{(n)} - \widehat{p}_u^{\mathrm{full}}\right\|_2 \leqslant \alpha_{\mathcal{A}}(\mathcal{S})^{-1} \left\|\epsilon_{\mathcal{A}}^{(n)}\right\|_2.$$

We report excess MAE on non-anchor evaluation models relative to the paired full-score reconstruction, together with $\left\|\epsilon_{\mathcal{A}}^{(n)}\right\|_2$ and the normalized unclipped change $\left\|\widehat{p}_u^{(n)} - \widehat{p}_u^{\mathrm{full}}\right\|_2 / \sqrt{N}$. We repeat

the procedure $b = 100$ times for each n and average results within each target-fold pair, weighting folds by size. Intervals come from 2,000 paired bootstrap resamples of the 97 target-level values.

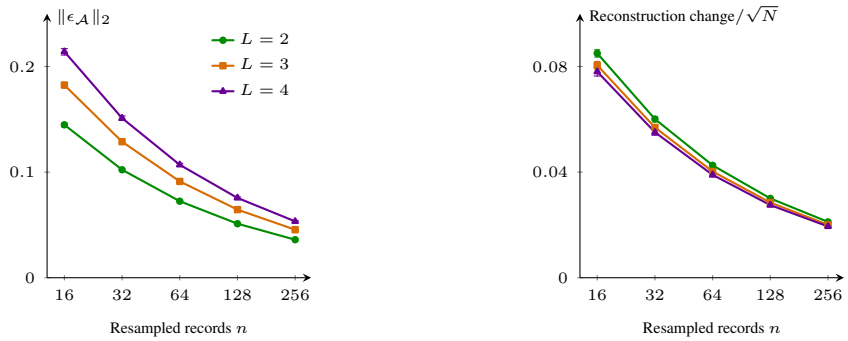


Figure 6: Intermediate diagnostics for controlled anchor-evaluation noise at $K = 2$. Left: empirical anchor-noise norm. Right: unclipped reconstruction change relative to the empirical full-score reconstruction. Curves show target means with bootstrap intervals.

To interpret the lower excess MAE at larger n and L in figure 2c, figure 6 reports the anchor-noise norm and its unclipped reconstruction effect. The anchor-noise norm decays approximately as $n^{-1/2}$, with fitted log-log slopes of -0.501 for all three budgets, while normalized reconstruction change falls roughly fourfold from $n = 16$ to 256. More coordinates raise the source norm, but mean realized amplification falls from 6.61 ($L = 2$) to 4.02 ($L = 4$), below inverse gains 7.58 and 5.21. All 727,500 trials satisfy the inverse-gain envelope within 10^{-10}, with a maximum nonzero realized-to-envelope ratio of 1.000. Here, larger n reduces score noise and larger L reduces amplification.

C.8 Observed Conditioning and Error Details

For each target-fold pair, let $x_{j,q}(\mathcal{A}) := \alpha_{\mathcal{A}}(\widehat{\mathcal{S}}^{(-j)})^{-1}$ be the inverse gain, where $\widehat{\mathcal{S}}^{(-j)}$ is fitted without task j. We center the fold-level clipped MAE by the random-set mean for the same pair:

$$y_{j,q}(\mathcal{A}) := \mathrm{MAE}_{j,q}(\mathcal{A}) - \frac{1}{100}\sum_{r=1}^{100} \mathrm{MAE}_{j,q}(\mathcal{A}^{(r)}).$$

The same mean centers all methods, so $y_{j,q} = 0$ marks the random baseline. Figure 3a shows 3,000 random-set points subsampled without replacement for legibility, while each deterministic marker is the arithmetic centroid of $(x_{j,q}, y_{j,q})$ across target-fold pairs. We compute Spearman correlation within each pair's random sets; the median correlation's 95% interval $[0.653, 0.669]$ comes from 2,000 bootstrap resamples of the pair-level correlations.

C.9 Controlled Common Amplification Details

We compare 5,000 anchor sets, including the greedy E-optimal set and uniformly sampled coordinate subsets with $\alpha_{\mathcal{A}}(\widehat{\mathcal{S}}) \geqslant 10^{-5}$. Their inverse gains range from 6.24 to 9.01×10^4. The same 128 random error directions per source are reused across sets and are independent of anchor choice.

For target-subspace mismatch, take unit vectors $s_r \in \widehat{\mathcal{S}}$ and $e_{\perp r} \in \widehat{\mathcal{S}}^{\perp}$, and set

$$p_r^{\mathrm{mis}} = s_r + \eta e_{\perp r}, \qquad \eta = 0.1.$$

With $\mathcal{S}^* = \widehat{\mathcal{S}}$ and noiseless anchor scores, this leaves only $r^* = \eta$.

For subspace displacement, take unit vectors $q_r \in \widehat{\mathcal{S}}$ and $b_{\perp r} \in \widehat{\mathcal{S}}^{\perp}$ and define

$$\mathcal{S}_r^* = \left(\widehat{\mathcal{S}} \cap q_r^{\perp}\right) \oplus \mathrm{span}\left\{\cos\theta\, q_r + \sin\theta\, b_{\perp r}\right\}, \qquad \theta = 0.1.$$

Taking $p_r^{\mathrm{sub}} = \cos\theta\, q_r + \sin\theta\, b_{\perp r} \in \mathcal{S}_r^*$ gives $r^* = 0$ and $d_{\mathrm{Gr}}(\widehat{\mathcal{S}}, \mathcal{S}_r^*) = \sin\theta$ with noiseless anchor scores. This isolates post-estimation amplification without perturbing the historical matrix.

For anchor-evaluation noise, take $p_r^{\mathrm{anc}} \in \widehat{\mathcal{S}}$ and $\epsilon_r \in \mathbb{R}^5$ with $\|\epsilon_r\|_2 = 0.1$. Add the same ϵ_r to each set's exact scores in increasing coordinate order; both mismatch terms vanish.

