WHEN IS GRADIENT-BASED DATA SELECTION A PROPERTY OF THE PREDICTOR?

Anonymous authors
Paper under double-blind review

ABSTRACT

Gradient-based data selection is commonly interpreted as identifying examples that are informative for a given predictor, yet neural-network parameters are not unique representations of the predictor. We study predictor-level well-definedness of the resulting discrete selection under function-preserving reparameterizations. Across BADGE, CRAIG, and GradMatch, equivalent predictors replace 18.8–63.0% of a query batch while their logits remain unchanged to numerical precision; in end-to-end LoRA fine-tuning, factor-coordinate selection replaces 30.4–43.7% of MiniLM queries and, in the analytical factor-basis comparison, 70.3% on average for ModernBERT. Per-example gradients transform as covectors, so a fixed parameter-space metric generally generally changes the geometry seen by the selector. We characterize when Gram geometry is preserved exactly or up to scale, and give selector-specific constructions in which the resulting distortion crosses a discrete selection boundary. For Fisher-based selection, we further show that a coordinate-fixed prior can break invariance even when the Fisher terms transform covariantly. Finally, repeated acquisition reveals that query non-identifiability and downstream utility are distinct: acquisition trajectories can diverge strongly while terminal accuracy remains similar in the settings studied. These results identify parameterization as a source of data-selection non-identifiability that is separate from conventional questions of subset utility.

1 INTRODUCTION

Gradient-based data selection uses per-example gradients to decide which data should be labeled, retained, or used for training. BADGE, CRAIG, GradMatch, and BAIT convert gradient norms, distances, matching objectives, or Fisher information into data decisions (Ash et al., 2020; Mirzasoleiman et al., 2020; Killamsetty et al., 2021a; Ash et al., 2021). Although typically evaluated by subset utility, their outputs are naturally interpreted as properties of the predictor: given the same predictor and candidate pool, one expects the same examples to be judged informative.

Neural-network parameters, however, are not unique. Hidden units can be rescaled without changing the represented function, while LoRA factors admit invertible basis changes that preserve the merged update. Such transformations leave predictions unchanged but alter parameter-gradient coordinates. This raises our central question: *is a gradient-based data-selection decision well defined for the represented predictor, or can functionally equivalent parameterizations select different data?*

Gradient coordinate dependence alone does not answer this question. Invariant optimization studies how parameter updates transform (ichi Amari, 1998; Neyshabur et al., 2015), implementation invariance asks related questions for explanatory measurements (Sundararajan et al., 2017), and recent work studies LoRA factor symmetries (Yen et al., 2025; Castin et al., 2026). Data selection adds a discrete step: geometric distortion matters only when it crosses a selector decision boundary.

We find substantial changes in selected identities under function-preserving reparameterizations. Across MNIST, Fashion-MNIST, and CIFAR-10, the evaluated BADGE, CRAIG, and GradMatch implementations replace 18.8–63.0% of a query batch while the largest logit discrepancy is 1.26×10^{-13}. In end-to-end Transformer PEFT, analytical LoRA basis changes replace 30.4–43.7% of MiniLM queries and 70.3% on average for rank-8 ModernBERT, whereas basis-invariant or canonical controls return identical queries.

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.

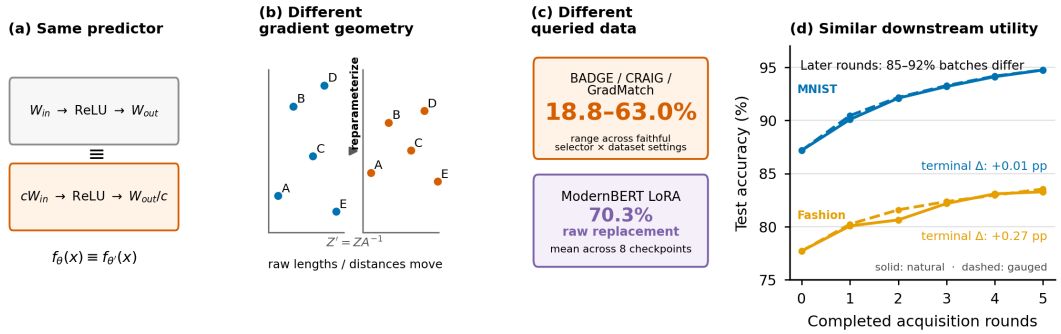

Figure 1: **Same predictor, different selected data, and potentially similar utility. (a)** Function-preserving reparameterizations leave the predictor unchanged. **(b)** They can alter the gradient geometry seen by a selector. **(c)** Raw query replacement ranges from 18.8–63.0% across BADGE, CRAIG, and GradMatch and reaches 70.3% for rank-8 ModernBERT LoRA. **(d)** Query divergence can persist across acquisition rounds while terminal accuracy remains similar.

We explain this dependence geometrically. Per-example gradients transform as covectors, so a fixed Euclidean metric need not preserve selector geometry. We characterize exact and projective Gram invariance, instance-visible distortion, and when such distortion can cross a selection boundary. Explicit constructions give first-choice reversals for BADGE, CRAIG, and blockwise positive OMP. For Fisher-based selection, Fisher terms transform covariantly but a coordinate-fixed prior need not: retaining BAIT's identity prior changes 36.2–42.6% of selected examples, whereas covariant controls preserve the query path.

Selection identifiability is also distinct from downstream utility. After five active-learning rounds, cumulative acquired sets differ by 47.9% on MNIST and 63.2% on Fashion-MNIST, while paired terminal accuracy differs by only $+0.01$ and $+0.27$ percentage points. Related Fisher, pretrained-feature, and Transformer comparisons likewise show no consistent negative accuracy direction.

Our contributions are:

- We **formulate selection identifiability as a well-definedness question** and demonstrate substantial selected-index changes under function-preserving reparameterizations.
- We **characterize when coordinate dependence changes a discrete selection decision** through covector transport, Gram invariance, instance-visible geometry, selector reversals, and Fisher-prior congruence.
- We **extend the phenomenon to end-to-end Transformer PEFT** and show that selection identifiability and downstream utility are empirically distinct in the settings studied.

2 When Is Data Selection a Property of the Predictor?

Gradient-based selectors are often interpreted as identifying examples that are informative for the represented predictor. Because neural-network parameters are non-unique, this interpretation raises a well-definedness question: should equivalent parameterizations induce the same data decision? We formulate this question relative to an explicit family of function-preserving transformations, rather than to all possible representations of the same function.

2.1 Function-Preserving Reparameterizations

Let f_θ denote the predictor and let a group G act on a specified parameter block such that $f_{g\cdot\theta}(x) = f_\theta(x)$ for all x and $g \in G$. The orbit $[\theta]_G = \{g \cdot \theta : g \in G\}$ therefore contains different parameter representatives of the same predictor. Examples used below are positive hidden-unit rescalings in homogeneous networks and invertible changes of basis in a fixed-rank LoRA factorization.

For gradient-based quantities, we require the group action to extend to a differentiable function-preserving reparameterization in a neighborhood of the parameter point, with model state trans-

ported; Section 5 states these assumptions precisely. Coordinate dependence is not undesirable when the estimand is coordinate specific. Our question concerns selectors interpreted as assigning importance to examples *for the represented predictor*, for which $[\theta]_G$ is the object.

2.2 Pathwise G-Identifiability

Let $S(\theta, \mathcal{D}, \xi)$ be the selected subset, where $\mathcal{D}$ is the candidate pool and ξ contains algorithmic randomness and deterministic side information, including tie breaking. We call the selector *pathwise G-identifiable* when

$$S(g \cdot \theta, \mathcal{D}, \xi) = S(\theta, \mathcal{D}, \xi) \qquad \forall g \in G, \tag{1}$$

with $\mathcal{D}$ and ξ fixed. For stochastic selectors, this asks whether the same realized path is preserved, not whether independent runs happen to agree or whether the marginal distribution over selected sets is unchanged. This pathwise question is operationally meaningful when one realized query triggers example-specific actions such as annotation purchase, human review, deletion, or provenance decisions; these downstream action costs motivate the definition but are not measured here. The definition is also group-relative: it makes no invariance claim outside the stated symmetry family.

2.3 Gradient-Based Selection as Geometric Decision Making

Let $z_i(\theta)$ denote the per-example row covector actually consumed by the selector. It need not contain gradients for every trainable parameter, but the consumed block must be specified and closed under the transformation under study. Stacking these rows as Z, a broad class of selectors can be written through the Gram geometry

$$K_M(\theta) = Z(\theta) M(\theta) Z(\theta)^\top, \qquad S(\theta, \mathcal{D}, \xi) = \Phi_{\mathcal{D}}(K_M(\theta), \xi), \tag{2}$$

where M is a symmetric covector bilinear form; $M = I$ gives ordinary Euclidean gradient geometry. Gradient-norm ranking uses its diagonal, BADGE and CRAIG use Euclidean norms or distances, and GradMatch uses gradient inner products blockwise, so for GradMatch Φ consumes $\{K_M^{(b)}\}_{b=1}^B$. Fisher-based selection additionally contains a prior or regularizer, treated separately in Section 5.4.

The selector-specific geometries, obstructions, and corresponding diagnostic controls are summarized in Appendix Table 1. Section 5 characterizes when the geometry is preserved and gives explicit selector-level decision reversals.

3 Auditing Selection Identifiability

We compare different parameter representatives of the same predictor while holding the candidate pool and all selector side information fixed. The evaluation separates three questions: whether the predictor is unchanged, whether the geometry consumed by the selector is unchanged, and whether any geometric change alters the discrete query.

3.1 Reparameterization Families

We use two function-preserving symmetries. For a positively homogeneous ReLU hidden layer, positive diagonal rescaling $W' = DW$, $b' = Db$, and $V' = VD^{-1}$ preserves the network function while changing the associated gradient coordinates. For a rank-r LoRA update $\Delta W = BA$, any $R \in GL(r)$ gives $B' = BR^{-1}$ and $A' = RA$, so $B'A' = BA$ while the internal factor basis changes. The positive homogeneity is the standard ReLU property (Nair & Hinton, 2010). The vision experiments use the former family; the PEFT experiments use the latter. Section 5 treats the more general Jacobian action of an invertible function-preserving reparameterization, and Appendix B.2 gives the experiment-specific transformations.

3.2 Invariant and Covariant Controls

The controls isolate coordinate dependence from predictor dependence. If $Z' = ZA^{-1}$ and the selector uses $K_M = ZMZ^\top$, transporting the bilinear form as $M' = AMA^\top$ gives $Z'M'Z'^\top = ZMZ^\top$; the corresponding necessity statement is proved in Section 5.2. We instantiate this principle with invariant or canonical gradient representations for the ReLU symmetry. For Fisher-based

selection, we compare a coordinate-fixed identity prior with transported or Fisher-derived covariant priors (Section 5.4). For LoRA, balanced coordinates and a basis-invariant merged representation remove the arbitrary factor basis. These LoRA controls diagnose basis dependence; the merged representation used in the Transformer experiments is not assumed to be an exact computational substitute for exact merged-gradient selection, whose fidelity is evaluated in Appendix G. Thus, when the raw selector changes but the corresponding invariant control does not, the difference is attributable to the representation used by the selector rather than to a changed predictor.

3.3 Evaluation Protocol

For equal-size query sets S, S' of size k, we report

$$\mathrm{Rep}(S, S') = 1 - \frac{|S \cap S'|}{k}, \tag{3}$$

so 0 denotes identical queries and 1 disjoint queries. For stochastic selectors, both parameterizations share the same random tape, candidate ordering, pseudo-label convention, class allocation, and tie rules, so the comparison measures the pathwise notion in Section 2.2. Reparameterizations are aggregated within each trained model or checkpoint, with uncertainty formed across independent models or checkpoints.

We verify predictor equivalence whenever parameters are transformed. We distinguish *analytical* reparameterizations, which transform selector-consumed quantities, from *physical* reparameterizations, which materialize transformed parameters and incur finite-precision effects (Appendix F).

4 Equivalent Predictors Can Select Different Data

4.1 Query Replacement Across Selection Paradigms

We test whether function-preserving reparameterizations alter the examples returned by three distinct gradient-based selectors: BADGE, CRAIG, and GradMatch. On MNIST, Fashion-MNIST, and CIFAR-10 (LeCun et al., 1998; Xiao et al., 2017; Krizhevsky, 2009), each setting uses five independently trained acquisition models, ten hidden-unit rescalings per model, a candidate pool of $1{,}000$, and query budget 50. All stochastic side information is coupled before and after reparameterization, so the comparison measures pathwise G-identifiability rather than independent-run variability.

Figure 2 shows substantial selected-index changes despite an unchanged predictor. Across selector–dataset pairs, raw query replacement ranges from 18.8% to 63.0%: BADGE replaces $51.0/62.5/63.0\%$, CRAIG $18.8/23.9/46.2\%$, and GradMatch $22.9/24.7/52.1\%$ on MNIST/Fashion-MNIST/CIFAR-10, respectively. Thus coordinate dependence reaches the final data decision rather than only an intermediate gradient statistic. The magnitude is method- and dataset-dependent; we do not interpret these values as a universal ordering of selector sensitivity. Replacement is already present at the mildest evaluated rescaling strength, with method- and dataset-dependent severity curves reported in Appendix C.1; the sweep is a robustness analysis over the specified transformations, not a population model of naturally occurring gauges.

4.2 What Changes and What Remains Fixed

The comparison changes the parameter representative, not the prediction problem: the candidate pool, budget, selector hyperparameters, pseudo-label convention, ordering, class allocation where applicable, and random tape are held fixed, and the maximum logit discrepancy is 1.26×10^{-13}. For CRAIG and GradMatch, fixed class allocation does not imply identical selected identities. Under the corresponding invariant or canonical representations, all three selector families yield zero query replacement across the evaluated datasets, localizing the discrepancy to the coordinate geometry consumed by the selector.

These results establish an empirical failure of pathwise G-identifiability in the evaluated model–data instances: equivalent predictors *can* induce substantially different queries. They do not imply that every reparameterization changes a query; the realized effect depends on the selector, data, and position within the parameter-equivalence class. Section 5 characterizes when coordinate changes alter the visible geometry and when such changes can cross a selection boundary.

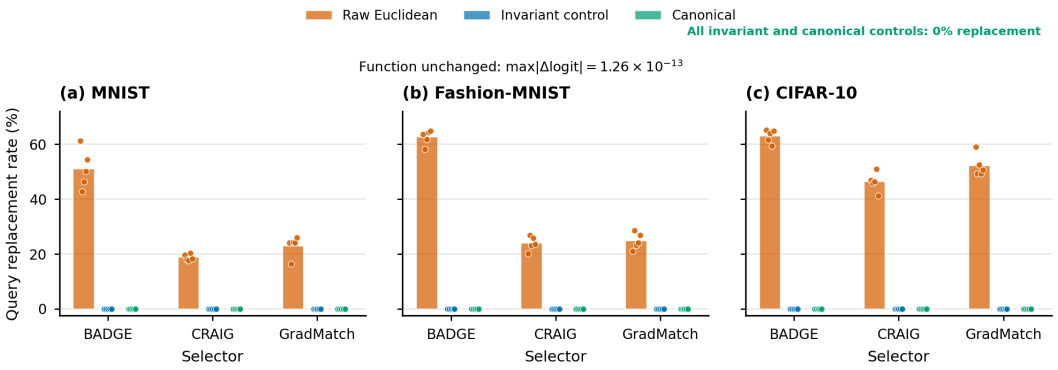

Figure 2: **Equivalent predictors can select different data.** Query replacement under function-preserving reparameterizations for BADGE, CRAIG, and GradMatch on MNIST, Fashion-MNIST, and CIFAR-10. Raw replacement ranges from 18.8–63.0%, while invariant/canonical controls return identical queries; the maximum logit discrepancy is 1.26×10^{-13}.

5 Why Does Selection Depend on Parameter Coordinates?

Equivalent predictors can return different queries only through two steps: reparameterization must change the selector-visible geometry, and that change must cross a decision boundary. Neither nonorthogonality nor non-conformality alone therefore implies query turnover.

Let $\varphi : U \to U'$ be an invertible differentiable reparameterization of the selector-consumed parameters, with Jacobian $A = D\varphi(\theta)$, such that

$$f_{\varphi(\vartheta)} = f_\vartheta \qquad \forall \vartheta \in U, \tag{4}$$

with loss-relevant state transported. Nonsmooth points use a fixed subgradient convention.

5.1 Covector Transport Under Reparameterization

Per-example gradients transform as covectors.

Lemma 1 (Covector transport). *If z_i is the row covector consumed by the selector, then*

$$z_i' = z_i A^{-1}, \qquad Z' = ZA^{-1}. \tag{5}$$

Thus coordinate change matters through the geometry used to compare these covectors.