For each set, we summarize the normalized errors in section 6.3 by their median and 90th percentile. Figure 3b groups sets into 18 logarithmic inverse-gain bins, plotting geometric-mean inverse gain against the median per-set statistic in each bin. The per-set 90th percentiles also enter the Spearman comparisons reported in the main text. The mismatch and subspace median curves nearly overlap in the plot. All realizations remain below $1/\alpha_{\mathcal{A}}(\widehat{\mathcal{S}})$; anchor noise nearly reaches it (maximum ratio 0.998), whereas mismatch and subspace reach 0.415 and 0.413 for the sampled directions.

C.10 Rank, Representation, and Stability

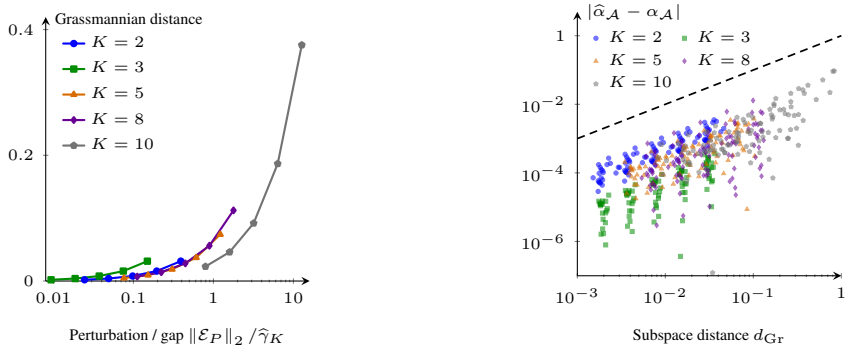
Purpose. Increasing K may reduce the representation residual but raises the latent dimension and minimum identifiable budget, while smaller anchor gain or spectral gap can weaken worst-case stability under historical-matrix perturbations, motivating our tests of subspace displacement under equal operator-norm perturbations and the gain-change bound in proposition 3.

Setup. On one task-family-stratified split with 66 historical tasks and 31 held-out targets as in appendix C.5, we estimate the leading subspace using only the historical-library block for each $K \in \{2, 3, 5, 8, 10\}$. For 20 Gaussian directions, we rescale the perturbation to

$$\|\mathcal{E}_P\|_2 / \sigma_1(P_{\mathrm{lib}}) \in 10^{-3} \times \{0, 0.25, 0.5, 1, 2, 4\}.$$

The plotted horizontal coordinate divides this common absolute perturbation by the rank-specific empirical gap $\widehat{\gamma}_K = \sigma_K(P_{\mathrm{lib}}) - \sigma_{K+1}(P_{\mathrm{lib}})$. We recompute the leading subspace after perturbation and measure its projection distance from the unperturbed subspace. For the anchor set selected on the clean subspace, we also record the absolute change in anchor gain.

Results. Figure 7a shows that ranks with smaller empirical gaps undergo larger subspace movement under the same absolute perturbation. The small-gap $K = 10$ subspace is especially sensitive because the K-th retained direction is weakly separated from the $(K+1)$-th discarded direction. Thus, increasing K may lower the rank-K representation residual while making the estimated subspace less stable when the spectrum is locally flat. Expressing the horizontal axis as $\|\mathcal{E}_P\|_2 / \widehat{\gamma}_K$ organizes these sensitivities around the perturbation quantity used by the theory. Figure 7b compares gain change with Grassmannian distance for all 500 rank-perturbation draws. Every point lies below the identity line, and the largest observed ratio is 0.131, consistent with the Lipschitz bound in proposition 3. Together with the higher latent dimension and possible loss of anchor gain, these results show that increasing K trades lower representation residual against a harder and potentially less stable reconstruction problem.



(a) (b)

Figure 7: Subspace and anchor-gain stability under Gaussian historical-matrix perturbations. (a) Performance-subspace movement across ranks and empirical gaps. (b) Absolute anchor-gain change against Grassmannian distance; the dashed identity line is the theoretical upper bound.

C.11 Reusable-Anchor Block Holdout

Setup. We test whether anchor choice still matters near $L = K$ when a subspace and anchor set constructed from historical tasks are reused across a future task block. We generate ten task-family-stratified splits as in appendix C.5, each with 66 historical tasks and 31 future tasks. For each split and fixed coordinate fold from appendix C.3, we fit the rank-5 subspace and select anchors at fixed L using only the historical library, then reuse both for every target in the held-out block. Anchors cannot enter the evaluation fold, giving the five selection rules from appendix C.2 identical evaluation support; random results average ten selections. We evaluate every budget $K \leqslant L \leqslant 2K$. Each target is averaged across the splits in which it is held out before applying target-level bootstrap.

Results. Figure 8a shows that E-optimal, D-optimal, and Q-DEIM substantially outperform leverage and random selection near the minimum budget, while the gaps narrow as L grows. The target-bootstrap intervals confirm that this separation is not driven by a few held-out tasks.

C.12 Composition and Cold-Start Stress Tests

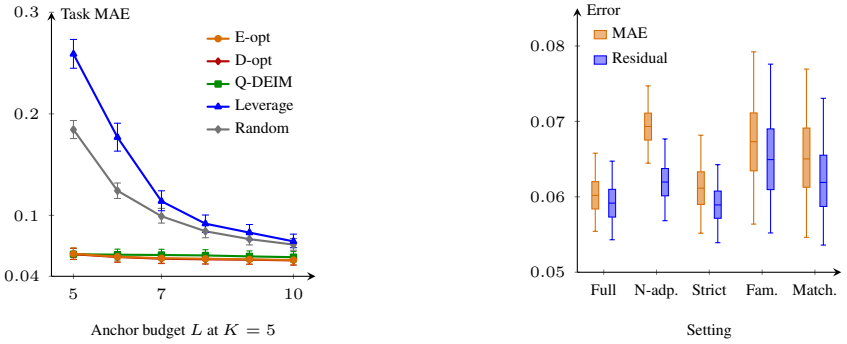
Setup. We test the sensitivity of minimum-budget reconstruction to adapter composition and the availability of related historical tasks. We use the leave-one-out reconstruction in appendix C.3 at $K = L = 5$, evaluating clipped non-anchor MAE. For each setting s, we also report $\left\|(I - \Pi_{\widehat{\mathcal{S}}_s})p_{u,s}\right\|_2 / \sqrt{N_s}$, where $p_{u,s} \in \mathbb{R}^{N_s}$ is the target vector on its N_s retained models and $\widehat{\mathcal{S}}_s \subseteq \mathbb{R}^{N_s}$ is the estimated subspace. After each historical exclusion, we re-estimate the subspace and select E-optimal anchors as in appendix C.2.

Let $\mathcal{F} := \{f_i\}_{i=1}^N$ and $\mathcal{T} := \{t_j\}_{j=1}^M$ denote the model and task collections in section 3.1. For task $t \in \mathcal{T}$, let $\mathrm{Fam}(t) \subseteq \mathcal{T}$ be the tasks in its family and $\mathrm{Adapt}(t) \subseteq \mathcal{F}$ the adapters fine-tuned on it, and set $\mathrm{Adapter} := \bigcup_{t \in \mathcal{T}} \mathrm{Adapt}(t)$. Let $\mathcal{J}_u$ be a uniformly sampled subset of $\mathcal{T} \setminus \{t_u\}$ with $|\mathcal{J}_u| = |\mathrm{Fam}(t_u)| - 1$. The five settings retain the model and historical-task subsets shown below:

Setting	Models	Historical tasks
Full	$\mathcal{F}$	$\mathcal{T} \setminus \{t_u\}$
Non-adapter	$\mathcal{F} \setminus \mathrm{Adapter}$	$\mathcal{T} \setminus \{t_u\}$
Strict new-task	$\mathcal{F} \setminus \mathrm{Adapt}(t_u)$	$\mathcal{T} \setminus \{t_u\}$
Family holdout	$\mathcal{F}$	$\mathcal{T} \setminus \mathrm{Fam}(t_u)$
Matched random	$\mathcal{F}$	$\mathcal{T} \setminus (\{t_u\} \cup \mathcal{J}_u)$

The Non-adapter setting retains the 52 non-adapter checkpoints in table 1. The bootstrap distributions in figure 8b use 97 target-level values for Full, Non-adapter, and Strict new-task. Family holdout and Matched random cover 95 targets in 15 families with at least two tasks. For Matched random, we average metrics over 10 independent draws of $\mathcal{J}_u$ per target, each matched to Family holdout in historical-library size. We then average targets within each family for these two settings and bootstrap the family means, giving every family equal weight.

Results. Figure 8b shows that Strict new-task nearly matches Full in both metrics, so the results are not driven by adapters fine-tuned on the held-out task. Reconstruction also remains viable on the 52 non-adapter checkpoints, although its MAE is higher at the minimum anchor budget, showing that the phenomenon extends beyond adapter duplication. Family holdout removes the same number of historical tasks as Matched random but yields slightly higher mean MAE and residual, supporting a role for task-related history beyond history size alone. The concurrent increase in residual and MAE agrees with the target-subspace mismatch mechanism in section 5.2. Paired family-level bootstrap gives positive mean differences for both MAE and residual, and the increases occur in $12/15$ and $11/15$ families, respectively. The remaining heterogeneity means these results describe an aggregate trend rather than uniform degradation.



(a) (b)

Figure 8: (a) Reusable-anchor reconstruction at $K = 5$ over ten historical / future task-block splits. (b) Bootstrap distributions of mean MAE and projection residual under composition controls at $K = L = 5$; whiskers mark 95% intervals.

C.13 Random Seeds

The stochastic parts of the task-side experiments use the seeds listed below. The rank-K fit and the E-optimal, D-optimal, Q-DEIM, and leverage selections are deterministic once their inputs are fixed. Some ranges denote repetition labels rather than literal generator seeds.

Rank-budget and anchor selection. Seed 2027 defines the model folds. In selection comparisons, random labels $0 \sim 99$ use base 70,000,000 with target and fold offsets, and the orders are reused across ranks and budgets. Bootstrap base 120,000 adds rank, budget, and method offsets.

Target-subspace mismatch. Labels $0 \sim 19$ define the task-family splits, and seed 6427 is used for the split-level correlation bootstrap.

Historical-matrix perturbation. Seed 2027 defines the reference split and the generator of the 8 independent Gaussian draws used to construct the perturbation directions. Bootstrap seeds use the same base with level and metric offsets.

Anchor-evaluation noise. Seed 2027 defines the model folds. Record resampling uses base 31,000,000 with target, fold, sample-size, and repetition offsets. The target bootstrap uses base 72,027 with budget, sample-size, and metric offsets.

Observed conditioning and error. Seed 2027 sets the model folds and plotted-point subsampling. Random labels $0 \sim 99$ use target and fold offsets; seed 4427 controls the correlation bootstrap.

Controlled common amplification. Seed 2027 sets the task split and sampled anchor sets; 31415 generates shared error sources and 31416 the in-subspace anchor-noise targets.

Rank, representation, and stability. Seed 2027 defines the task split. Base 83,000 plus labels $0 \sim 19$ generates Gaussian directions. Bootstrap bases 84,000, 85,000, and 85,500 add rank and perturbation-level offsets.

Reusable-anchor block holdout. Labels $0 \sim 9$ index splits and random draws. Random anchor choices use base 70,000,000 with split and fold offsets, while seed 2027 defines the model folds. Bootstrap base 81,000 adds budget offsets.