5.2 Universal and Instance-Visible Gram Geometry

For $K_M = ZMZ^\top$, exact universal invariance is characterized by metric transport.

Theorem 1 (Universal Gram congruence). *For invertible A and symmetric M, M',*

$$Z'M'Z'^\top = ZMZ^\top \quad \forall Z \qquad \Longleftrightarrow \qquad M' = AMA^\top. \tag{6}$$

The universal quantifier is essential: congruence is necessary for every possible Z, not for one realized dataset. Conversely, Gram equality is sufficient but not necessary for identical selection, since a changed Gram can remain within the same decision region. Proofs are in Appendix A.2.

For Euclidean metric, universal invariance requires $AA^\top = I$, while projective invariance satisfies

$$K' = \alpha K \quad \forall Z \qquad \Longleftrightarrow \qquad AA^\top = \alpha^{-1} I. \tag{7}$$

Hence conformal transformations preserve scale-homogeneous primitives such as max-norm/D^2 BADGE, squared-Euclidean CRAIG, and unregularized OMP, up to rounding and ties.

Global anisotropy also need not be visible to the realized data. Let $C = A^{-1}A^{-\top}$ and $\mathcal{W} = \mathrm{span}\{z_i\}$. For fixed Z,

$$K' = \alpha K \qquad \Longleftrightarrow \qquad u^\top C v = \alpha\, u^\top v \quad \forall u, v \in \mathcal{W}. \tag{8}$$

Thus even non-projective geometry changes alter the query only when they cross a selector decision boundary.

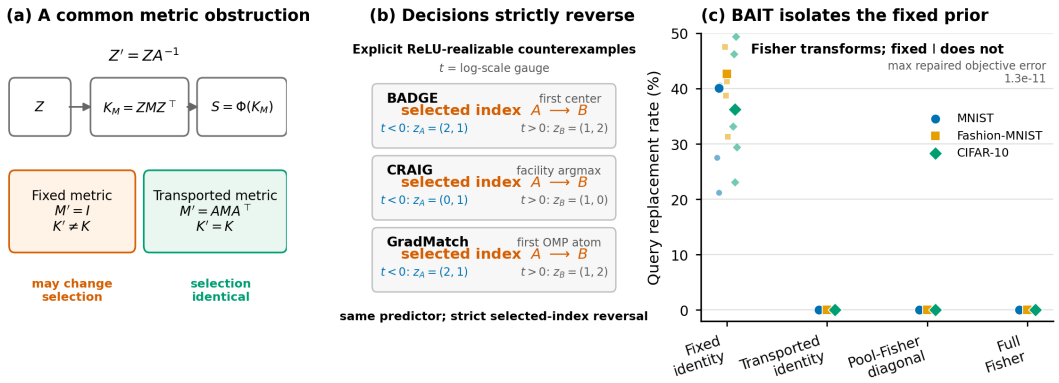

Figure 3: **Parameter coordinates affect selection through selector-visible geometry.** **(a)** Covectors transform as $Z' = ZA^{-1}$; congruent transport preserves Gram geometry, while visibility depends on the realized gradient subspace. **(b)** ReLU-realizable constructions yield first-choice reversals for BADGE, CRAIG, and blockwise positive OMP. **(c)** A fixed BAIT identity prior changes 36.2–42.6% of queries, while covariant controls preserve the path with objective discrepancy below 1.3×10^{-11}.

5.3 Selector-Level Decision Reversals

For a one-unit ReLU rescaling, any fixed affine Gram comparison has the form

$$\Delta(t) = ae^{-2t} + be^{2t} + c, \tag{9}$$

so a sign change reverses that comparison.

Proposition 1 (Selector-level reversals)**.** *There exist ReLU-realizable, function-preserving rescalings that change the max-norm first center of BADGE, the first Euclidean facility-location choice of CRAIG, and the first positive-OMP atom in a GradMatch block.*

These are first-choice existence results, not universal turnover guarantees. Explicit constructions appear in Appendix A.8; larger-budget changes are demonstrated in Figure 2.

5.4 Fisher Selection and the Role of the Prior

Fisher selection exposes a related obstruction. Let $H_S = F_L + \sum_{x \in S} I_x$ and

$$J_R(S; F_U) = \operatorname{tr}\left[(R + H_S)^{-1} F_U\right]. \tag{10}$$

Under an invertible coordinate map T, $F'_U = TF_UT^\top$ and $H'_S = TH_ST^\top$; the prior must transform covariantly as well.

Theorem 2 (Universal-probe criterion for fixed priors)**.** *For fixed $R, R' \succ 0$, $H \succeq 0$, and invertible T,*

$$\operatorname{tr}\left[(R' + THT^\top)^{-1} TGT^\top\right] = \operatorname{tr}\left[(R + H)^{-1} G\right] \quad \forall G \succ 0 \iff R' = TRT^\top. \tag{11}$$

This necessity statement concerns universal-probe objective equality, not equality for one realized batch. More generally, a data-dependent regularizer is sufficient when it transforms covariantly. Full-Fisher constructions satisfy this under arbitrary invertible maps, while diagonal-Fisher constructions do so only for the corresponding monomial transformations; a fixed identity prior is universally compatible only with orthogonal maps.

Figure 3c confirms the mechanism empirically. Across MNIST, Fashion-MNIST, and CIFAR-10, the fixed identity prior changes 40.1%, 42.6%, and 36.2% of selected examples, whereas transported or Fisher-derived covariant controls return identical queries, with objective discrepancy below 1.3×10^{-11}.

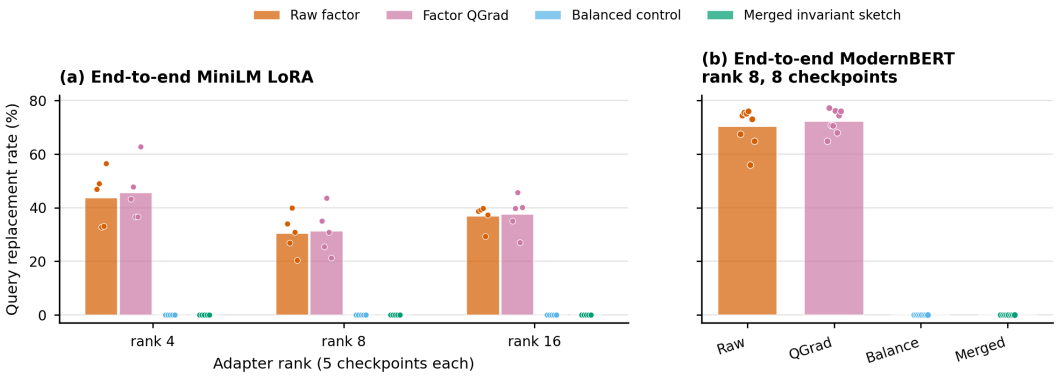

Figure 4: **Selection non-identifiability appears in the evaluated end-to-end Transformer encoder PEFT settings.** Under analytical function-preserving LoRA basis changes on SNLI, raw and factor-QGrad selection change substantially in MiniLM and ModernBERT, while balanced and merged analytical diagnostic controls remain invariant. Transformations are aggregated within each independent model instance.

6 From Classical Selectors to Modern PEFT

We next test whether selection non-identifiability appears beyond the classical settings above, whether query differences remain through repeated acquisition, and whether large query turnover implies downstream utility loss.

6.1 End-to-End Transformer PEFT

LoRA (Hu et al., 2022) provides a distinct symmetry: for any $R \in GL(r)$, $BA = (BR^{-1})(RA)$, so the represented update is unchanged while the factor basis—and hence factor-coordinate gradients—changes. With matrix gradients defined by the Frobenius pairing $d\ell = \langle G_B, dB\rangle_F + \langle G_A, dA\rangle_F$, the corrected analytical transport law is

$$G_{B'} = G_B R^\top, \qquad G_{A'} = R^{-\top} G_A, \qquad B' = BR^{-1},\ A' = RA. \tag{12}$$

Appendix B.3 derives the row/column convention; all Transformer results use this transport law.

We evaluate end-to-end LoRA fine-tuning on SNLI (Bowman et al., 2015). MiniLM-L12-H384 (Wang et al., 2020b) uses ranks $r \in \{4, 8, 16\}$ with five independently trained model instances per rank, 500 candidates, and budget 50. Under nonorthogonal basis changes, raw factor-coordinate selection replaces 43.7%, 30.4%, and 36.9% of the query at ranks 4, 8, and 16, respectively; elementwise factor-QGrad replaces 45.5%, 31.3%, and 37.6%. Balanced and merged analytical controls return identical queries under the evaluated transformations.

In the analytical factor-basis comparison, the same effect appears in rank-8 ModernBERT-base (Warner et al., 2024) with LoRA in 44 attention and MLP projections across eight independently trained model instances: raw factor-coordinate selection replaces 70.3% of queries on average and factor-QGrad 72.3%, while the analytical diagnostic controls remain invariant. A separate physical balanced-SVD writeback changes four of five held-out raw queries, with replacement 76/24/48/0/66% (mean 42.8%). This checkpoint-dependent mean falls below the predeclared 50% gate and is distinct from the 70.3% analytical rate. Thus invariance to coordinatewise rescaling does not imply invariance to dense basis mixing. The claim is limited to the encoder PEFT settings studied here; we do not extrapolate to substantially larger generative models. The merged Transformer control is basis-invariant but only an approximation to exact merged-gradient selection; its fidelity–cost tradeoff is quantified in Appendix G.

6.2 Selection Instability Across Acquisition Rounds

A one-shot difference could disappear after retraining, so we run five rounds of BADGE acquisition on MNIST and Fashion-MNIST with five independent acquisition models per dataset and 250 ex-

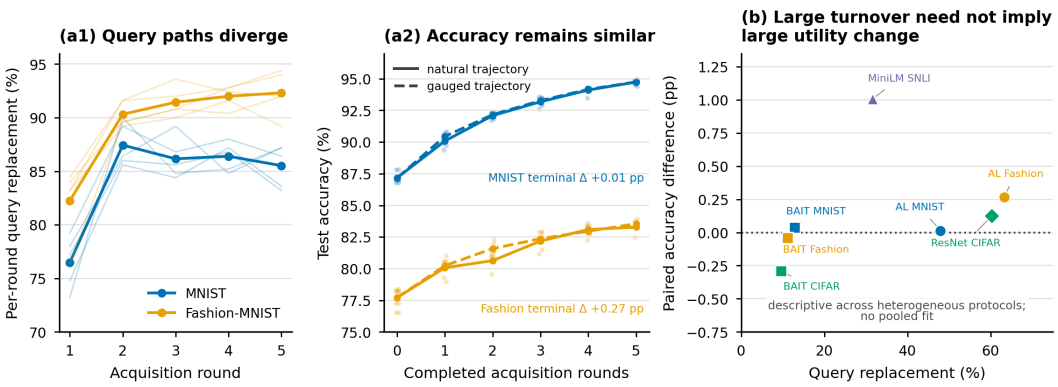

Figure 5: **Large query turnover need not imply large utility change. (a)** Five-round BADGE acquisition shows large batch and cumulative replacement on MNIST and Fashion-MNIST, while terminal accuracy changes remain small. **(b)** Across heterogeneous selection settings, query turnover shows no consistent negative downstream-accuracy direction; points are paired within protocol and shown descriptively without pooled regression.

amples acquired per round. Candidate ordering and selector randomness are coupled between the original and reparameterized trajectories.

The first-round query replacement is 76.5% on MNIST and 82.2% on Fashion-MNIST; later-round batches differ by 85–92%. After five rounds, cumulative acquired-set replacement remains 47.9% and 63.2%, respectively (Figure 5a). Thus the selection difference persists through adaptive acquisition rather than vanishing after one query, although this says nothing about predictive utility.

6.3 Large Turnover Need Not Imply Large Utility Change

Despite these divergent acquisition histories, paired terminal accuracy differs by only $+0.01$ percentage points on MNIST ($[-0.20, +0.18]$) and $+0.27$ on Fashion-MNIST ($[-0.08, +0.68]$). Figure 5b places these results alongside paired Fisher-based, pretrained-feature, and MiniLM continuation experiments. Because these settings differ in architecture, selector, and training protocol, we treat the panel descriptively rather than fit a pooled relationship. Across them, large query replacement shows no consistent negative downstream-accuracy direction.

Selection identifiability and downstream utility are distinct estimands: the former asks whether a predictor determines a unique query under the stated symmetry, whereas the latter asks whether different queries produce different trained models under a budget and training procedure. High turnover alone is not evidence of semantic severity or accuracy loss; randomness and turnover-matched baselines are reported in Appendix D, with full downstream results in Appendix H.

7 Related Work

Our work connects data selection, invariant geometry, training-data attribution, and low-rank parameter symmetries, but studies a different object: whether the *selected data identities* are well defined under functionally equivalent parameterizations.

Gradient-based data selection. Active learning and subset selection use geometric, information-theoretic, representation-based, and learned criteria (Settles, 2009; Sener & Savarese, 2018; Kirsch et al., 2019; Sinha et al., 2019; Yoo & Kweon, 2019). BADGE, CRAIG, GradMatch, and BAIT specifically use per-example gradients or Fisher information (Ash et al., 2020; Mirzasoleiman et al., 2020; Killamsetty et al., 2021a; Ash et al., 2021), while related methods use bilevel selection, gradient-based importance, forgetting, targeted instruction selection, importance resampling, or mixture optimization (Killamsetty et al., 2021b; Paul et al., 2021; Toneva et al., 2019; Xia et al., 2024; Xie et al., 2023b;a). These works primarily study subset utility; we ask whether a fixed criterion returns the same examples for the same represented predictor.

Invariant geometry and measurement. Parameterization dependence of gradients and Euclidean parameter-space geometry motivates natural gradient, K-FAC, Path-SGD, and symmetry-aware optimization (ichi Amari, 1998; Martens & Grosse, 2015; Neyshabur et al., 2015; Badrinarayanan et al., 2015; Dinh et al., 2017). Related studies implementation invariance, saliency reliability, and gauge dependence of representation measurements (Sundararajan et al., 2017; Kindermans et al., 2017; Adebayo et al., 2018; Cain, 2026; Karasan, 2026). Our focus is not coordinate dependence, but when it reaches a *discrete selection decision*, from geometric distortion to selected-index reversals.

Gradient sensitivity and training-data attribution. Weight-times-gradient sensitivity appears in pruning (Lee et al., 2019; Wang et al., 2020a), while influence functions, TracIn, representer points, Data Shapley, datamodels, TRAK, DataInf, and LESS quantify training-example effects (Koh & Liang, 2017; Pruthi et al., 2020; Yeh et al., 2018; Ghorbani & Zou, 2019; Ilyas et al., 2022; Park et al., 2023; Kwon et al., 2024; Xia et al., 2024). Recent work further studies parameter weighting and the semantics captured by gradient similarity (Li et al., 2026; Kim et al., 2026). We instead study invariance of the resulting *selected identity* under functional equivalence.

LoRA symmetry and low-rank geometry. LoRA uses nonunique low-rank factors (Hu et al., 2022), connecting naturally to fixed-rank quotient and matrix-manifold geometry (Mishra et al., 2014; Vandereycken, 2013; Absil et al., 2008). Recent work exploits these symmetries through invariant optimization, balanced factorization, and equivariant processing (Yen et al., 2025; Castin et al., 2026; Putterman et al., 2025). We do not claim the $GL(r)$ symmetry itself as new; we ask whether it changes *which examples are selected* in end-to-end Transformer PEFT.

Data redundancy and utility. Different subsets can induce similar training behavior when examples or aggregate gradients are substitutable (Mirzasoleiman et al., 2020; Killamsetty et al., 2021a; Toneva et al., 2019; Paul et al., 2021; Ilyas et al., 2022). This motivates treating selection identifiability and downstream utility as separate estimands.

8 Discussion and Limitations

Our G-identifiability notion applies when selected examples are interpreted as properties of the predictor; coordinate dependence may be appropriate for parameterization-specific estimands.

Identifiability and reproducibility. Geometric non-invariance need not change a query: the realized distortion must be selector-visible and cross a decision boundary. Our reversal results are existence statements, while empirical turnover is instance dependent. For stochastic selectors, the main claim is pathwise under a shared random tape, not equality of the full selection distribution.

Invariance, approximation, and utility. Invariance does not imply informativeness, numerical reproducibility, exact approximation, or downstream utility. Singular strata, finite-precision checkpoint rewriting, and approximate merged representations introduce separate boundaries; likewise, large query turnover shows no consistent negative accuracy effect in the settings studied.