Composition and cold-start controls. Matched-random labels $0 \sim 9$ use base 20,000,000 with target offsets. Bootstrap bases 93,000 for MAE and 94,000 for residual add setting offsets; base 95,000 for paired family contrasts adds metric offsets.

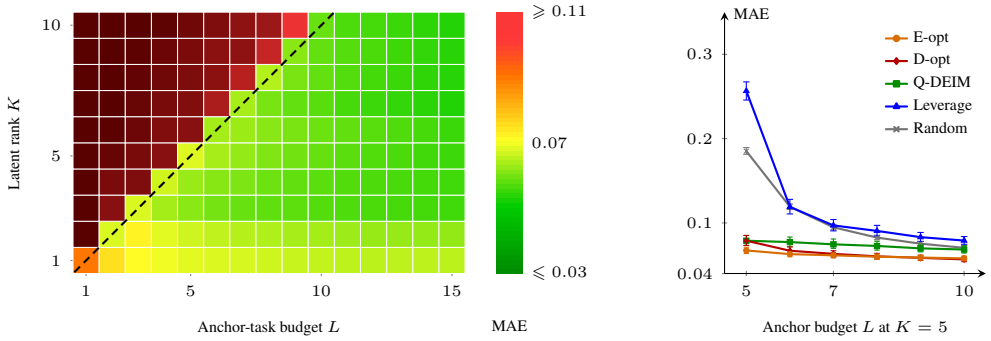
D Model-Side Experiments on the Transposed Matrix

For a new model, model-side anchor reconstruction uses its scores on a small set of anchor tasks to estimate its full performance vector across tasks. We apply each task-side experiment in section 6 and appendix C to P^{T} with the same hyperparameters and protocol choices after exchanging the two matrix axes. We therefore report only the model-side results.

D.1 Reconstruction across Rank, Budget, and Anchor Selection

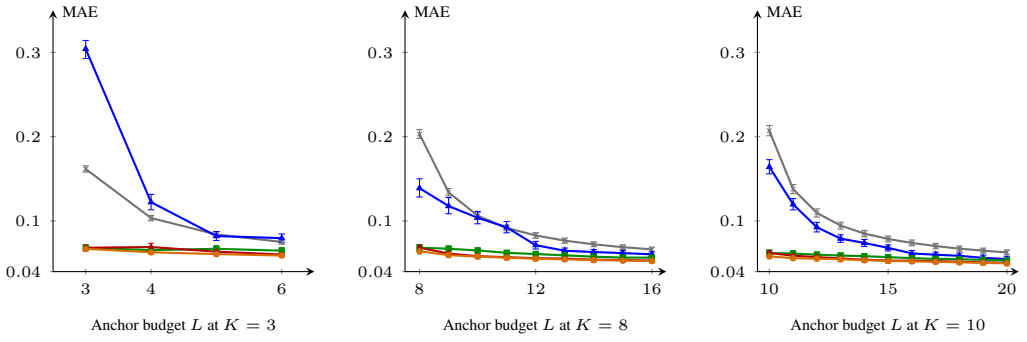
Reconstruction across rank and budget. Figure 9a reproduces the dimensional pattern in figure 1a: MAE is high for $L < K$, drops sharply around $L = K$, and generally improves more gradually thereafter. The variation for $L \geqslant K$ again shows that removing the dimensional obstruction does not by itself provide well-conditioned coordinates.

The importance of anchor choice. Figure 9b shows that E-optimal, D-optimal, and Q-DEIM clearly outperform random and leverage selection near the minimum budget, while the gaps narrow as L grows. The additional ranks in figure 10 show the same broad pattern, although the ordering of the first three rules varies. Hence, model-side reconstruction also depends most strongly on anchor choice near the minimum feasible budget.



(a) (b)

Figure 9: Model-side counterpart of figure 1 on P^{T}.



(a) (b) (c)

Figure 10: Model-side counterpart of figure 4 on P^{T}.

Clipping and ranking preservation. Table 7 reports the metrics in appendix C.3 at $K = 5$ and $L \in \{5, 10\}$. As in table 5, clipping mainly benefits leverage and random selection at $L = 5$, with smaller corrections at $L = 10$. The geometric rules also have higher ranking Spearman than

the naive baselines. Following appendix C.3, we exclude 27 undefined correlations among 66,000 random draws at $L = 5$ (0.041%); all other settings have none.

Table 7: Model-side counterpart of table 5 at $K = 5$.

L	Rule	Out-of-bounds (%)	Mean exceedance	Unclipped MAE	Clipped MAE	Ranking Spearman
5	E-optimal	0.45	0.1017	0.06783	0.06737	0.781
5	D-optimal	0.70	0.1050	0.07962	0.07889	0.765
5	Q-DEIM	0.70	0.1050	0.07962	0.07889	0.765
5	Leverage	32.30	8.8583	3.11798	0.25654	0.381
5	Random	17.21	4.1374	0.89743	0.18518	0.472
10	E-optimal	0.29	0.1098	0.05829	0.05798	0.806
10	D-optimal	0.16	0.0975	0.05689	0.05673	0.829
10	Q-DEIM	0.36	0.0853	0.06875	0.06844	0.802
10	Leverage	0.71	0.0790	0.07964	0.07908	0.784
10	Random	1.01	0.1526	0.07199	0.07045	0.762

D.2 Reconstruction Error Pathways

Target-subspace mismatch. Across 20 model-family-stratified splits and 800 held-out-model pairs, normalized projection residual and reconstruction MAE have a median within-split Spearman correlation of 0.850, and all split-level correlations are positive. Table 8 reports the individual values. Figure 11a therefore recovers the mismatch-error association when the target is a new model.

Table 8: Model-side counterpart of table 6 on P^{T}.

s	ρ_s	s	ρ_s	s	ρ_s	s	ρ_s	s	ρ_s
0	0.829	4	0.807	8	0.733	12	0.930	16	0.816
1	0.925	5	0.824	9	0.875	13	0.902	17	0.835
2	0.857	6	0.731	10	0.896	14	0.839	18	0.848
3	0.848	7	0.887	11	0.884	15	0.851	19	0.870

Historical-matrix perturbation. We follow the signed protocol in appendix C.6, using 88 historical and 44 held-out models. Figure 11b shows that larger perturbations generally worsen reconstruction for both signs. All 160 systems remain identifiable, and all 7,040 empirical-reference bound checks pass, with maximum, mean, and median E/B of 0.265, 0.090, and 0.081. Nevertheless, realized error decreases in 38.7% of successive comparisons despite increasing d_{Gr}. Figure 12a shows error reductions despite larger target residuals in 676 pairs, while figure 12b shows that nonzero target-specific projector action can also accompany lower error. These diagnostics highlight the role of perturbation direction and target alignment, rather than subspace-change magnitude alone.

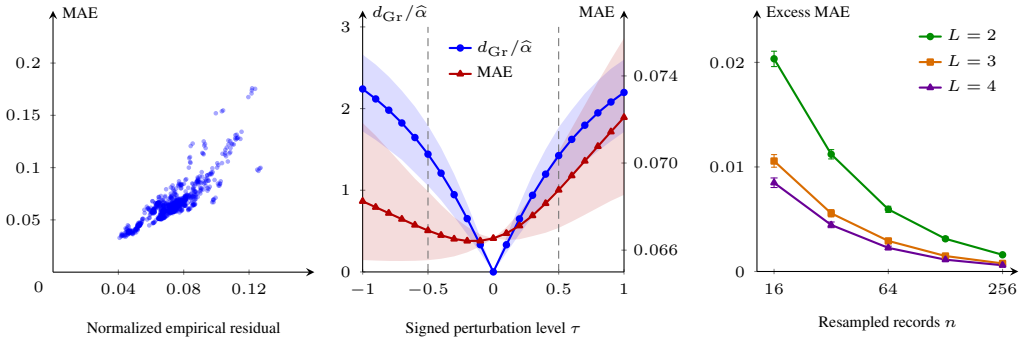
Anchor-evaluation noise. Figure 11c shows that finite-sample excess MAE relative to full-score reconstruction decreases for all L as n increases and is lower for larger anchor sets. Across all five sample sizes, the $L = 2$ curve remains clearly above the other two curves, suggesting that most of the finite-sample benefit comes from adding redundancy beyond $L = K = 2$. The anchor-noise norm follows the expected $n^{-1/2}$ scaling, and all 990,000 nonzero trials satisfy the inverse-gain envelope. Figure 13 shows the corresponding source norm and propagated reconstruction change. Together, these results recover the same three error pathways for model-side reconstruction.

D.3 Anchor Gain and Common Amplification

Observed conditioning and error. After the corresponding random-set centering, inverse gain and MAE have a median within-pair Spearman correlation of 0.640. Figure 14a places the deterministic rules in the same plane and shows the same tendency for poorer conditioning to accompany larger realized error, without making inverse gain a pointwise error predictor.

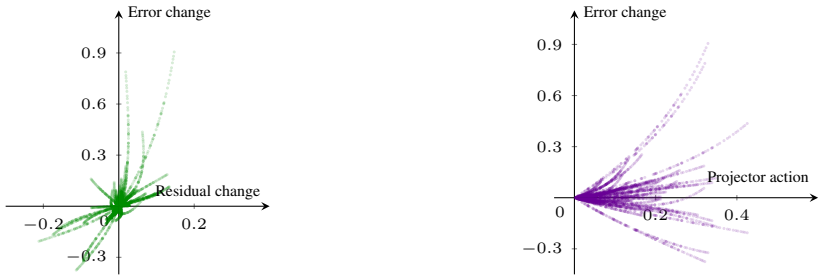
Controlled common amplification. Figure 14b groups anchor sets into 18 logarithmic inverse-gain bins and plots median normalized errors with 90th-percentile bands. All three curves increase with inverse gain while remaining below its envelope. Across anchor sets, inverse gain and the

per-set 90th-percentile amplification have Spearman correlations of 0.996, 0.996, and 0.998, respectively, and all 1.92 million realization-level checks satisfy the envelope. Thus, inverse gain plays the same common amplification role after transposition.



(a) (b) (c)

Figure 11: Model-side counterpart of figure 2 on P^{T}.



(a) (b)

Figure 12: Model-side counterpart of figure 5 on P^{T}.

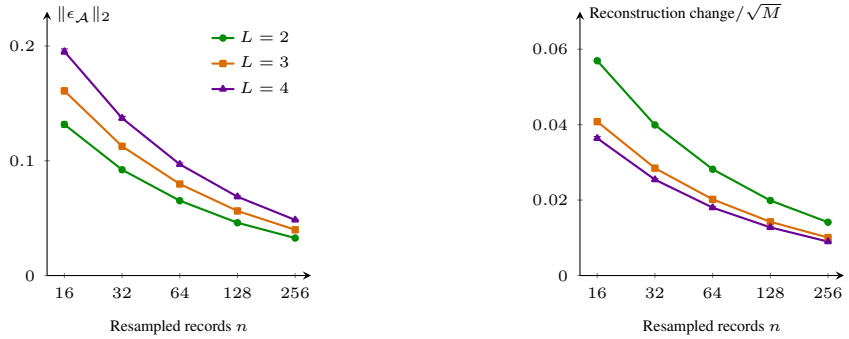
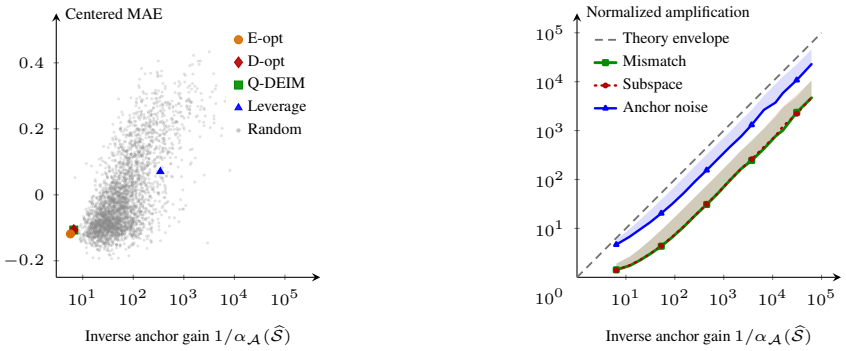


Figure 13: Model-side counterpart of figure 6 on P^{T}.