Scope. Our claims concern positive hidden-unit rescaling and fixed-rank LoRA basis changes in the evaluated selector and encoder settings. They do not establish prevalence in larger generative models, full-parameter fine-tuning, or unmeasured fairness, privacy, or economic outcomes.

9 Conclusion

Gradient-based data selection raises a prior question to subset quality: whether the selected data are determined by the represented predictor. Under the function-preserving symmetries studied here, equivalent parameterizations can induce different selector-visible geometry and sample identities when the resulting distortion crosses a decision boundary. Fisher-based selection shows that coordinate dependence can also enter through a fixed prior despite covariant Fisher terms. Empirically, large query divergence need not produce consistent downstream-accuracy loss, separating selection identifiability from utility. A natural next question is which data-selection estimands should be defined on parameter-equivalence classes.

AI USE STATEMENT

We used generative AI tools to assist with literature retrieval and reference organization, manuscript editing, organization of experimental results, and research and mathematical reasoning. All AI-assisted text, citations, code, results, and proofs were reviewed and verified by the authors, who made all final methodological and scientific decisions. We take responsibility for the complete content of this work.

ETHICS STATEMENT

This work studies statistical properties of language-model responses and does not involve human participants or the collection of personal data. The experiments use existing benchmark tasks and pretrained language models. The certificates studied here characterize response distributions under registered representation groups; they should not be interpreted as guarantees of factual correctness, fairness, or safety.

REPRODUCIBILITY STATEMENT

We provide an anonymized supplementary archive containing the implementation, experiment configurations, definitions, random seeds, result artifacts used in the reported analyses, and scripts for reproducing the tables and figures. The supplement also includes environment information and integrity manifests for the experimental artifacts. Experimental settings and evaluation metrics are specified in the main paper and Appendix, and complete proofs of the theoretical results are provided in the Appendix.

REFERENCES

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008. URL https://www.jstor.org/stable/j.ctt7smmk.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html.

David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027–1035, 2007. URL https://research.google/pubs/k-means-the-advantages-of-careful-seeding/.

Jordan Ash, Surbhi Goel, Akshay Krishnamurthy, and Sham Kakade. Gone fishing: Neural active learning with fisher embeddings. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/4afe044911ed2c247005912512ace23b-Abstract.html.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*, 2020. URL https://mlanthology.org/iclr/2020/ash2020iclr-deep/.

Vijay Badrinarayanan, Bamdev Mishra, and Roberto Cipolla. Symmetry-invariant optimization in deep networks, 2015. URL https://arxiv.org/abs/1511.01754.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642. Association for Computational Linguistics, 2015. URL https://aclanthology.org/D15-1075/.

Jericho Cain. Gauge freedom and metric dependence in neural representation spaces, 2026. URL https://arxiv.org/abs/2603.06774.

Valérie Castin, Kimia Nadjahi, Pierre Ablin, and Gabriel Peyré. Balanced lora: Removing parameter invariance to accelerate convergence. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.31484.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1019–1028, 2017. URL https://proceedings.mlr.press/v70/dinh17b.html.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 2019. URL https://proceedings.mlr.press/v97/ghorbani19c.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Shun ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998. doi: 10.1162/089976698300017746.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Predicting predictions from training data. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 2022. URL https://proceedings.mlr.press/v162/ilyas22a.html.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 448–456, 2015. URL https://proceedings.mlr.press/v37/ioffe15.html.

Abdullah Karasan. Representation measurements under function-preserving reparameterizations, 2026. URL https://arxiv.org/abs/2608.27020.

Krishnateja Killamsetty, Durga S, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5464–5474, 2021a. URL https://proceedings.mlr.press/v139/killamsetty21a.html.

Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. Glister: Generalization based data subset selection for efficient and robust learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9):8110–8118, 2021b. doi: 10.1609/aaai.v35i9.16988.

Sunwoo Kim, Seokwon Jung, Sohyung Kim, Seong Joon Oh, and Alice Oh. Form over content in gradient-based data attribution methods, 2026. URL https://arxiv.org/abs/2609.19589.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un)reliability of saliency methods, 2017. URL https://arxiv.org/abs/1711.00867.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/95323660ed2124450caaac2c46b5ed90-Abstract.html.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894, 2017. URL https://proceedings.mlr.press/v70/koh17a.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Yongchan Kwon, Eric Wu, Kevin Wu, and James Y. Zou. Datainf: Efficiently estimating data influence in lora-tuned llms and diffusion models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5e84a0f233611a1dc8fb794dc52415a3-Abstract-Conference.html.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Namhoon Lee, Thalaiyasingam Ajanthan, and Philip H. S. Torr. Snip: Single-shot network pruning based on connection sensitivity. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1VZqjAcYX.

Shuangqi Li, Hieu Le, Jingyi Xu, and Mathieu Salzmann. Learning to weight parameters for training data attribution. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/72fb9ab442fc60b7ae5d53bf6b478273-Abstract-Conference.html.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2408–2417, 2015. URL https://proceedings.mlr.press/v37/martens15.html.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6950–6960, 2020. URL https://proceedings.mlr.press/v119/mirzasoleiman20a.html.

Bamdev Mishra, Gilles Meyer, Silvère Bonnabel, and Rodolphe Sepulchre. Fixed-rank matrix factorizations and riemannian low-rank optimization. *Computational Statistics*, 29(3):591–621, 2014. doi: 10.1007/s00180-013-0464-z.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*, 2010. URL https://icml.cc/2010/papers/432.pdf.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14:265–294, 1978. doi: 10.1007/BF01588971.

Behnam Neyshabur, Ruslan Salakhutdinov, and Nati Srebro. Path-sgd: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/eaa32c96f620053cf442ad32258076b9-Abstract.html.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. Trak: Attributing model behavior at scale. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27074–27113, 2023. URL https://proceedings.mlr.press/v202/park23c.html.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2107.07075.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/e6385d39ec9394f2f3a354d9d2b88eec-Abstract.html.

Theo Putterman, Derek Lim, Yoav Gelberg, Michael M. Bronstein, Stefanie Jegelka, and Haggai Maron. Gl equivariant metanetworks for learning on low rank weight spaces. In *Learning on Graphs Conference*, 2025. URL https://openreview.net/forum?id=fHRRYBZAmS.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1aIuk-RW.

Burr Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison, 2009. URL https://burrsettles.com/publications.

Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5972–5981, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Sinha_Variational_Adversarial_Active_Learning_ICCV_2019_paper.html.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJlxm30cKm.

Joel A. Tropp and Anna C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007. doi: 10.1109/TIT.2007.909108.

Bart Vandereycken. Low-rank matrix completion by riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013. doi: 10.1137/110845768.

Chaoqi Wang, Guodong Zhang, and Roger Grosse. Picking winning tickets before training by preserving gradient flow. In *International Conference on Learning Representations*, 2020a. URL https://iclr.cc/virtual_2020/poster_SkgsACVKPH.html.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, 2020b. URL https://proceedings.neurips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024. URL https://arxiv.org/abs/2412.13663.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: Selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54104–54132, 2024. URL https://proceedings.mlr.press/v235/xia24c.html.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. URL https://arxiv.org/abs/1708.07747.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. In *Advances in Neural Information Processing Systems*, volume 36, 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/dcba6be91359358c2355cd920da3fcbd-Paper-Conference.pdf.

Table 1: **Gradient-based selection through a common geometric lens.** For each selector, we summarize its decision geometry, coordinate-dependent obstruction, and invariant or covariant control. Reversal entries denote existence results, not universal query changes.

Selector / primitive	Decision geometry	Potential obstruction	Diagnostic control
Gradient-norm ranking	$z_i^\top z_i$	fixed Euclidean metric	transported metric / $w \odot g$
BADGE (max-norm + D^2)	gradient Gram / distances	fixed Euclidean metric	invariant Gram
CRAIG (facility location)	pairwise gradient distances	fixed Euclidean metric	invariant Gram
GradMatch (positive OMP)	blockwise gradient inner products	fixed Euclidean metric	invariant block Grams
BAIT	Fisher + prior	non-covariant prior	covariant prior
LoRA factor selection	factor-coordinate Gram	factor-basis freedom	merged / balanced control

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. Data selection for language models via importance resampling. In *Advances in Neural Information Processing Systems*, volume 36, 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6b9aa8f418bde2840d5f4ab7a02f663b-Abstract-Conference.html.

Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K. Ravikumar. Representer point selection for explaining deep neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/8a7129b8f3edd95b7d969dfc2c8e9d9d-Abstract.html.

Jui-Nan Yen, Si Si, Zhao Meng, Felix Yu, Venkata Sai Surya Subramanyam Duvvuri, Inderjit Dhillon, Cho-Jui Hsieh, and Sanjiv Kumar. Lora done rite: Robust invariant transformation equilibration for lora optimization. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/bcbc0f660d2dde42f9d1d0ecb14a6f9a-Abstract-Conference.html.

Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 93–102, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Yoo_Learning_Loss_for_Active_Learning_CVPR_2019_paper.html.

A Theory Proofs and Counterexamples

This appendix provides the proofs and boundary cases underlying Section 5. Throughout, all invariance statements are exact mathematical statements unless finite precision is explicitly discussed. The object Z always denotes the per-example covectors actually consumed by the selector; it need not contain gradients with respect to every trainable parameter. We assume that the consumed parameter block is closed under the reparameterization under consideration.

Our results distinguish three levels. First, a reparameterization acts on the covectors and hence on their induced geometry. Second, the realized data may or may not see a non-projective distortion of that geometry. Third, even a visible distortion changes a discrete selection only if it crosses a decision boundary. The counterexamples in Section A.10 show why none of these levels can in general be collapsed into the next.

A.1 Covector Transport

We first give the full argument for Lemma 1. The neighborhood assumption in the lemma is essential: equality of the predictors at two isolated parameter points does not by itself determine how their derivatives are related.

Let

$$\varphi : U \to U'$$

be a local diffeomorphism between parameter neighborhoods, and suppose that the transported model and all state relevant to the per-example loss satisfy

$$\ell_i(\varphi(\vartheta)) = \ell_i(\vartheta) \qquad \text{for every } \vartheta \in U.$$

This condition includes the requirement that no parameter-dependent term in the audited loss is left untransported. At nonsmooth points, the statement requires a fixed subgradient convention; below we work at a differentiability point.

Let

$$A = D\varphi(\theta).$$

Writing $z_i(\theta)$ as a row covector, differentiation of the identity above gives

$$z_i(\varphi(\theta))\, A = z_i(\theta).$$

Since A is invertible,

$$z_i(\varphi(\theta)) = z_i(\theta)A^{-1}.$$

Stacking the per-example covectors row-wise yields

$$Z(\varphi(\theta)) = Z(\theta)A^{-1},$$

which proves Lemma 1.

The same argument applies to a selector-consumed parameter block provided the group action closes on that block. If a reparameterization mixes consumed and unconsumed coordinates, the restricted covectors cannot in general be transported independently of the omitted coordinates. Likewise, model state that enters the loss—for example, running statistics of a stateful normalization layer—must be regarded as part of the transported model description.

A.2 Universal Gram Congruence

We next prove Theorem 1. Recall that

$$K_M(Z) = ZMZ^\top$$

for a symmetric covector bilinear form M, and that under the reparameterization

$$Z' = ZA^{-1}.$$

Proof of Theorem 1. Suppose first that

$$M' = AMA^\top.$$

Then

$$\begin{aligned} Z'M'Z'^\top &= ZA^{-1}M'A^{-\top}Z^\top \\ &= ZA^{-1}(AMA^\top)A^{-\top}Z^\top \\ &= ZMZ^\top, \end{aligned}$$

so the congruence rule is sufficient.

For necessity, suppose instead that

$$ZA^{-1}M'A^{-\top}Z^\top = ZMZ^\top \qquad \text{for every } Z.$$

Define the symmetric matrix

$$Q = A^{-1}M'A^{-\top} - M.$$

The universal equality is equivalent to

$$ZQZ^\top = 0 \qquad \text{for every } Z.$$

In particular, taking Z to consist of a single arbitrary row vector z gives

$$zQz^\top = 0 \qquad \text{for every } z.$$

Because Q is symmetric, its associated quadratic form uniquely determines the bilinear form. Equivalently, polarization gives

$$uQv^\top = \frac{1}{4}\left[(u+v)Q(u+v)^\top - (u-v)Q(u-v)^\top\right] = 0$$

for all u, v. Hence $Q = 0$, and therefore

$$M' = AMA^\top.$$

□

The universal quantifier over Z is part of the theorem. For a particular finite dataset, it is possible that

$$ZA^{-1}M'A^{-\top}Z^\top = ZMZ^\top$$

even when the two bilinear forms disagree outside the span of the realized covectors. Section A.4 characterizes this instance-dependent case.

We also record explicitly the connection from geometry to a deterministic selection path.

Corollary 1 (Gram preservation implies pathwise selection preservation)**.** *Let a selector with fixed side information and random tape ξ have the form*

$$S = \Phi(K, \xi).$$

If $K' = K$, then

$$\Phi(K', \xi) = \Phi(K, \xi).$$

Proof. With ξ and all deterministic side information fixed, Φ is a deterministic map of its geometric input. Equal inputs therefore produce equal outputs. □

The converse of Corollary 1 does not hold: different Gram matrices can lie in the same decision region of Φ and therefore yield the same selected set. We give explicit examples in Section A.10.

A.3 Projective and Conformal Geometry

We now specialize to a fixed Euclidean geometry,

$$M = M' = I.$$

The transformed Gram is

$$K' = ZA^{-1}A^{-\top}Z^\top.$$

The exact invariance criterion follows immediately from Theorem 1: $K' = K$ for every Z if and only if

$$AA^\top = I.$$

Thus universal exact Euclidean-Gram preservation is equivalent to orthogonality.

Proof of Corollary . Suppose first that

$$AA^\top = \alpha^{-1}I$$

for some $\alpha > 0$. Then

$$A^{-1}A^{-\top} = \alpha I,$$

and hence

$$K' = ZA^{-1}A^{-\top}Z^\top = \alpha ZZ^\top = \alpha K$$

for every Z.

Conversely, suppose

$$ZA^{-1}A^{-\top}Z^\top = \alpha ZZ^\top \qquad \text{for every } Z.$$

Applying the same one-row argument as in Theorem 1 gives

$$A^{-1}A^{-\top} = \alpha I.$$

Taking inverses yields

$$AA^\top = \alpha^{-1} I.$$

□

The last condition is equivalent to

$$A = cQ, \qquad c > 0, \qquad QQ^\top = I,$$

with $c = \alpha^{-1/2}$. Thus a conformal but nonorthogonal coordinate change changes every Euclidean Gram matrix only by one common positive scale. This case is important because it separates failure of *exact* Gram equality from distortion of the relative geometry.

A.4 Instance-Level Visible Subspaces

Universal non-conformality does not imply that a particular dataset sees a non-projective distortion. We now prove Proposition.

Let

$$C = A^{-1}A^{-\top}, \qquad \mathcal{W} = \operatorname{span}\{z_1, \dots, z_n\},$$

where the z_i are the rows of the realized selector-consumed matrix Z.

Proof of Proposition . Suppose first that

$$K' = \alpha K.$$

Entrywise, this means

$$z_i C z_j^\top = \alpha z_i z_j^\top \qquad \text{for every } i, j.$$

Every $u, v \in \mathcal{W}$ can be written as linear combinations

$$u = \sum_i a_i z_i, \qquad v = \sum_j b_j z_j.$$

By bilinearity,

$$\begin{aligned} uCv^\top &= \sum_{i,j} a_i b_j z_i C z_j^\top \\ &= \alpha \sum_{i,j} a_i b_j z_i z_j^\top \\ &= \alpha u v^\top. \end{aligned}$$

Therefore the bilinear form induced by C on $\mathcal{W}$ is a common multiple of the Euclidean form. Equivalently,

$$P_{\mathcal{W}} C P_{\mathcal{W}} = \alpha P_{\mathcal{W}}$$

as a bilinear form on $\mathcal{W}$.

Conversely, if the displayed relation holds on $\mathcal{W}$, then every row of Z lies in $\mathcal{W}$, so

$$z_i C z_j^\top = \alpha z_i z_j^\top$$

for every pair of rows. Therefore

$$ZCZ^\top = \alpha ZZ^\top,$$

which is exactly $K' = \alpha K$. □

The statement does not require

$$C\mathcal{W} \subseteq \mathcal{W}.$$

Only the restriction of the *bilinear form* to pairs of vectors in $\mathcal{W}$ matters.