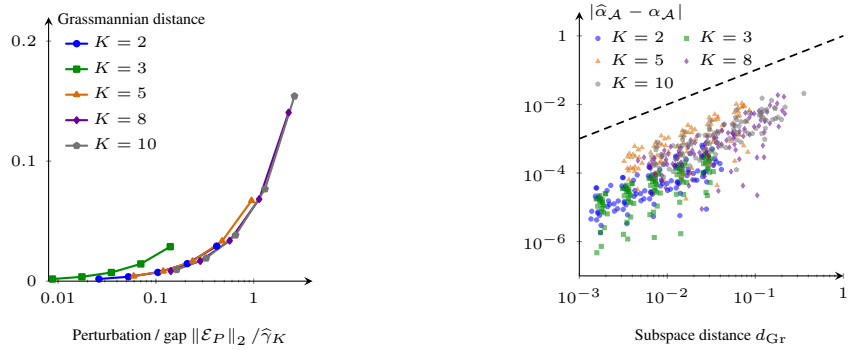


(a) (b)

Figure 14: Model-side counterpart of figure 3 on P^{T}.

D.4 Additional Model-Side Diagnostics

Rank, representation, and stability. Figure 15 shows that, under equal operator-norm perturbations, smaller empirical spectral gaps produce larger movements of the task-coordinate subspace, while every observed anchor-gain change remains below the corresponding Grassmannian distance. This reproduces the representation-stability tradeoff in appendix C.10 after transposition.



(a) (b)

Figure 15: Model-side counterparts of figure 7 on P^{T}.

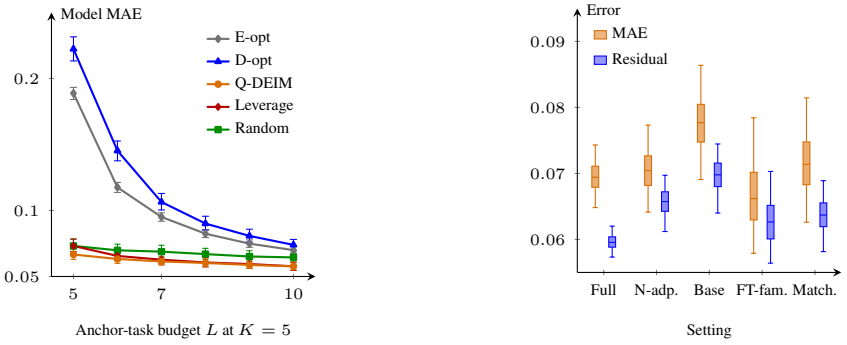
Reusable anchor tasks. Figure 16a shows that the structured rules outperform the naive baselines near the minimum budget, while the gaps narrow as L grows. Thus, anchor geometry remains important near the minimum budget when the system is reused across new models.

Historical-composition settings. Using the notation in appendix C.12, let $\mathrm{BaseFam}(f) \subseteq \mathcal{F}$ contain the models in the same base-model family as f. For an adapter f with fine-tuning task $\tau(f) \in \mathcal{T}$, define $\mathrm{TuneFam}(f) := \bigcup_{t \in \mathrm{Fam}(\tau(f))} \mathrm{Adapt}(t)$. For a target model f_u, let $\mathcal{J}_u$ be a uniformly sampled subset of $\mathcal{F} \setminus \{f_u\}$ with $|\mathcal{J}_u| = |\mathrm{BaseFam}(f_u)| - 1$. All settings retain $\mathcal{T}$ and use the target and historical-model subsets below:

Setting	Target models	Historical models
Full	$\mathcal{F}$	$\mathcal{F} \setminus \{f_u\}$
Non-adapter	$\mathcal{F} \setminus \mathrm{Adapter}$	$(\mathcal{F} \setminus \mathrm{Adapter}) \setminus \{f_u\}$
Base-family holdout	$\mathcal{F}$	$\mathcal{F} \setminus \mathrm{BaseFam}(f_u)$
Fine-tune-family holdout	$\mathrm{Adapter}$	$\mathcal{F} \setminus \mathrm{TuneFam}(f_u)$
Matched random	$\mathcal{F}$	$\mathcal{F} \setminus (\{f_u\} \cup \mathcal{J}_u)$

Matched random averages 10 draws of $\mathcal{J}_u$ per target. Bootstrap units are targets for Full and Non-adapter. We aggregate within 18 base-model families for Base-family holdout and Matched random, and within five task families for Fine-tune-family holdout.

Figure 16b shows that Non-adapter has mean MAE near Full but a higher residual. Their target sets differ, so this comparison is descriptive, although reconstruction remains viable without adapters. Base-family holdout exceeds size-matched random removal in MAE and residual, suggesting that related model history matters beyond library size and agreeing with the target-subspace mismatch mechanism in section 5.2. Fine-tune-family holdout shows no corresponding MAE increase, but evidence is limited to five families and its intervals overlap Full.



(a) (b)

Figure 16: Model-side counterpart of figure 8 on P^{T}.