A simple example makes this distinction concrete. Take

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \qquad C = \begin{bmatrix} 1/4 & 0 \\ 0 & 1 \end{bmatrix}.$$

The transformation is globally non-conformal. Nevertheless, if all consumed covectors lie in

$$\mathcal{W} = \mathrm{span}(e_1),$$

then

$$P_{\mathcal{W}} C P_{\mathcal{W}} = \frac{1}{4} P_{\mathcal{W}},$$

and hence

$$K' = \frac{1}{4} K$$

for this dataset. A scale-homogeneous selector therefore sees no change despite the global anisotropy.

A.5 Scale-Homogeneous Selectors

We next spell out the bridge between projective Gram invariance and selection invariance.

Proof of Corollary . If

$$K' = \alpha K, \qquad \alpha > 0,$$

and the selection primitive satisfies

$$\Phi(\alpha K, \xi) = \Phi(K, \xi)$$

for fixed side information and random tape, then

$$S' = \Phi(K', \xi) = \Phi(\alpha K, \xi) = \Phi(K, \xi) = S.$$

□

For completeness, we verify scale homogeneity for the exact mathematical primitives used in the paper.

Max-norm-first D^2 clustering. If

$$K' = \alpha K,$$

then every squared norm is multiplied by α:

$$\|z_i'\|^2 = \alpha \|z_i\|^2.$$

Thus the max-norm first center is unchanged, provided the same deterministic tie rule is used. Likewise every squared pairwise distance satisfies

$$d_{ij}' = \alpha d_{ij}.$$

At a later D^2 step, if

$$d_i = \min_{j \in S} d_{ij},$$

then

$$d_i' = \alpha d_i.$$

The normalized sampling probabilities therefore obey

$$\frac{d_i'}{\sum_j d_j'} = \frac{\alpha d_i}{\alpha \sum_j d_j} = \frac{d_i}{\sum_j d_j}.$$

With a common random tape, every subsequent sampled center is therefore the same by induction.

Squared-Euclidean facility location. For a representative set S, write the distance-form objective as

$$L(S) = \sum_i \min_{j \in S} d_{ij},$$

or equivalently use a facility-location similarity obtained from the same pairwise distances by a common affine shift. Under $d_{ij}' = \alpha d_{ij}$,

$$L'(S) = \alpha L(S).$$

Hence every exact greedy marginal gain is multiplied by the same positive constant, and the greedy path is unchanged under fixed tie handling.

Unregularized positive OMP. Consider one OMP block with candidate atoms and target vector represented in a common inner-product geometry. Let G_S be the Gram matrix of the currently selected atoms and let c_S contain their correlations with the target. A common positive rescaling of the inner product gives

$$G'_S = \alpha G_S, \qquad c'_S = \alpha c_S.$$

Whenever the least-squares system is nonsingular,

$$w'_S = (G'_S)^{-1} c'_S = (\alpha G_S)^{-1}(\alpha c_S) = w_S.$$

The refitted residual correlations are then all multiplied by the same positive factor, so the next positive-correlation maximizer is unchanged. Induction gives the same exact OMP path, provided the stopping criterion is itself scale free.

These are statements about the exact mathematical primitives. Fixed-decimal rounding, finite-precision accumulation, absolute stopping thresholds, and tie handling can destroy bitwise scale equivariance. In particular, the CRAIG and GradMatch implementations used in the experiments include fixed-decimal rounding. This numerical boundary does not alter the exact projective result above.

A.6 BLOCKWISE GRADMATCH

The faithful GradMatch implementation operates on class- and output-specific gradient blocks rather than one global Gram matrix. We make the corresponding sufficiency statement explicit.

For block b, let

$$Z^{(b)}$$

contain the candidate atoms used by positive OMP, and let $r^{(b)}$ denote the target gradient for the same block. It is convenient to augment the candidate matrix with the target:

$$\widetilde{Z}^{(b)} = \begin{bmatrix} Z^{(b)} \\ r^{(b)} \end{bmatrix}, \qquad \widetilde{K}^{(b)} = \widetilde{Z}^{(b)} M^{(b)} \widetilde{Z}^{(b)\top}.$$

This augmented Gram contains both candidate–candidate inner products and the candidate–target correlations required by OMP.

Proposition 2 (Blockwise preservation for GradMatch)**.** *Suppose that, for every block b,*

$$\widetilde{K}^{(b)\prime} = \widetilde{K}^{(b)},$$

and that the class partition, output-column definitions, per-class allocation, candidate ordering, positive-OMP convention, fill rule, tie rule, and random side information are unchanged. Then the final GradMatch selection is unchanged.

Proof. Within each block, the same augmented Gram implies the same initial candidate–target correlations. Hence the first OMP atom is the same. Given an identical selected prefix, the least-squares refit depends only on the corresponding Gram submatrix and target correlations, which are also identical. Induction therefore gives the same OMP path and weights within each block. Since the block allocation and final fill/union rules are fixed, the combined selected set is identical. □

A sufficient geometric condition is obtained blockwise from Theorem 1. If

$$Z^{(b)\prime} = Z^{(b)} A_b^{-1}$$

and

$$M^{(b)\prime} = A_b M^{(b)} A_b^{\top}$$

for each consumed block, then every block Gram is preserved. The same statement applies projectively to the exact OMP primitive when each augmented block Gram is scaled by a positive block-specific constant.

A.7 ReLU Orbit Geometry and Crossing Constructions

We now derive the orbit formula used to explain how an exact function-preserving ReLU rescaling can move a fixed geometric comparison through a decision boundary.

Consider one hidden unit whose incoming parameters are multiplied by e^t and whose outgoing parameters are multiplied by e^{-t}. Positive homogeneity preserves the represented function. By covector transport, the per-example gradient coordinates associated with the incoming parameters are multiplied by e^{-t}, while those associated with the outgoing parameters are multiplied by e^t. Unaffected parameter blocks remain unchanged.

Write the concatenated consumed covector as

$$z_i(t) = \big(e^{-t} z_{i,\mathrm{in}}, e^{t} z_{i,\mathrm{out}}, z_{i,\mathrm{rest}}\big).$$

Because the blocks occupy disjoint parameter coordinates,

$$\begin{aligned} z_i(t) z_j(t)^\top = e^{-2t} z_{i,\mathrm{in}} z_{j,\mathrm{in}}^\top + e^{2t} z_{i,\mathrm{out}} z_{j,\mathrm{out}}^\top \\ + z_{i,\mathrm{rest}} z_{j,\mathrm{rest}}^\top. \end{aligned}$$

Stacking all candidates gives

$$K(t) = e^{-2t} K_{\mathrm{in}} + e^{2t} K_{\mathrm{out}} + K_{\mathrm{rest}}.$$

Let L be any fixed affine functional of the Gram entries. Then

$$\Delta(t) = L(K(t)) = a e^{-2t} + b e^{2t} + c$$

for constants a, b, c determined by L and the three block Grams. If

$$\Delta(t_-)\Delta(t_+) < 0,$$

continuity gives a $t^\star \in (t_-, t_+)$ such that

$$\Delta(t^\star) = 0,$$

and the corresponding ordering is reversed on the two sides of the crossing.

This derivation applies to a *fixed* affine comparison. It is not a global closed form for an arbitrary sequential selector. Later BADGE steps depend on the already selected centers and a minimum-distance operation; CRAIG uses piecewise facility gains; and GradMatch repeatedly changes its residual after refitting. The next subsection therefore gives direct first-choice constructions matched to the corresponding selector representations.

A.8 Selector-Specific First-Choice Reversals

We now prove Proposition 1 using explicit two-coordinate constructions. All three examples can be embedded in an ordinary ReLU classifier.

Consider two positive hidden coordinates and the rescaling

$$D(t) = \begin{bmatrix} e^{-t} & 0 \\ 0 & e^{t} \end{bmatrix}.$$

Rescale the two hidden units by $D(t)$ and compensate the following classifier weights by $D(t)^{-1}$. The predictor is unchanged. To realize the two-dimensional gradients below, take a one-hidden-layer ReLU model in a region where the chosen hidden activations are positive, and choose classifier weights so that the relevant examples share the same nonzero classifier-error factor. The classifier-weight gradient is then an outer product of this common error factor with the hidden activation. Consequently, its norm, inner products, and pairwise distances differ from those of the two-dimensional vectors below only by a common positive multiplicative factor. Coordinates outside this two-dimensional block may be set identically across candidates.

BADGE: max-norm first center. Take two candidate gradient coordinates

$$z_A = (2, 1), \qquad z_B = (1, 2).$$

After reparameterization,

$$z_A(t) = (2e^{-t}, e^t), \qquad z_B(t) = (e^{-t}, 2e^t).$$

Their squared norms are

$$\|z_A(t)\|^2 = 4e^{-2t} + e^{2t},$$

and

$$\|z_B(t)\|^2 = e^{-2t} + 4e^{2t}.$$

Hence

$$\|z_A(t)\|^2 - \|z_B(t)\|^2 = 3\left(e^{-2t} - e^{2t}\right).$$

For every $t < 0$, candidate A has the larger norm, while for every $t > 0$, candidate B has the larger norm. Thus the max-norm first center changes strictly along a function-preserving orbit. With budget one, the selected set therefore changes.

CRAIG: budget-one Euclidean facility choice. Consider three candidate gradient coordinates

$$z_A = (1, 1), \qquad z_B = (1, 4), \qquad z_C = (2, 2).$$

Under the same transformation,

$$d_{AB}(t) = 9e^{2t},$$
$$d_{AC}(t) = e^{-2t} + e^{2t},$$

and

$$d_{BC}(t) = e^{-2t} + 4e^{2t},$$

where $d_{ij}(t) = \|z_i(t) - z_j(t)\|^2$.

For a budget-one Euclidean facility objective, selecting one representative is equivalent to minimizing the total squared distance to that representative. The three costs are therefore

$$L_A(t) = e^{-2t} + 10e^{2t},$$
$$L_B(t) = e^{-2t} + 13e^{2t},$$

and

$$L_C(t) = 2e^{-2t} + 5e^{2t}.$$

Candidate B is always worse than A. Comparing A and C gives

$$L_A(t) - L_C(t) = -e^{-2t} + 5e^{2t}.$$

The sign changes at

$$t^\star = -\frac{1}{4}\log 5.$$

For $t < t^\star$, A is the unique facility choice; for $t > t^\star$, C is the unique facility choice. Hence the budget-one CRAIG decision reverses along the same function-preserving family. An equivalent construction applies to the usual shifted-similarity form of the facility-location objective because a common affine shift does not alter the argmax.

GradMatch: first positive-OMP atom. Take

$$z_A = (2, 1), \qquad z_B = (1, 2),$$

and let the target gradient in the block be their sum,

$$r = z_A + z_B = (3, 3).$$

After reparameterization,

$$r(t) = (3e^{-t}, 3e^t).$$

The first-step correlations are

$$\langle z_A(t), r(t)\rangle = 6e^{-2t} + 3e^{2t},$$

and

$$\langle z_B(t), r(t)\rangle = 3e^{-2t} + 6e^{2t}.$$

Their difference is

$$3\left(e^{-2t} - e^{2t}\right).$$

Thus positive OMP selects A first for $t < 0$ and B first for $t > 0$. Both correlations are positive, and the corresponding one-atom least-squares coefficient is positive because

$$\frac{\langle z_i(t), r(t)\rangle}{\|z_i(t)\|^2} > 0.$$

This gives a strict first-atom and budget-one reversal within one faithful GradMatch block.

These constructions establish existence, not prevalence. They do not imply that every reparameterization changes the first choice, nor that a first-choice change forces two larger unordered batches to differ. The large-budget replacement rates in the main experiments are empirical properties of the corresponding model–data instances.

A.9 BAIT Trace Criterion and Regularizer Rules

We first give the full proof of Theorem 2. The theorem concerns a fixed pair of priors and a family of *independent* SPD probe matrices.

Let

$$\mathcal{J}_{R,H}(G) = \operatorname{tr}\left[(R+H)^{-1}G\right],$$

with

$$R \succ 0, \qquad H \succeq 0, \qquad G \succ 0.$$

Proof of Theorem 2: sufficiency. Suppose

$$R' = TRT^\top.$$

Then

$$R' + THT^\top = T(R+H)T^\top,$$

and hence

$$(R' + THT^\top)^{-1} = T^{-\top}(R+H)^{-1}T^{-1}.$$

Therefore

$$\begin{aligned}\operatorname{tr}\left[(R' + THT^\top)^{-1}TGT^\top\right] &= \operatorname{tr}\left[T^{-\top}(R+H)^{-1}T^{-1}TGT^\top\right] \\ &= \operatorname{tr}\left[T^{-\top}(R+H)^{-1}GT^\top\right] \\ &= \operatorname{tr}\left[(R+H)^{-1}G\right],\end{aligned}$$

where the last equality follows from cyclicity of the trace.

Necessity. Now suppose

$$\operatorname{tr}\left[(R' + THT^\top)^{-1}TGT^\top\right] = \operatorname{tr}\left[(R+H)^{-1}G\right]$$

for every independent $G \succ 0$. Cyclicity gives

$$\operatorname{tr}\left[\left(T^\top(R' + THT^\top)^{-1}T - (R+H)^{-1}\right)G\right] = 0$$

for every SPD G.

The SPD cone has nonempty interior in the vector space of symmetric matrices. A linear functional that vanishes on this open set vanishes on the entire space. Hence

$$T^\top(R' + THT^\top)^{-1}T = (R+H)^{-1}.$$

Taking inverses yields

$$T^{-1}(R' + THT^\top)T^{-\top} = R + H.$$

Expanding and cancelling H gives

$$T^{-1}R'T^{-\top} = R,$$

or equivalently

$$R' = TRT^\top.$$

□

The independent-probe quantifier is necessary for the converse. Equality for one fixed target matrix, one candidate pool, or one selected set does not identify the prior uniquely.

Data-dependent regularizer rules. Suppose now that the regularizer is generated by a rule

$$R = \mathcal{R}(F)$$

from a Fisher-type matrix F. We use only the following sufficiency result.

Corollary 2 (Equivariant regularizer rules)**.** *Suppose*

$$F' = TFT^\top$$

and

$$\mathcal{R}(TFT^\top) = T\mathcal{R}(F)T^\top.$$

If every other Fisher term in the objective transforms by the same congruence, then the regularized trace objective is pointwise invariant, provided the required inverse exists.

Proof. The regularizer produced in transformed coordinates is precisely the congruence transform of the original regularizer, so the sufficiency direction of Theorem 2 applies pointwise. □

Three cases used in the paper follow immediately.

Proportional full-Fisher regularization. For

$$\mathcal{R}(F) = \lambda F,$$
$$\mathcal{R}(TFT^\top) = \lambda TFT^\top = T(\lambda F)T^\top$$

for every invertible T.

Diagonal-Fisher regularization. For

$$\mathcal{R}(F) = \mathrm{Diag}(F),$$

the rule is congruence equivariant under the corresponding monomial group $T = PD$, with P a permutation and D an invertible diagonal matrix:

$$\mathrm{Diag}(TFT^\top) = T\,\mathrm{Diag}(F)T^\top.$$

This statement does not extend to a general dense T.

Fixed identity prior. For a coordinate-fixed prior

$$R = I$$

that is reset to I after the coordinate change, congruence compatibility requires

$$I = TIT^\top = TT^\top.$$

Thus a fixed identity prior is universally compatible with the transformation only on the orthogonal subgroup.

Singular Fisher support. If the relevant Fisher matrices are singular, the same statements can be applied on a subspace on which the required inverse is well defined. If the original support is $\mathcal{V}$, its transformed counterpart is

$$\mathcal{V}' = T\mathcal{V},$$

and the relevant map is the isomorphism

$$T : \mathcal{V} \to \mathcal{V}'.$$

The support should not in general be treated as the same coordinate subspace before and after a dense transformation.

From objective equality to a greedy path. If every objective value compared by an exact greedy BAIT step is preserved, and the candidate ordering and tie rule are fixed, then the selected candidate at that step is preserved. Induction gives the same greedy path. The converse does not hold: two objectives may differ while producing the same argmin over the candidates encountered by one particular run.

A.10 Counterexamples and Claim Boundaries

We close with a collection of counterexamples that delimit the statements in the main text.

Nonorthogonal does not imply selector-visible distortion. Take

$$A = 3I.$$

This transformation is nonorthogonal, but

$$A^{-1}A^{-\top} = \frac{1}{9}I,$$

so

$$K' = \frac{1}{9}K$$

for every Z. Every exact scale-homogeneous selector therefore returns the same path.

This is not merely an algebraic possibility. For the normalized linear predictor

$$f_w(x) = \left\langle \frac{w}{\|w\|}, x \right\rangle, \qquad w \neq 0,$$

the transformation

$$w \mapsto cw, \qquad c > 0,$$

is exactly function preserving. Its Jacobian is cI, giving precisely the conformal-but-nonorthogonal case above.