E Task-Side Experiments on the Open LLM Leaderboard

We test whether the mechanisms in section 6 extend to an independent public performance matrix from the Open LLM Leaderboard v2 (Fourrier et al., 2024).

E.1 Dataset and Task-Side Protocol

Hugging Face's Open LLM Leaderboard v2 evaluates open LLMs on six benchmarks using the EleutherAI Language Model Evaluation Harness (Fourrier et al., 2024). We extract the released evaluation records from its public `open-llm-leaderboard/contents` dataset, retaining the normalized scores for IFEval, BBH, MATH Level 5, GPQA, MuSR, and MMLU-PRO. Excluding the reported average and raw-score variants and dividing percentage scores by 100 yields a $4{,}576 \times 6$ matrix in $[0, 1]$, with rows representing evaluation records rather than necessarily distinct models. We group records connected by a shared `fullname` or valid model commit SHA into entities and assign each entity wholly to one of five evaluation folds. For empirical reconstruction, we follow the task-side protocols in appendices C.2 and C.3, treating each benchmark in turn as the target and the other five as historical tasks. Table 9 reports the rapidly decaying singular spectrum.

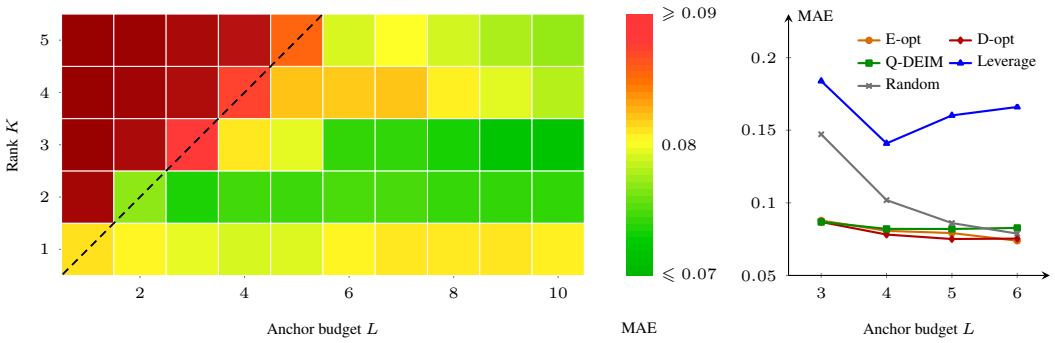
Table 9: Singular-spectrum diagnostics, with cumulative share reported as in table 4.

K	c_K (%)	ρ_K^{svd}	$\sigma_K - \sigma_{K+1}$
1	67.98	0.2469	36.9822
2	81.74	0.1501	3.2588
3	90.71	0.0783	3.3742
4	94.75	0.0532	0.7789
5	97.64	0.0337	0.3602

E.2 Reconstruction across Rank, Budget, and Anchor Selection

Reconstruction across rank and budget. The leave-one-benchmark-out matrix has rank at most five. Figure 17a reports E-optimal reconstruction for $K \leqslant 5$ and $1 \leqslant L \leqslant 10$, including $L < K$ as an underdetermined diagnostic. MAE is high throughout $L < K$ and drops sharply at $L = K$. Within this feasible region, the changes are not monotone at every intermediate budget, but redundant anchors generally help at the larger ranks.

The importance of anchor choice. At $K = 3$ and $K \leqslant L \leqslant 2K$, figure 17b shows similar MAE for E-optimal, D-optimal, and Q-DEIM, while leverage performs worst. Random selection outperforms leverage and trails the other three rules at small L, but approaches them as L grows.



(a) (b)

Figure 17: Leaderboard version of figure 1, with $K \leqslant 5$ and $1 \leqslant L \leqslant 10$ for rank-budget reconstruction and $K = 3$, $K \leqslant L \leqslant 2K$ for anchor selection.

Clipping and ranking preservation. Using the metrics in appendix C.3, table 10 reports results at $K = 3$ and $L \in \{3, 6\}$. We retain the plotted comparison's exclusion of one non-identifiable random set at $L = 3$ and average retained draw–fold results within targets, then across targets. Clipping changes MAE little for the three geometric rules but substantially benefits leverage and random selection, especially at $L = 3$. The geometric rules also have higher ranking Spearman than the naive baselines at both budgets; no ranking correlations are undefined.

Table 10: Leaderboard counterpart of table 5 at $K = 3$.

L	Rule	Out-of-bounds (%)	Mean exceedance	Unclipped MAE	Clipped MAE	Ranking Spearman
3	E-optimal	5.64	0.0150	0.08868	0.08783	0.797
3	D-optimal	4.86	0.0156	0.08756	0.08680	0.785
3	Q-DEIM	4.86	0.0156	0.08756	0.08680	0.785
3	Leverage	28.93	5.6580	1.82065	0.18388	0.488
3	Random	21.88	3.7803	0.97434	0.14718	0.514
6	E-optimal	3.23	0.0100	0.07423	0.07390	0.813
6	D-optimal	2.65	0.0079	0.07559	0.07538	0.803
6	Q-DEIM	5.20	0.0144	0.08346	0.08271	0.787
6	Leverage	50.27	0.3201	0.32696	0.16602	0.376
6	Random	6.18	0.1345	0.08712	0.07880	0.727

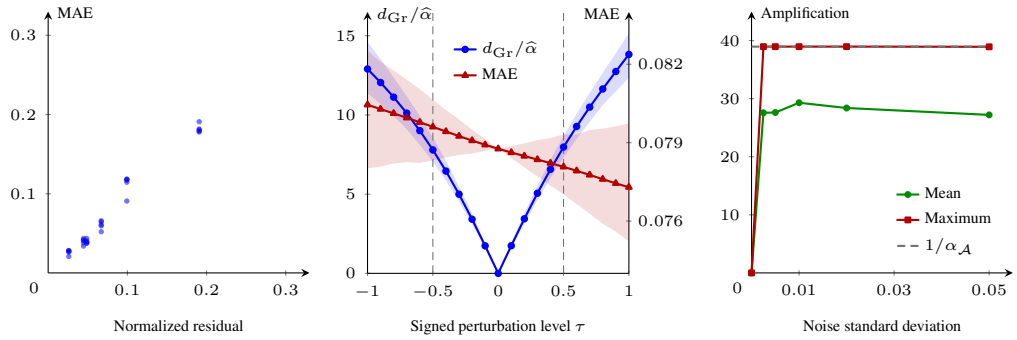
E.3 Reconstruction Error Pathways

At $K = L = 2$, we adapt the three pathway diagnostics in section 6.2 to the leaderboard matrix.