Global non-conformality need not be visible to the realized data. As in Section A.4, take

$$A = \mathrm{diag}(2, 1).$$

The ambient transformation is non-conformal, but if

$$\mathcal{W} = \mathrm{span}(e_1),$$

then

$$K' = \frac{1}{4}K$$

for every realized Z whose rows lie in $\mathcal{W}$. Thus ambient anisotropy alone does not imply a non-projective distortion of the data actually consumed by the selector.

A changed Gram need not change the selected set. Consider max-norm selection with

$$z_1 = (10, 0), \qquad z_2 = (1, 1),$$

and

$$A = \mathrm{diag}(2, 1).$$

The transformed covectors are

$$z_1' = (5, 0), \qquad z_2' = (1/2, 1),$$

so the Gram matrix changes. Nevertheless,

$$\|z_1\|^2 = 100 > \|z_2\|^2 = 2$$

and

$$\|z_1'\|^2 = 25 > \|z_2'\|^2 = \frac{5}{4}.$$

The selected index is therefore unchanged. This demonstrates directly why Gram invariance is sufficient, but not necessary, for a particular selection to remain fixed.

The ReLU three-term orbit formula is not a global sequential-score formula. every fixed affine functional of that Gram has the three-term form in Equation equation 9. A sequential selector, however, changes its internal branch when a previous decision changes. For example, later BADGE distances depend on the currently selected centers, CRAIG contains a max over representatives, and GradMatch recomputes its residual after each refit. Their full paths are therefore generally piecewise functions of the orbit parameter rather than one fixed expression $ae^{-2t} + be^{2t} + c$.

Fixed-prior necessity does not imply rule-level necessity. The necessity direction of Theorem 2 keeps R and R' fixed while varying an independent SPD probe. If instead

$$R = \mathcal{R}(F)$$

changes when F changes, that proof cannot be reused by simultaneously treating F as the independent probe. For example,

$$\mathcal{R}(F) = F$$

is manifestly congruence equivariant. The appropriate general statement for a data-dependent rule is therefore the sufficiency condition in Corollary 2; a rule-level necessity theorem would require additional richness assumptions.

Dense transformations move singular supports. If

$$\mathcal{V} = \mathrm{span}(e_1)$$

and T is a generic dense invertible matrix, then

$$T\mathcal{V} \neq \mathcal{V}.$$

Restricting both objectives to the same coordinate support after transformation is therefore not a congruence-covariant operation. Singular support must itself be transported.

Zero coordinates are a boundary of multiplicative representations. A zero parameter does not make an invertible positive-rescaling map noninvertible. Instead, zero coordinates create different difficulties:

$$u_j = \log |\theta_j|$$

is undefined,

$$\mathrm{Diag}(\theta^2)$$

loses positive definiteness, and

$$q_j = \theta_j g_j$$

can erase a nonzero raw gradient whenever $\theta_j = 0$. Some zero strata may also have nontrivial stabilizers, but non-freeness does not follow from a zero coordinate alone.

Exact scale homogeneity is not bitwise implementation invariance. The scale-homogeneity arguments in Section A.5 concern exact mathematical primitives. Fixed-decimal rounding in CRAIG and GradMatch, floating-point accumulation, absolute tolerances, and near ties can produce different discrete paths after a common rescaling. These effects belong to numerical reproducibility rather than to the exact projective theorem; we examine them empirically in Appendix F.

Taken together, these examples delineate the logic used in the main text:

$$\begin{aligned} &\text{coordinate change} \longrightarrow \text{geometry visible on the realized subspace} \\ &\longrightarrow \text{decision-boundary crossing.} \end{aligned}$$

The first arrow can be characterized geometrically, whereas the second is instance- and selector-dependent. The theory therefore establishes conditions for well-defined geometry and explicit failures of selector identifiability, rather than asserting that every function-preserving reparameterization must change every query.

B Experimental Protocols and Statistical Units

This section specifies the experimental conventions needed to interpret the results in the main paper. We focus on four aspects that are particularly relevant to an identifiability study: the exact selection primitive, the function-preserving transformation being applied, the metric used to compare selected sets, and the level at which observations are statistically independent. Architecture- and task-specific settings that are only relevant to a particular robustness experiment are given with the corresponding appendix section.

B.1 SELECTOR IMPLEMENTATIONS

The main experiments use the specified geometric primitives associated with the corresponding published selection rules rather than a common surrogate selector. This distinction is important because parameterization dependence must be evaluated at the level of the actual decision rule, not only through an intermediate gradient statistic.

Gradient-norm ranking. For candidate i, gradient-norm ranking uses

$$s_i = \|z_i\|_2^2$$

and returns the k largest scores. We use this primitive primarily to expose the simplest dependence of a scalar gradient statistic on parameter coordinates and in the training-stage analyses of Appendix C.2.

BADGE. For BADGE (Ash et al., 2020), each unlabeled candidate is assigned its current predicted label and represented by the corresponding classifier-weight gradient embedding. The complete classifier-weight embedding is used in the faithful vision experiments rather than a random projection; the classifier bias is not included in this embedding. Batch construction follows the reference max-norm-first convention: the first center is the candidate with largest embedding norm, after which the remaining centers are sampled according to the D^2 primitive used by k-means++ (Arthur & Vassilvitskii, 2007). Thus both the initial norm ranking and the subsequent D^2 geometry can depend on the parameter-space representation.

For paired reparameterization comparisons, the same random numbers are used for the D^2 sampling steps. This coupling isolates changes induced by the gradient geometry from ordinary variability across independent BADGE runs. Independent-tape behavior is examined separately in Appendix D.2.

CRAIG. CRAIG (Mirzasoleiman et al., 2020) is implemented using exact squared Euclidean distances between the classifier-gradient representations consumed by the selector, followed by greedy facility-location selection, a standard monotone-submodular setting for greedy maximization (Nemhauser et al., 1978). The representation includes the classifier weight and bias terms used by the selection objective. Selection is performed within the prescribed class allocation before the classwise selections are combined. Consequently, identical class counts across two runs do not imply identical selected examples; our primary comparison remains sample identity.

GradMatch. GradMatch (Killamsetty et al., 2021a) is implemented blockwise. Candidates are partitioned by class and by the output-gradient blocks consumed by the method, and positive orthogonal matching pursuit (Tropp & Gilbert, 2007) is applied within each block to approximate the corresponding aggregate gradient. The block allocation, candidate order, output-column definitions, fill rule, and tie handling are kept fixed across paired parameterizations. This implementation motivates the blockwise formulation in Appendix A.6: preserving a single global Gram matrix is not the appropriate implementation-level abstraction.

Both CRAIG and GradMatch apply fixed-decimal rounding in their numerical decision paths. The exact scale-homogeneity statements in Appendix A.5 therefore describe the underlying mathematical primitives rather than a bitwise guarantee for the finite- precision implementations.

BAIT. For BAIT (Ash et al., 2021), we form exact pointwise Fisher information using the reduced-softmax parameterization. The batch-selection procedure follows the forward-$2B$/backward-B structure of the reference method. In the mechanism experiments of Figure 3, the candidate pool contains 160 examples and the final batch size is $B = 8$, so the forward stage retains 16 candidates before backward elimination.

The published form uses a coordinate-fixed identity regularizer. We compare it with three geometrically matched alternatives: a transported identity, a Fisher-diagonal regularizer in the transformation family where that rule is equivariant, and a full-pool-Fisher regularizer. These variants change the regularization geometry, not the candidate data or Fisher point estimates.

Table 2: **Operational mapping for the audited classical selectors.** The consumed gradient block, selection primitive, random side information, and numerical convention are fixed within every paired comparison. This table defines the implementations used by the empirical claims; "specified" is an implementation description, not a claim of unique fidelity among all software realizations of the published methods.

Selector	Consumed representation	Selection primitive	Randomness / initialization	Rounding and ties
BADGE	complete classifier-weight pseudo-label gradient; classifier bias excluded	maximum-norm first center, then squared-distance D^2 sampling	common random tape for every paired D^2 step	no projection; candidate order and deterministic tie handling fixed
CRAIG	classifier weight-and-bias gradients within the prescribed class allocation	greedy squared-Euclidean facility-location selection	candidate ordering and class allocation fixed; no new random draw in the paired path	fixed-decimal rounding and tie handling fixed
GradMatch	class- and output-specific candidate gradients and their aggregate block target	blockwise positive OMP followed by the fixed allocation/fill rule	block allocation, output columns, candidate order, and other side information fixed	fixed-decimal rounding and tie handling fixed

LoRA factor-coordinate selection. The PEFT experiments compare several representations of the same fixed-rank LoRA update. Raw selection uses per-example gradients expressed in the coordinates of the LoRA factors. Factor-QGrad applies the corresponding elementwise factor-times-gradient construction. The balanced control first maps an equivalent factorization to a balanced representative, whereas the merged control removes the internal factor basis by constructing the selection geometry from the represented update.

The exact realization of the merged control depends on the experiment. The dense-LoRA experiment in Appendix C.4 computes exact merged-gradient geometry. For MiniLM and ModernBERT, the analytical merged control is a fixed basis-invariant merged sketch. It is invariant to the evaluated LoRA basis transformations but is not identical to exact merged-gradient selection; Appendix G measures that approximation error directly.

Table 2 makes the operational content of the three classical selector implementations inspectable in one place.

B.2 Reparameterization Families

All paired comparisons change a parameter representative while preserving the mathematical object that determines the predictor. We use different symmetry families for homogeneous networks and low-rank adaptation.

Positive hidden-unit rescaling. For a positively homogeneous hidden layer, let D be a positive diagonal matrix. We apply

$$W' = DW, \qquad b' = Db, \qquad V' = VD^{-1}, \tag{13}$$

where (W, b) are the incoming parameters of the hidden layer and V contains the outgoing weights. Positive homogeneity gives

$$V'\sigma(W'x + b') = V\sigma(Wx + b).$$

The transformation therefore changes the parameter coordinates while preserving the represented predictor in exact arithmetic.

The faithful BADGE, CRAIG, and GradMatch experiments use multiple positive rescalings nested within each independently trained model. The main comparison uses ten such transformations per model; Appendix C.1 varies the magnitude of the rescaling to characterize how replacement depends on transformation strength. The BAIT experiments use the corresponding positive coordinate-rescaling family on the Fisher parameter block.

LoRA changes of basis. For a rank-r update

$$\Delta W = BA,$$

we use the fixed-rank transformation

$$B' = BR^{-1}, \qquad A' = RA, \qquad R \in GL(r), \tag{14}$$

for which

$$B'A' = BA.$$

The MiniLM and ModernBERT experiments apply nonorthogonal basis changes to each trained model instance while leaving the represented LoRA update unchanged. MiniLM additionally includes orthogonal basis changes, for which raw factor-coordinate Euclidean geometry is itself invariant; these comparisons are reported in Appendix D.4.

The analytical PEFT comparisons transport the per-example gradient quantities according to the exact factor transformation. The physical experiments in Appendix F.1 instead materialize an equivalent factorization in the model and recompute the forward and backward quantities. The latter therefore include finite-precision effects in addition to the mathematical change of coordinates.

Canonical and balanced representatives. Some controls remove the coordinate freedom by selecting one representative from an equivalence class rather than explicitly transporting a metric. For positive hidden-unit rescaling, the canonical construction fixes the relevant hidden-unit scale before computing selection geometry. For fixed-rank LoRA, the balanced construction uses a balanced factorization of the same merged update. These operations are used as invariance controls; their role is to remove the specified representation freedom, not to define a universally preferred parameterization.

B.3 LoRA Factor-Gradient Transport

For clarity, we state the matrix convention behind Equation equation 12. Let the scalar per-example loss depend on the LoRA factors through $\Delta W = BA$, and define matrix gradients by the Frobenius differential

$$d\ell = \operatorname{tr}(G_B^\top dB) + \operatorname{tr}(G_A^\top dA).$$

Under $B' = BR^{-1}$ and $A' = RA$, the inverse coordinate relations are $B = B'R$ and $A = R^{-1}A'$. Hence $dB = dB'R$ and $dA = R^{-1}dA'$, so

$$G_{B'} = G_B R^\top, \tag{15}$$

$$G_{A'} = R^{-\top} G_A. \tag{16}$$

Here G_B has the same row–column layout as B and G_A the same layout as A; no implicit transpose or row-vectorization convention is used. If the two matrices are subsequently flattened, the implementation uses the same fixed flattening order before and after transformation. These are the transport equations used by the retained Transformer experiments and by the direct-autograd checks; superseded outputs obtained with a different transpose convention are excluded from every quantitative claim.

B.4 Evaluation Metrics

Query replacement. Our primary metric compares the identities of two equal-cardinality selected sets. For $|S| = |S'| = k$, we define

$$\operatorname{Rep}(S, S') = 1 - \frac{|S \cap S'|}{k}. \tag{17}$$

Thus $\operatorname{Rep} = 0$ denotes identical query sets and $\operatorname{Rep} = 1$ denotes disjoint sets. If

$$J(S, S') = \frac{|S \cap S'|}{|S \cup S'|}$$

is the Jaccard index, then for equal-size sets

$$\operatorname{Rep}(S, S') = \frac{1 - J(S, S')}{1 + J(S, S')}. \tag{18}$$

We report replacement rather than $1 - J$ because it has the direct interpretation "fraction of the query batch replaced."

Function discrepancy. When transformed parameters are materialized, we evaluate the discrepancy between the original and reparameterized predictors on the relevant candidate or evaluation inputs. We report the maximum absolute logit discrepancy,

$$\Delta_{\text{logit}} = \max_{i,c} |f_\theta(x_i)_c - f_{\theta'}(x_i)_c| .$$

This quantity distinguishes changes in selection geometry from changes in the represented predictor. Where pseudo-labels are used, we additionally check whether the transformation changes the predicted class.

Gram discrepancy. For comparisons between two geometric representations, we use the relative Frobenius error

$$\varepsilon_K = \frac{\|K' - K\|_F}{\|K\|_F}, \tag{19}$$

whenever the reference Gram is nonzero. Exact-invariance experiments use this quantity primarily as a numerical closure check; the held-out sketch study in Appendix G.1 instead uses it as an approximation metric relative to exact merged-gradient geometry.

Batch-composition differences. For analyses that compare the class composition of two selected batches, we use total variation between their empirical label distributions,

$$\text{TV}(p, q) = \frac{1}{2} \sum_c |p_c - q_c|. \tag{20}$$

For selectors with an explicit per-class allocation, class-level TV can be zero by construction even when the selected identities differ. We therefore never use label TV as a substitute for query replacement.

Downstream utility. Downstream classification results are reported as paired accuracy differences in percentage points,

$$\Delta_{\text{acc}} = 100 \big(\text{Acc}(S') - \text{Acc}(S)\big).$$

Training randomness is paired whenever the protocol permits, so that the reported difference compares models trained under matched stochastic conditions. Accuracy is analyzed separately from query replacement; no monotone relationship between the two metrics is assumed.

Multiround replacement. In repeated acquisition, we distinguish the query selected at a particular round from the cumulative acquired set. If S_t and S'_t are the batches queried at round t, per-round replacement is

$$\text{Rep}(S_t, S'_t).$$

For cumulative labeled additions

$$C_t = \bigcup_{\tau \le t} S_\tau, \qquad C'_t = \bigcup_{\tau \le t} S'_\tau,$$

we analogously report

$$\text{Rep}(C_t, C'_t).$$

The two quantities answer different questions: the former measures divergence of the current acquisition decision, whereas the latter measures divergence of the data histories accumulated so far.

B.5 Statistical Units and Aggregation

The independent statistical unit is the independently trained acquisition model or model checkpoint, not an individual reparameterization. Multiple transformations of one parameter vector probe different representatives of the same predictor orbit and are therefore nested observations. Likewise, multiple selection tapes, annotation tapes, sketch dimensions, or paired downstream fits associated with the same trained model do not create additional independent model instances.

For each quantitative comparison, we first compute the relevant metric for the paired original and transformed conditions. We then aggregate all nested transformations or repeated tapes within an independently trained model or checkpoint. Means, intervals, and the small points shown in the main figures are formed from these model-level or checkpoint-level quantities.

Table 3 summarizes the principal experimental families and the corresponding nesting structure.

Table 3: **Independent statistical units and nested repetitions.** Reparameterizations, selector tapes, annotation tapes, downstream refits, and representation variants are treated as repeated measurements within a trained model or checkpoint when they share that underlying model instance. Reported uncertainty across model instances is therefore not inflated by counting these nested measurements as independent samples.

Experiment family	Independent unit	Nested measurements	Main comparison
Faithful ReLU selectors	trained acquisition model	10 reparameterizations per model; fixed selector tape	natural vs. reparameterized BADGE / CRAIG / GradMatch
BAIT geometry	trained acquisition model	20 reparameterizations per model; regularizer variants	fixed identity vs. covariant Fisher/prior geometry
BAIT downstream	trained acquisition model	paired downstream training seeds	query change vs. paired accuracy change
Five-round active learning	acquisition model	acquisition rounds and paired terminal refits	trajectory replacement and terminal utility
Dense LoRA	trained adapter model	10 dense $GL(r)$ basis changes per model	factor-coordinate vs. basis-invariant geometry
MiniLM LoRA	trained model instance at a fixed rank	nonorthogonal and orthogonal basis changes; continuation tapes nested within model	factor-basis dependence across ranks
ModernBERT LoRA	trained model instance	9 nonorthogonal basis changes per model	factor-basis dependence across model instances
Merged-gradient fidelity	ModernBERT model instance	sketch dimensions evaluated within model	approximate sketch vs. exact merged-gradient selection

Paired quantities. Whenever two conditions originate from the same trained model, candidate pool, or acquisition history, we preserve that pairing in the reported difference. For example, a downstream accuracy delta is computed within a model instance before model-level aggregation, rather than subtracting two independently averaged accuracies. The same convention applies to function discrepancies and query replacement.

Randomized selectors. The main identifiability comparisons deliberately couple the random tape: the scientific variable is the parameter representative, not a new random draw. Independent-tape variability is a separate estimand and is analyzed in Appendix D.2. Likewise, the marginal inclusion-frequency analysis in Appendix D.1 treats repeated tapes as Monte Carlo samples for estimating a selection law, rather than as independent trained models.

Heterogeneous utility comparisons. The turnover–utility map in Figure 5 combines several experimental families only for descriptive comparison. Each plotted point pairs turnover and utility from the same protocol, but the tasks, selectors, and training procedures differ across families. We therefore do not pool these observations into a common regression, correlation, or cross-family significance test.

C Robustness Across Reparameterization Strength, Training Stage, and Architecture

The main experiments establish that function-preserving reparameterizations can change the examples returned by several gradient-based selectors. We now ask how broadly this effect persists across three dimensions that could otherwise provide alternative explanations: the magnitude of the reparameterization, the stage of model training, and the representation on which selection is performed. These experiments are not used to define a universal prevalence rate; rather, they characterize when the effect becomes visible and how its magnitude depends on the selector and model setting.

C.1 Sensitivity to Reparameterization Strength

The headline experiments use a fixed family of positive hidden-unit rescalings, but a large replacement rate under one transformation magnitude would not establish that the phenomenon remains visible for milder coordinate changes. We therefore vary the magnitude of the reparameterization while holding the trained model, candidate pool, selector, and random tape fixed.

For each of MNIST, Fashion-MNIST, and CIFAR-10, we use ten independently trained models. Each model is evaluated at five reparameterization strengths, with ten rescaling directions nested within the model at each strength.

BADGE is already sensitive at the smallest evaluated strength. At $\sigma = 0.05$, the mean replacement rates are 38.8%, 54.9%, and 53.9% on MNIST, Fashion-MNIST, and CIFAR-10, respectively. Increasing the strength to $\sigma = 0.50$ raises these values to 53.0%, 63.6%, and 63.4%. Thus, for BADGE, much of the observed set-level instability appears before the largest transformation magnitude considered here.

CRAIG and GradMatch show a different profile. Their replacement rates increase more gradually with the transformation strength, reaching approximately 18.8–50.8% for CRAIG and 21.9–50.5% for GradMatch at $\sigma = 0.50$, depending on the dataset. The contrast indicates that reparameterization sensitivity is not characterized by a single universal severity curve: the rate at which geometric distortion reaches a discrete decision depends on both the selection primitive and the model–data instance.

The same experiment also allows us to compare different representatives on the same parameter orbit rather than always using the trained parameterization as one endpoint. At the largest evaluated strength, pairwise reparameterization-to-reparameterization replacement is often at least as large as trained-representation-to-reparameterization replacement; for CIFAR-10, the corresponding CRAIG and GradMatch comparisons reach approximately 62% and 57%. This behavior is consistent with treating the trained parameter vector as one representative of the equivalence class rather than as an intrinsic geometric center of that class.

C.2 Training-Stage Dependence

A second possible explanation is that reparameterization sensitivity is mainly a property of poorly trained or weakly organized representations. We therefore measure the same coordinate dependence at initialization, after three epochs, and after thirty epochs of training in an MNIST MLP. Each stage uses five independently trained model instances and twenty reparameterizations nested within each instance.

For this comparison, we report Jaccard overlap,

$$J(S, S') = \frac{|S \cap S'|}{|S \cup S'|},$$

so larger values indicate more stable selections. The corresponding acquisition-model accuracies increase from 0.097 at initialization to 0.787 after three epochs and 0.871 after thirty epochs, allowing us to compare parameterization sensitivity as the predictive model becomes substantially better trained.

The result rules out an interpretation in which the observed coordinate dependence is purely an initialization phenomenon. At the same time, the different trajectories across selection rules argue against a single training-stage mechanism. Training can change score margins, relative block energies, and gradient geometry simultaneously; these measurements therefore establish persistence through training rather than attributing the change to one specific internal mechanism.

The invariant representations used in this comparison remain unchanged across the reparameterizations at all three stages, showing that the stage dependence belongs to the coordinate-sensitive geometry rather than to a loss of the underlying function-preserving relation.

C.3 Pretrained Representations

The experiments above train both the feature representation and the selector head on the target task. To test whether the effect instead disappears when selection is built on a pretrained representation, we use ImageNet-pretrained ResNet-18 features for CIFAR-10 and train only a $512 \to 64 \to 10$ ReLU classification head. The ResNet-18 representation is kept fixed. We evaluate five independently trained heads and ten hidden-unit reparameterizations per head.

BADGE still exhibits substantial query replacement in this setting: the mean replacement rate is 60.3%, with model-level range $[57.6, 62.4]\%$. The true-label composition of the selected batch also changes, with mean label-distribution total variation of 12.6%. By contrast, the corresponding invariant and canonical representations return identical query sets under the same reparameterizations.

The downstream accuracy differences remain small. Relative to the natural parameterization, the paired accuracy changes are approximately $+0.13$ percentage points for the reparameterized raw selection, $+0.24$ percentage points for the multiplicative invariant representation, and -0.02 percentage points for the canonical representation. We therefore use this experiment to establish robustness of the *selection* phenomenon to a pretrained feature representation, not to argue for either a performance benefit or harm.

This setting also provides an intermediate architectural regime between the small fully trained vision models and the end-to-end Transformer PEFT experiments in Section 6.1. The coordinate freedom here remains confined to the trained ReLU head; the Transformer experiments address parameter redundancy inside the pretrained model itself.

C.4 Dense LoRA Basis Transformations

The ReLU experiments involve coordinatewise positive rescalings. A natural question is whether an elementwise invariant representation remains sufficient when the parameter redundancy mixes coordinates densely. We examine this question using a rank-8 LoRA update on fixed pretrained visual features, where the same merged update admits the basis transformation

$$(B, A) \mapsto (BR^{-1}, RA), \qquad R \in GL(8).$$

We train five independent adapter models and apply ten dense basis changes within each model. Selection is performed from a pool of 500 candidates with budget 25, and the geometric calculations use float64 arithmetic. Figure 6 compares four representations of the same merged update.

Raw factor-coordinate selection replaces 45.4% of the queried examples on average, with model-level range $[39.5, 50.5]\%$. The elementwise factor-QGrad representation does not remove this dependence; its replacement rate is 52.0% with range $[47.8, 55.4]\%$. In contrast, the balanced representation and the exact merged-gradient representation return identical query sets across all evaluated basis changes.

The numerical agreement of the invariant geometries is also at machine precision: the relative Gram discrepancy is below 8.5×10^{-15}, while the maximum logit discrepancy between equivalent models is 4.26×10^{-14}. The failure of factor-QGrad is therefore not explained by a change in the represented predictor.

This experiment delineates the group dependence of an invariant construction. Coordinatewise multiplication can cancel diagonal rescalings, but a dense $GL(r)$ transformation mixes axes and cannot in general be removed by the same elementwise rule. The exact merged representation eliminates the factor basis entirely, while balancing selects a representative within the fixed-rank equivalence class. The end-to-end MiniLM and ModernBERT experiments in Section 6.1 test the same basis-dependence question when LoRA modules are embedded throughout a Transformer.

D Randomness and Alternative Explanations

The main experiments couple selector randomness so that the parameter representative is the only intended source of variation. This appendix asks whether the resulting query differences could instead be explained as ordinary stochastic dispersion, whether they imply a comparably large change

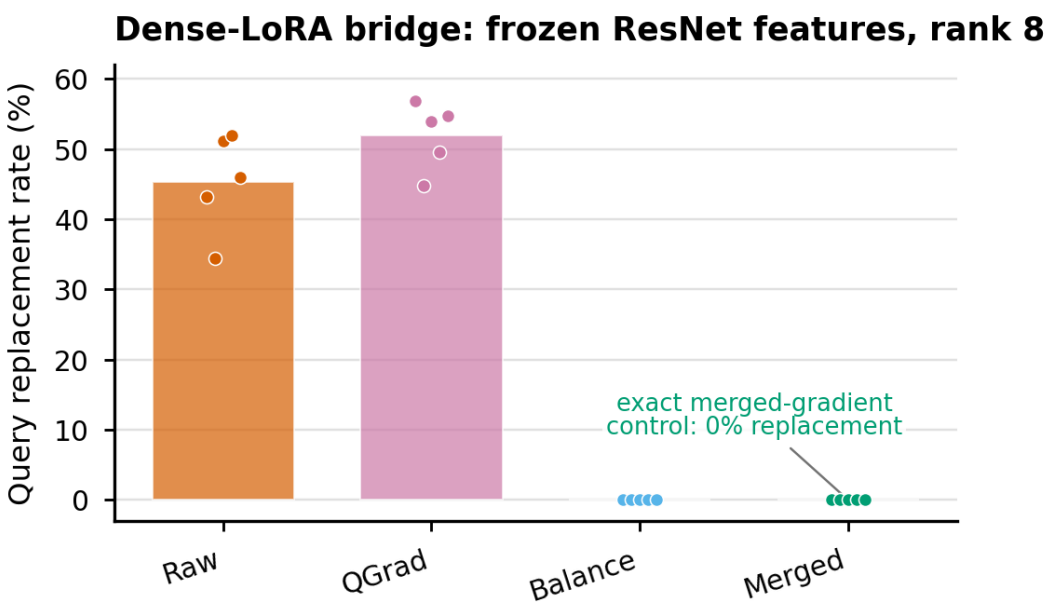

Figure 6: **Dense LoRA basis changes expose the boundary of elementwise factor invariance.** Five independently trained rank-8 adapters are evaluated under ten dense $GL(8)$ basis changes per model using 500 candidates and query budget 25. Raw factor-coordinate selection replaces 45.4% of queries on average, and factor-QGrad replaces 52.0%. In contrast, balanced factors and exact merged-gradient geometry yield identical selections under the evaluated basis changes. The relative Gram error of the exact invariant control is below 8.5×10^{-15} and the maximum logit discrepancy is 4.26×10^{-14}. Dense coordinate mixing therefore lies outside the invariance provided by a purely elementwise factor-times-gradient representation.

in the selection distribution, and whether large identity turnover necessarily corresponds to unusually large changes in batch properties. We also report orthogonal and zero-turnover controls that delimit the claim that reparameterization *can*, rather than must, change a selection.

D.1 Pathwise Versus Marginal Selection Variability

Our primary notion of identifiability is pathwise:

$$S(\theta, \mathcal{D}, \xi) \stackrel{?}{=} S(g \cdot \theta, \mathcal{D}, \xi)$$

for the same realized random tape ξ. For a stochastic selector, this is stronger than asking whether the two parameterizations induce the same long-run inclusion frequencies.

To examine the latter distinction for BADGE, we estimate candidate-level marginal inclusion frequencies from two equally sized Monte Carlo blocks of 256 selector tapes. The same number of tapes is used in the natural and reparameterized conditions, and a natural-versus-natural split-half comparison provides a Monte Carlo baseline at the same precision. Ten independently trained models are used for each of MNIST, Fashion-MNIST, and CIFAR-10; reparameterizations remain nested within model.

The estimated total-variation discrepancy between the natural and reparameterized marginal inclusion frequencies is 8.1%, 9.7%, and 12.4% on MNIST, Fashion-MNIST, and CIFAR-10, respectively. The corresponding natural split-half baselines are 7.6%, 8.9%, and 9.7%. Thus the excess over equal-precision Monte Carlo variation is only 0.5, 0.8, and 2.7 percentage points.

These values are much smaller than the same-tape set replacement reported in the main experiments. The distinction is important: a fixed selector tape can follow a different discrete path after reparameterization even when the estimated marginal inclusion frequencies change only modestly.

This experiment does *not* establish equality of the full distribution over unordered query sets. Candidate-wise inclusion frequencies are only one-dimensional marginals, and the full batch law may differ through higher-order dependence even when these marginals are similar. We therefore use this result to narrow the empirical claim to strong pathwise non-identifiability, rather than to assert a comparably large distributional shift.

D.2 Independent Selector Randomness

BADGE contains stochastic D^2 sampling after its deterministic max-norm first center. Large same-tape replacement could therefore appear practically less distinctive if independent selector runs already produce very different batches. We quantify this baseline directly on MNIST and Fashion-MNIST using ten independently trained models per dataset and eight independent tape pairs per model.

We compare three quantities:

1. the natural and reparameterized models under the *same* tape;
2. two independent tapes at the natural parameterization; and
3. independent tapes at the reparameterized parameterization.

The same-tape reparameterization replacement is 53.4% on MNIST and 63.4% on Fashion-MNIST. By comparison, two independent natural BADGE runs replace 63.3% and 72.3%, while two independent reparameterized runs replace 63.5% and 73.1%. The additional independent-draw dispersion associated with using the reparameterized model is therefore only 0.2 and 0.5 percentage points in these two settings.

The deterministic first center also changes under reparameterization in approximately 38% of MNIST models/comparisons and 50% of Fashion-MNIST models/comparisons. Thus the pathwise effect cannot be attributed solely to the stochastic part of D^2 sampling: the deterministic beginning of the selection path itself can change.

The interpretation is therefore asymmetric. Ordinary BADGE randomness is large enough that two independent runs may differ more than a paired reparameterization comparison. This does not remove the identifiability failure: holding the tape fixed, changing only the parameter representative still changes the realized query. It instead shows that reparameterization is not an unusually large additional source of *independent-run* dispersion in these experiments.

Because the independent-tape experiment uses eight tape pairs per model, we treat it as an operational calibration of set-level stochastic variability rather than as a high-precision estimate of the complete batch distribution.

D.3 Turnover-Matched Random Reselection

A second possible interpretation of large query replacement is that the new batch must differ from the original batch in an unusually meaningful way. Identity turnover alone does not establish this: in a redundant candidate pool, many alternative batches can differ substantially in membership while having similar class composition, difficulty, or downstream effect.

We therefore compare each observed reparameterized batch with random reselections constrained to match its turnover relative to the reference batch. For each observed comparison, 200 turnover-matched random reselections are generated. An additional $1{,}000$ random-pair baseline is used to calibrate the broader range of batch-to-batch variability.

We compare batch properties including label-distribution total variation, classification-error composition, predictive entropy, and prediction margin. Rather than asking whether the gauge-induced change is nonzero, we ask where its magnitude lies within the distribution produced by random reselection with the same number of replaced identities.

The matched-random controls are frequently at least as different as the gauge-induced batches. For BADGE, the gauge-induced label-TV changes occupy approximately the 29th, 35th, and 54th percentiles of the turnover-matched random distributions across the three evaluated datasets. The corresponding error-composition shifts lie near the 3rd, 11th, and 36th percentiles. Thus, in these

Table 4: **Geometry-preserving and zero-turnover controls.** MiniLM orthogonal basis changes leave the raw factor-coordinate Euclidean selection unchanged at all three evaluated ranks, while factor-QGrad remains basis dependent. The ModernBERT row is a held-out physical nonorthogonal writeback that nevertheless produces zero raw query replacement, providing an instance-level example that coordinate distortion need not cross a discrete selection boundary.

Setting	Transformation	Raw repl.	Factor-QGrad repl.
MiniLM, $r = 4$	orthogonal	0.0%	42.3%
MiniLM, $r = 8$	orthogonal	0.0%	33.1%
MiniLM, $r = 16$	orthogonal	0.0%	38.7%
ModernBERT, held-out instance	nonorthogonal physical	0.0%	–

summaries, gauge-induced turnover is not systematically associated with unusually large changes in these batch properties.