Target-subspace mismatch. Repeating the empirical residual diagnostic in appendix C.5, figure 18a shows that larger normalized residuals generally accompany higher reconstruction MAE. Because only six target benchmarks enter each coordinate fold, this is only a descriptive diagnostic.

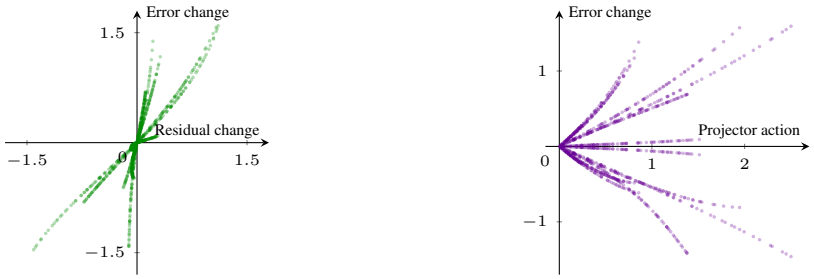
Historical-matrix perturbation. We follow the signed protocol in appendix C.6 for each held-out benchmark, fixing anchors and target scores. We average over the six benchmarks within each direction before applying the same bootstrap. Figure 18b shows worsening reconstruction on the negative side but improvement on the positive side as perturbations grow, despite increasing subspace displacement on both sides. All 960 perturbed systems remain identifiable and pass the empirical-reference bound checks. Figure 19a links error reductions to smaller target residuals, while figure 19b shows that nonzero projector action can accompany either improvement or degradation. Realized error thus depends on perturbation direction and target alignment, not only on subspace-change magnitude.

Anchor-evaluation noise. The finite-sample resampling experiment in appendix C.7 is unavailable since the leaderboard provides only aggregate scores. Instead, we draw 64 i.i.d. anchor-noise vectors $\epsilon_{\mathcal{A}} \sim \mathcal{N}(0, \sigma^2 I_L)$ for each σ and measure $\|D_{\mathcal{SA}}\epsilon_{\mathcal{A}}\|_2 \,/\, \|\epsilon_{\mathcal{A}}\|_2$. Figure 18c reports the mean and maximum sampled ratios, while the dashed line is the exact worst-case value $\|D_{\mathcal{SA}}\|_2 = 1/\alpha_{\mathcal{A}}$. This ratio is scale invariant, so variation across nonzero σ reflects the sampled directions rather than noise magnitude; the zero-noise ratio is plotted as zero by convention. Thus, this diagnostic tests the inverse-gain envelope rather than how benchmark sample size controls score error.



(a) (b) (c)

Figure 18: Leaderboard version of figure 2 at $K = L = 2$, except that (c) measures synthetic Gaussian anchor-score noise amplification rather than finite-sample excess MAE.



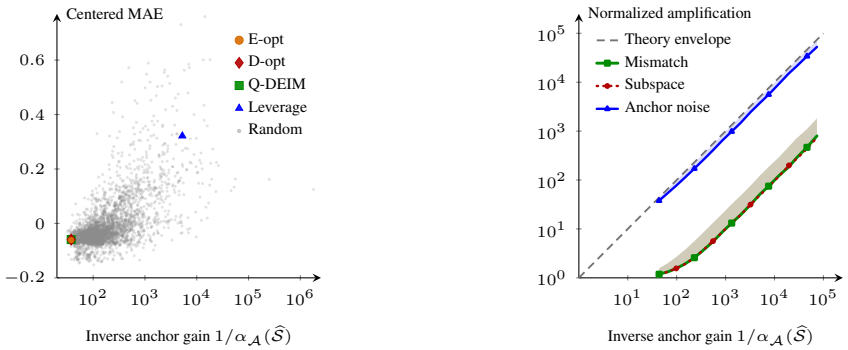
(a) (b)

Figure 19: Leaderboard counterpart of figure 5 at $K = L = 2$.

E.4 Anchor Gain and Common Amplification

Observed conditioning and error. We repeat the conditioning-error comparison of section 6.3 at $K = L = 2$. Figure 20a places E-optimal, D-optimal, and Q-DEIM near the low-inverse-gain, low-error end of the random cloud, while leverage has poorer average conditioning and error. Among random designs, inverse gain and MAE have a median within-pair Spearman correlation of 0.619.

Controlled common amplification. We apply the protocol of appendix C.9 to one historical benchmark split at $K = L = 2$, comparing 5,000 identifiable anchor sets. Figure 20b shows that the three normalized amplifications increase with inverse gain and that all sampled values stay below its theoretical envelope.



(a) (b)

Figure 20: Leaderboard version of figure 3 at $K = L = 2$.

E.5 Scope and Limitations

Six benchmarks cap the leave-one-out rank at five, precluding the broad rank study in appendix C.10. A reusable-anchor split leaves only two future benchmarks, too few for the block-holdout test in appendix C.11; we omit both extensions. Aggregate scores cannot support the per-example resampling in appendix C.7, and the dataset lacks adapter and task-family annotations for appendix C.12. Connected-entity folds reduce but do not eliminate dependence among records. Thus, these results support only the low-rank mechanisms tested above.