This negative result places an important boundary on the interpretation of query replacement. A 50% change in sample identity should not be read as a 50% change in semantic content, difficulty, or expected utility. The selection-decision discrepancy is real, but the severity of its consequences depends on what properties of the selected examples matter for the downstream application.

The control also helps interpret Figure 5: large identity replacement can coexist with small accuracy differences when many turnover-matched alternatives are similarly usable by the subsequent training procedure.

D.4 Orthogonal and Zero-Turnover Controls

The theory predicts two complementary boundaries. First, a coordinate transformation may preserve the raw Euclidean geometry exactly. Second, even a transformation that changes the geometry can remain inside the same discrete decision region for a particular model–data instance. We observe both cases.

Orthogonal LoRA basis changes. For the LoRA factorization, an orthogonal change of basis preserves the raw factor-coordinate Euclidean Gram geometry. MiniLM therefore provides a direct empirical control for the distinction between arbitrary dense basis changes and geometry-preserving ones.

Across ranks 4, 8, and 16, the raw factor-coordinate selection is unchanged under the evaluated orthogonal gauges. In contrast, the elementwise factor-QGrad representation replaces 42.3%, 33.1%, and 38.7% of the query batch, respectively. This contrast is informative because the orthogonal transformations have condition number one: factor-QGrad's failure cannot be attributed solely to ill-conditioning or extreme singular values. Rather, the elementwise construction singles out coordinate axes and is not invariant to dense orthogonal mixing.

A nonorthogonal zero-turnover instance. The converse boundary also occurs. In the held-out ModernBERT physical writeback study, five model instances yield raw query replacement rates of 76%, 24%, 48%, 0%, and 66%. One model instance therefore returns exactly the same raw query under the evaluated nonorthogonal reparameterization.

This zero-turnover case is consistent with the theory in Appendix A.4 and Appendix A.10. Nonorthogonality is sufficient to rule out universal *exact* Euclidean-Gram equality, but it is not sufficient to force a selection change for every realized set of covectors. The data may see only projective distortion on its occupied subspace, or a non-projective change may fail to cross the relevant decision boundary. Table 4 places the orthogonal and zero-turnover controls side by side.

We do not promote the separate conformal-scaling numerical check to an empirical result here because the corresponding canonical experiment artifact is not part of the final evidence set. The conformal boundary is instead established analytically in Appendix A.3 and Appendix A.5.

D.5 Negative and Boundary Result Ledger

Table 5 consolidates the principal results that limit stronger interpretations of query replacement. Each row is already analyzed in the referenced subsection; the ledger makes these outcomes visible together rather than treating them as isolated caveats.

Table 5: **Negative and boundary results.** These results preserve the pathwise identifiability finding while delimiting claims about marginal selection laws, excess stochastic variability, semantic batch change, downstream harm, and physical checkpoint turnover.

Analysis	Observed result	Claim boundary
BADGE marginal inclusion frequencies	Natural–reparameterized marginal TV is $8.1/9.7/12.4\%$ on MNIST/Fashion-MNIST/CIFAR-10, versus split-half Monte Carlo baselines $7.6/8.9/9.7\%$.	Same-tape set turnover does not establish a comparably large marginal-law shift (Appendix D.1).
Independent selector tapes	Same-tape replacement is $53.4/63.4\%$ on MNIST/Fashion-MNIST, whereas independent natural tapes replace $63.3/72.3\%$ and independent reparameterized tapes $63.5/73.1\%$.	Reparameterization is not an unusually large additional source of independent-run dispersion in these settings (Appendix D.2).
Turnover-matched reselection	Gauge-induced BADGE label-TV changes fall near the $29/35/54$th percentiles of matched random reselection; error-composition shifts fall near the $3/11/36$th percentiles.	Identity turnover alone does not imply unusually large semantic or difficulty changes (Appendix D.3).
Multiround downstream replication	An initial Fashion-MNIST result of -1.16 pp becomes $+0.27$ pp in the disjoint five-model, five-round experiment, while cumulative query divergence remains large.	The evaluated evidence supports trajectory non-identifiability, not a stable downstream-harm direction (Appendix H.2).
Physical ModernBERT writeback	Held-out raw replacement is $76/24/48/0/66\%$ (mean 42.8%), below the predeclared 50% mean-replacement gate; one checkpoint has zero turnover.	The 70.3% analytical factor-basis rate is not a universal physical-writeback rate (Appendix F.1).

E Modern PEFT Details

Section 6.1 establishes that selection non-identifiability appears in the evaluated end-to-end Transformer encoder PEFT settings. This appendix provides the corresponding rank-, architecture-, and continuation-level details. The experiments use LoRA basis transformations

$$(B, A) \mapsto (BR^{-1}, RA), \qquad R \in GL(r),$$

which preserve the represented update BA while changing the factor coordinates in which per-example gradients are expressed.

Throughout this section, reparameterizations are nested within independently trained model instances. Query replacement is therefore first aggregated over basis changes within a model, and model instances—rather than individual basis changes—form the independent statistical units.

E.1 MiniLM Across LoRA Ranks

We first examine whether factor-basis dependence is specific to one LoRA rank. The MiniLM-L12-H384 experiments use SNLI and insert LoRA modules into the query and value projections of all attention layers, giving 24 in-model LoRA modules. We evaluate ranks

$$r \in \{4, 8, 16\},$$

with five independently trained model instances per rank. For each model, selection is performed from 500 candidates with query budget 50. Nine nonorthogonal and three orthogonal factor-basis transformations are evaluated within each model.

Under the nonorthogonal transformations, raw factor-coordinate selection replaces 43.7%, 30.4%, and 36.9% of the query set at ranks 4, 8, and 16, respectively. Elementwise factor-QGrad does not remove the basis dependence: the corresponding replacement rates are 45.5%, 31.3%, and 37.6%.

Table 6: **Query replacement across LoRA settings.** All numeric entries are percentages of the query batch replaced after a function-preserving change of LoRA factor basis. Reparameterizations are nested within independent models or checkpoints and are averaged before reporting each setting. Dense LoRA uses exact merged-gradient geometry; the Transformer merged column uses the basis-invariant merged-sketch analytical control. Zeros therefore denote invariance to the evaluated factor transformation, not equivalence between the sketch and exact merged-gradient selection.

Model	Adapter	Units	Raw	Factor-QGrad	Balanced	Merged
Dense LoRA / ResNet features	rank 8	5	45.4	52.0	0.0	0.0
MiniLM	rank 4	5	43.7	45.5	0.0	0.0
MiniLM	rank 8	5	30.4	31.3	0.0	0.0
MiniLM	rank 16	5	36.9	37.6	0.0	0.0
ModernBERT	rank 8	8	70.3	72.3	0.0	0.0

The absence of a monotone rank trend is itself informative. The observed replacement is not determined by rank alone: rank changes both the dimension of the factor redundancy and the realized gradient geometry and decision margins. We therefore interpret the three ranks as evidence that the effect is not confined to one factor dimension, rather than as evidence for a monotone relationship between rank and selection sensitivity.

The balanced and merged analytical controls yield zero query replacement at all three ranks. These two controls remove the factor basis in different ways. Balancing selects a representative within the fixed-rank factorization class, whereas the merged construction forms a basis-invariant representation of the represented update. For these Transformer experiments, the latter is the fixed merged-sketch representation; it is exactly invariant to the audited basis transformations but is not identical to exact merged-gradient selection.

The orthogonal controls further distinguish raw Euclidean factor geometry from elementwise factor-QGrad. Because an orthogonal change of factor basis preserves the raw Euclidean Gram, raw factor-coordinate selection has zero replacement under the evaluated orthogonal transformations. Factor-QGrad, however, replaces 42.3%, 33.1%, and 38.7% of queries at ranks 4, 8, and 16. Thus its sensitivity to dense basis mixing is not merely a consequence of ill-conditioning; even a condition-number-one rotation can alter the coordinatewise multiplicative representation. The corresponding orthogonal-control analysis is discussed further in Appendix D.4.

E.2 ModernBERT Across Checkpoints

We next transfer the same analysis to ModernBERT-base. Unlike the MiniLM setting, which adapts attention query and value projections, the ModernBERT experiment places rank-8 LoRA modules in both attention and MLP projections: the 22 Transformer layers contribute 44 adapted fused-QKV and MLP-input projections. LoRA parameters and the classifier are trained jointly on SNLI.

We evaluate eight independently trained model instances. Each instance is tested under nine nonorthogonal factor-basis transformations, with the same candidate-pool and selection conventions used for the MiniLM experiments. The mean raw factor-coordinate replacement is 70.3%, while factor-QGrad replaces 72.3% of the selected examples. The corresponding reported model-level summaries are $[65.5, 74.3]\%$ and $[69.4, 75.1]\%$, respectively. Thus the dependence is present across the evaluated model instances rather than being attributable to a single checkpoint.

As in MiniLM, the balanced and merged analytical controls produce zero replacement under the audited basis transformations. The comparison therefore isolates the factor representation as the varying object: the underlying rank-8 update is unchanged, but the raw and elementwise factor-gradient geometries depend on the chosen internal basis.

Table 6 summarizes the PEFT experiments together with the dense-LoRA bridge from Appendix C.4. The dense-LoRA row uses exact merged-gradient geometry; the Transformer rows use the fixed basis-invariant merged-sketch analytical control. The two should therefore not be interpreted as identical computational constructions.

Table 7: **MiniLM continuation after rank-8 LoRA-based querying.** Accuracy differences are paired against continuation from the natural raw query and are reported in percentage points. The experiment uses five independent MiniLM model instances, three candidate-keyed annotation tapes per model, and one additional continuation epoch. These results are consequence checks rather than comparisons designed to establish a superior querying method.

Query condition	Accuracy difference (pp)	Reported interval
Reparameterized raw	$+1.01$	$[+0.51, +1.38]$
Balanced	$+0.85$	$[-0.01, +1.75]$
Merged	$+0.41$	$[-0.48, +1.49]$

The ModernBERT result should be interpreted as analytical evidence about factor-basis dependence, not as a prediction that every physical checkpoint rewrite will replace roughly 70% of the query. Materializing an equivalent factorization introduces numerical effects, and physical replacement varies substantially across model instances. We separate that question in Appendix F.1 and Appendix F.2.

E.3 Continuation Training

Large differences in query identity do not by themselves determine whether the resulting data have different downstream utility. We therefore perform a paired continuation experiment using the rank-8 MiniLM setting.

The comparison uses five independently trained MiniLM model instances and three annotation tapes per model. Annotation randomness is keyed by candidate identity: if the same candidate appears in two different query sets, it receives the same realized annotation outcome under the paired tape, independently of its position in the selected batch. This coupling prevents selection order from introducing an additional annotation-randomness difference. The full comparison contains 60 continuation fits, each adding one further training epoch after the queried data are incorporated.

We compare continuation from queries produced by the natural raw factor-coordinate selector with queries obtained after reparameterization and with the corresponding balanced and merged controls. Under the reported continuation condition, the raw reparameterized query differs from the natural raw baseline by $+1.01$ percentage points in downstream accuracy, with reported interval $[+0.51, +1.38]$. The balanced and merged conditions give $+0.85$ $[-0.01, +1.75]$ and $+0.41$ $[-0.48, +1.49]$ percentage points, respectively. Table 7 reports the paired continuation summary.

The number of noisy annotations is essentially unchanged across the paired conditions, so the observed accuracy differences are not accompanied by a corresponding systematic change in noisy-label count. More importantly, the direction of these continuation results does not support the hypothesis that large factor-basis-induced query turnover necessarily harms downstream accuracy. The raw condition is positive in this setting, while the balanced and merged intervals include little or no clear directional effect.

We do not interpret the positive raw result as evidence that parameterization-dependent selection improves training. The experiment uses five independent model instances and was designed to test the consequence of different query identities, not to compare selection algorithms for superiority. Together with the active-learning and Fisher experiments in Section 6.3, it instead supports the narrower conclusion that selection identifiability and downstream predictive utility can vary independently.

Finally, the continuation result should be distinguished from the invariance of the analytical controls. Zero replacement for balanced or merged geometry answers whether a factor-basis transformation changes the query. Continuation accuracy answers whether different queried data alter subsequent training. The former is an identifiability property; the latter is an empirical utility property, and neither implies the other.

Table 8: **Physical ModernBERT checkpoint reparameterization.** A balanced-SVD rewrite preserves the represented rank-8 LoRA updates while changing their factor coordinates. Raw replacement compares the factor-coordinate selection before and after the physical rewrite. Four of five held-out checkpoints change their selected examples, but the 42.8% mean misses the predeclared 50% gate and the magnitude is strongly checkpoint dependent. Analytical and physically recomputed balanced selections agree for all five checkpoints.

Checkpoint	Raw replacement	Physical/analytical balanced agreement
73	76%	yes
79	24%	yes
83	48%	yes
89	0%	yes
97	66%	yes
Mean	42.8%	$5/5$

F Numerical and Physical Reproducibility

The theoretical results in Section 5 concern exact function-preserving reparameterizations. In an implementation, however, an equivalent parameterization can be studied in two different ways. An *analytical* comparison applies the exact transformation law directly to the selector-consumed quantities, whereas a *physical* comparison writes the transformed parameters into a checkpoint and recomputes the forward and backward passes. The latter additionally contains floating-point materialization and recomputation error.

This distinction matters because selection is discrete. A numerically small perturbation of logits or Gram entries can leave the represented predictor effectively unchanged while crossing a score or sampling boundary of a sequential selector. We therefore separate three questions below: whether a physical equivalent checkpoint can change the query, whether the remaining analytical–physical discrepancy is explained by numerical precision, and whether the basis-invariant merged representation used in the Transformer experiments is exact or approximate.

F.1 Physical Checkpoint Reparameterization

The main ModernBERT results use analytical LoRA basis transformations in order to isolate the geometric consequence of factor non-uniqueness. We additionally test whether the effect survives when an equivalent factorization is materialized as an actual model checkpoint.

We use five held-out rank-8 ModernBERT model instances and rewrite each checkpoint using an equivalent balanced-SVD factorization of the same LoRA updates. The candidate pool contains 500 examples. After the rewrite, all selector-consumed quantities are recomputed from the materialized model rather than obtained by analytically transforming previously computed gradients.

The raw factor-coordinate query changes for four of the five checkpoints. Replacement rates are 76%, 24%, 48%, 0%, and 66%, giving a mean of 42.8%. This mean does not meet the predeclared 50% physical replacement gate. The variation is substantial: the physical effect is large for some checkpoints but absent for another. This result is therefore consistent with the instance-visible theory of Appendix A.4: a nonorthogonal coordinate change permits the selector-visible geometry to change, but does not require every realized query to cross a decision boundary.

The 0% case is scientifically useful rather than anomalous. It provides a physical example of the distinction made in Section 8: nonorthogonality rules out universal exact Euclidean-Gram invariance, but does not imply that every checkpoint and candidate pool must yield a different selected set.

Equally important, the balanced representation computed from the physical rewrite agrees with its analytical counterpart in all five held-out checkpoints. Thus the physical experiment supports the geometric interpretation while also showing that the magnitude of raw deployment-level turnover should not be identified with the larger analytical ModernBERT average in Figure 4.

F.2 Finite-Precision Effects

A physical reparameterization is only exactly function preserving over real arithmetic. In finite precision, storing the transformed factors and recomputing matrix products, activations, and gradients introduces a small perturbation. For a smooth score this perturbation may be negligible; for a discrete sequential selection path it can be amplified when a decision margin is small.

The ModernBERT precision diagnostics make this separation explicit. In the float32 physical rewrite, the maximum observed logit discrepancy is approximately

$$1.43 \times 10^{-4}.$$

No candidate changes its pseudo-label in the evaluated comparison. Thus the observed query differences cannot be attributed to a discrete change in the predicted class. Nevertheless, the recomputed gradient geometry is not bitwise identical to its exact analytical counterpart.

For a sequential D^2 selector, let d_i denote the current squared distance of candidate i to the selected center set and let

$$p_i = \frac{d_i}{\sum_j d_j}$$

be its sampling probability. A physical perturbation changes d_i to $d_i + \varepsilon_i$, thereby perturbing the cumulative distribution used with the fixed random tape. Even when every ε_i is small, the selected index can change if the fixed random number lies sufficiently close to a cumulative-probability boundary. The same principle applies to deterministic argmax decisions when the score gap is comparable to the numerical perturbation.

To distinguish this numerical sensitivity from an error in the analytical transport law, we recompute a reduced ModernBERT comparison using CPU float64. In this setting, the maximum discrepancy between directly recomputed and analytically transported Gram quantities is

$$5.63 \times 10^{-14},$$

and the resulting selections are identical. This closure supports the exact transport relation underlying the analytical experiments.

The finite-precision boundary is visible even for an invariant representation. For one held-out ModernBERT instance, the physically recomputed merged sketch differs from its analytical counterpart sufficiently to produce 34% query replacement after the checkpoint rewrite. This does not contradict gauge invariance of the analytical sketch: the analytical construction is exactly basis invariant, whereas the physical computation evaluates a slightly perturbed numerical model.

The relevant distinction can therefore be summarized as

$$\begin{array}{ccc} \text{exact reparameterization} & & \text{finite-precision checkpoint rewrite} \\ \Downarrow & \not\equiv & \Downarrow \\ \text{identical invariant geometry} & & \text{bitwise identical selection path.} \end{array}$$

Whether the second implication holds additionally depends on the numerical error and the margins of the discrete selector. We therefore use analytical comparisons to test the mathematical identifiability question and physical comparisons to characterize implementation-level reproducibility.

F.3 Exact and Approximate Merged Representations

The term *merged representation* refers to geometry defined from the represented LoRA update rather than from a particular factorization. This basis invariance should be separated from a second property: fidelity to the exact per-example merged-gradient geometry.

For a LoRA update

$$\Delta W = BA,$$

the factor transformation

$$(B, A) \mapsto (BR^{-1}, RA)$$

leaves ΔW unchanged. A statistic computed exactly from the merged update and its per-example differential therefore removes the internal $GL(r)$ basis freedom. The dense-LoRA experiment in

Table 9: **Roles of the merged representations used in the PEFT experiments.** Exact merged-gradient geometry removes the LoRA factor basis and serves as the reference data-selection geometry where computationally feasible. The fixed Transformer merged sketch is also basis invariant, but approximates that exact geometry. Physical recomputation adds a further finite-precision layer and therefore need not reproduce the analytical sketch bitwise.

Representation	**Factor-basis invariant**	**Exact merged-gradient geometry**	**Role in this work**
Exact merged gradient	yes	yes	Dense-LoRA reference and held-out fidelity reference
Fixed merged sketch	yes, analytically	no	MiniLM/ModernBERT invariant analytical control
Physically recomputed merged sketch	up to numerical realization	no	Finite-precision reproducibility diagnostic

Appendix C.4 uses such an exact merged-gradient representation. Its Gram geometry agrees across the evaluated basis changes to machine precision and its query replacement is zero.

Computing exact merged-gradient geometry throughout a Transformer is more expensive. The MiniLM and ModernBERT analytical experiments therefore use a fixed merged sketch constructed from basis-invariant merged quantities. The same sketch is applied before and after a factor-basis transformation. It is therefore exactly invariant to the audited LoRA gauge at the analytical level: if two factor pairs represent the same update, they produce the same sketched representation and hence the same analytical query.

However,

$$\text{gauge invariance} \quad \not\Rightarrow \quad \text{exact merged-gradient fidelity}.$$

The sketch can preserve the equivalence class while approximating the geometry that an exact merged-gradient selector would use. These are logically separate properties.

Table 9 distinguishes the three merged representations used throughout the PEFT experiments.

This distinction also explains why zero query replacement for the merged analytical control in Figure 4 should not be read as evidence that the sketch reproduces exact merged-gradient selection. The former comparison changes the *factor basis while holding the represented update fixed*; the latter comparison changes the *approximation used to represent the same merged-gradient geometry*.

We evaluate the latter question independently on five held-out ModernBERT checkpoints in Appendix G.1. Across sketch dimensions, the exact-query replacement ranges from roughly 10% to 28% on average and is strongly checkpoint dependent. That experiment therefore quantifies an approximation error that is orthogonal to the gauge-identifiability result.

Taken together, the three analyses in this section separate:

$$\underbrace{\text{factor-basis dependence}}_{\text{identifiability}}, \qquad \underbrace{\text{physical recomputation error}}_{\text{numerical reproducibility}}, \qquad \underbrace{\text{sketch–exact disagreement}}_{\text{approximation fidelity}}.$$

Conflating these effects would either overstate the exact invariance result or incorrectly interpret a finite-precision or sketching discrepancy as a failure of the underlying quotient geometry.

G Approximation Fidelity and Computational Cost

The merged representation used in the Transformer experiments is designed to remove dependence on the internal LoRA factor basis. Gauge invariance, however, does not imply that a computationally compressed representation reproduces the selection obtained from exact merged-gradient geometry. We therefore evaluate these two properties separately.

This section compares the fixed merged sketches used as invariant controls with an exact streamed merged-gradient reference on held-out ModernBERT checkpoints. The experiment addresses a prac-

Table 10: **Held-out merged-gradient fidelity–cost frontier.** Results use five held-out ModernBERT checkpoints, 64 candidates, and query budget 16. "Exact-query replacement" measures disagreement between the query returned by the merged sketch and that returned by exact merged-gradient geometry. Ranges are across the five checkpoints. Gram error is relative Frobenius error with respect to the exact merged Gram. Gram-only speedup measures the geometry-computation stage, whereas capture+Gram speedup also includes the shared feature/gradient capture cost. The sketches are invariant to the audited LoRA factor basis, but are not exact substitutes for the merged selector.

Per-side dim.	Exact-query replacement	Gram error	Gram-only speedup	Capture+Gram speedup	Storage
8	26.25% [12.50, 50.00]	0.125	6.1×	1.43×	0.7 MiB
16	27.50% [0.00, 56.25]	0.102	5.7×	1.44×	2.8 MiB
32	18.75% [6.25, 43.75]	0.086	4.4×	1.40×	11.0 MiB
64	11.25% [0.00, 18.75]	0.074	3.0×	1.33×	44.0 MiB
128	10.00% [0.00, 25.00]	0.053	1.8×	1.20×	176.0 MiB

tical question distinct from the identifiability result: how much exact-selection fidelity is retained as the merged representation is compressed, and what computational and storage costs are associated with increasing that fidelity?

G.1 Held-Out Merged-Gradient Fidelity

We evaluate five held-out ModernBERT checkpoints using 64 candidates and query budget 16. For each checkpoint, we construct the exact merged-gradient Gram matrix and compare its selected query with queries obtained from basis-invariant merged sketches. We vary the per-side sketch dimension over

$$d \in \{8, 16, 32, 64, 128\}.$$

The same candidate pool and selection convention are used for the exact and approximate geometries.

Table 10 reports both geometric and decision-level fidelity. At $d = 8$, the sketch differs from exact merged-gradient selection on 26.25% of the query on average, with checkpoint-level values ranging from 12.50% to 50.00%. At $d = 16$, mean replacement is 27.50% with an even wider range of $[0.00, 56.25]\%$. Thus increasing the sketch dimension does not produce a monotone improvement in the discrete selected set at the smallest dimensions.

From $d = 32$ onward, the mean replacement decreases to 18.75%, 11.25%, and 10.00% for dimensions 32, 64, and 128, respectively. Even the largest evaluated sketch is therefore not identical to exact merged-gradient selection. The checkpoint-level ranges also remain nonzero, reaching up to 25.00% replacement at $d = 128$.

The geometric approximation behaves more smoothly. Relative Gram error decreases monotonically from 0.125 at $d = 8$ to 0.053 at $d = 128$. The difference between this monotone geometric trend and the nonmonotone set-level behavior at small dimensions is consistent with the discrete nature of the selector: a smaller global Gram error need not preserve the selected batch if the perturbation crosses one or more decision boundaries.

We therefore interpret sketch dimension as controlling a *fidelity–cost tradeoff*, not as providing a threshold beyond which the approximate selector becomes exact. No post-hoc fidelity threshold is imposed in this analysis.

The held-out comparison is materially less optimistic than a single-checkpoint fidelity diagnostic. This is why all approximation claims in the paper use the five-checkpoint results in Table 10. The main conclusion is not that one sketch dimension is uniformly sufficient, but that exact-query fidelity is checkpoint dependent and improves only gradually as more of the merged geometry is retained.

G.2 Storage and Runtime Tradeoffs

Increasing the sketch dimension improves the geometric approximation but reduces the computational advantage. At $d = 8$, the Gram computation is $6.1\times$ faster than the exact merged reference and requires only 0.7 MiB of stored sketch data. At $d = 128$, the Gram-only speedup falls to $1.8\times$, while storage grows to 176 MiB.

These kernel-level numbers overstate the acceleration of the full selection pipeline because both exact and sketched representations require a shared feature and gradient-capture stage. In the held-out benchmark, the exact streamed merged-Gram computation takes 0.623 s on average, while the shared input/output capture itself takes approximately 1.050 s. Once this common cost is included, the measured capture-plus-Gram speedup is only $1.43\times$, $1.44\times$, $1.40\times$, $1.33\times$, and $1.20\times$ for sketch dimensions 8, 16, 32, 64, and 128, respectively.

The memory measurements exhibit the complementary tradeoff. Storage increases from 0.7 MiB at dimension 8 to 2.8, 11.0, 44.0, and 176.0 MiB as the per-side dimension increases. In the exact streamed computation, the largest temporary module embedding observed in this benchmark is approximately 432 MiB, and the measured peak resident memory is approximately 3.31 GiB.

These measurements suggest two different notions of computational advantage. The first is the cost of forming and manipulating the Gram representation once the required quantities have been captured; under this measure, low-dimensional sketches provide substantial acceleration. The second is the end-to-end cost of obtaining those quantities and constructing the geometry; under this measure, the shared capture stage substantially reduces the realized speedup.

Accordingly, we do not describe the low-dimensional merged sketches as drop-in system-level substitutes for exact merged-gradient selection. They provide a basis-invariant approximation whose operating point can be moved along a three-way tradeoff:

$$\text{selection fidelity} \quad \leftrightarrow \quad \text{runtime} \quad \leftrightarrow \quad \text{storage}.$$

Within the evaluated range, larger sketches reduce Gram error and generally improve exact-query agreement, but their storage cost rises rapidly and their runtime advantage narrows. Which operating point is appropriate therefore depends on whether the application prioritizes exact selected identities, memory, or selection-time computation.

H Downstream Utility and Negative Results

The central empirical result of this work concerns the identity of the selected examples. Downstream predictive utility is a separate estimand: two query sets can differ substantially while inducing similar trained models, or conversely a modest query difference can matter under a sensitive training regime. We therefore report the downstream comparisons without using query replacement as a proxy for effect size.

Across the final experimental families, we do not observe a consistent negative accuracy direction associated with parameterization-dependent selection. This statement is deliberately narrower than claiming that query identity is practically irrelevant. The selected examples determine which data are acquired, labeled, or retained, and the utility experiments cover only the tasks, budgets, and predictive metrics studied here.

H.1 Complete Downstream Utility Summary

Table 11 collects the downstream comparisons used to interpret the selection results. The independent unit is the trained acquisition model or checkpoint; repeated downstream fits or annotation tapes are paired within that unit before aggregation.

The clearest comparisons come from five-round active learning. After five rounds, the cumulative acquired sets differ by 47.9% on MNIST and 63.2% on Fashion-MNIST, yet the paired terminal accuracy differences are only $+0.01$ and $+0.27$ percentage points, respectively. The reported intervals include zero in both cases.

The pretrained-feature experiment gives the same qualitative picture. Reparameterized BADGE queries replace 60.3% of the selected CIFAR-10 examples, while the corresponding downstream

Table 11: **Downstream utility summary.** Accuracy differences are paired within independently trained acquisition models or checkpoints and are reported in percentage points. A dash indicates that the corresponding interval is not used as a headline summary in the paper. The comparisons differ in model, selector, data budget, and training procedure, so the table is a ledger of paired results rather than a pooled effect-size analysis.

Setting	Dataset	Comparison	Accuracy Δ (pp)	Reported interval
Five-round BADGE	MNIST	reparameterized vs. reference acquisition trajectory	$+0.01$	$[-0.20, +0.18]$
Five-round BADGE	Fashion-MNIST	reparameterized vs. reference acquisition trajectory	$+0.27$	$[-0.08, +0.68]$
Pretrained ResNet features	CIFAR-10	reparameterized raw vs. natural raw query	$+0.13$	–
Pretrained ResNet features	CIFAR-10	multiplicative invariant vs. natural raw query	$+0.24$	–
Pretrained ResNet features	CIFAR-10	canonical representation vs. natural raw query	-0.02	–
BAIT	MNIST	Fisher-diagonal repair vs. natural published BAIT	$+0.05$	–
BAIT	Fashion-MNIST	Fisher-diagonal repair vs. natural published BAIT	-0.42	–
BAIT	CIFAR-10	Fisher-diagonal repair vs. natural published BAIT	$+0.31$	–
MiniLM, rank 8	SNLI	reparameterized raw vs. natural raw query	$+1.01$	$[+0.51, +1.38]$
MiniLM, rank 8	SNLI	balanced-query continuation vs. natural raw	$+0.85$	$[-0.01, +1.75]$
MiniLM, rank 8	SNLI	merged-query continuation vs. natural raw	$+0.41$	$[-0.48, +1.49]$

accuracy difference is $+0.13$ percentage points. The multiplicative and canonical invariant constructions yield accuracy differences of $+0.24$ and -0.02 percentage points, respectively. These values do not support a consistent performance penalty associated with the large change in raw query identity.

The MiniLM continuation experiment provides a complementary Transformer setting. At rank 8, raw factor-coordinate selection changes by 30.4% under the evaluated nonorthogonal LoRA basis transformations. Continuation from the reparameterized raw query produces a $+1.01$ percentage-point accuracy difference relative to the natural raw query, with reported interval $[+0.51, +1.38]$. The balanced and merged continuation conditions give $+0.85$ and $+0.41$ percentage points, respectively. These positive values are not interpreted as improvements caused by non-identifiability; the experiment was not designed as a selector-superiority comparison.

Finally, the BAIT downstream experiments show mixed directions. Relative to the natural published BAIT selection, the Fisher-diagonal covariant construction changes downstream accuracy by $+0.05$, -0.42, and $+0.31$ percentage points on MNIST, Fashion-MNIST, and CIFAR-10, respectively. The sign therefore varies across datasets.

The table should not be interpreted as evidence that the effects are centered exactly at zero across tasks. Rather, the scientifically supported conclusion is that the evaluated settings do not show a stable negative direction that tracks query replacement. This is why Figure 5 is descriptive and does not fit a pooled regression or correlation across the heterogeneous experimental families.

H.2 Multiround Nonreplication

The multiround experiments also contain an informative negative result. An initial smaller experiment used two acquisition models and three active-learning rounds. In Fashion-MNIST, that comparison produced a terminal raw accuracy difference of

$$-1.16 \text{ percentage points},$$

Table 12: **Small-sample multiround result versus the larger experiment.** The negative Fashion-MNIST accuracy effect observed in the initial two-model, three-round comparison does not reproduce in the disjoint five-model, five-round experiment. This comparison is reported explicitly to separate the stable query-identifiability result from an unstable downstream harm interpretation.

Dataset	Experiment	Models / rounds	Terminal Δ (pp)
MNIST	smaller	$2/3$	$+0.59$
MNIST	larger	$5/5$	$+0.01$
Fashion-MNIST	smaller	$2/3$	-1.16
Fashion-MNIST	larger	$5/5$	$+0.27$

with reported interval

$$[-1.35, -0.96].$$

Taken in isolation, this result would suggest that parameterization-dependent acquisition can produce a substantial negative downstream effect.

The larger experiment does not reproduce that direction. Using a disjoint set of five acquisition models, five rounds, and 250 newly acquired examples per round, the Fashion-MNIST terminal difference becomes

$$+0.27 \quad [-0.08, +0.68] \text{ percentage points.}$$

The same larger experiment gives

$$+0.01 \quad [-0.20, +0.18]$$

on MNIST. The corresponding cumulative query replacement remains large: 47.9% on MNIST and 63.2% on Fashion-MNIST after five rounds. Table 12 records the initial and larger experiments together.

This nonreplication changes the appropriate scientific interpretation. The query trajectories themselves remain strongly different, so it does not weaken the identifiability result. It does weaken a stronger consequence claim that such divergence reliably degrades predictive accuracy. Several mechanisms could produce small or variable utility effects—including interchangeable examples, acquisition budget, and training variability—but the present experiments do not isolate which mechanism is responsible.

For this reason, the main paper reports the larger multiround experiment as the downstream evidence and treats the earlier negative result as a useful boundary on claims of stable harm.

H.3 BAIT Downstream Results

We separately evaluate downstream consequences for the Fisher-based setting. These experiments use five independent acquisition models per dataset, three paired downstream training seeds within each acquisition model, 500 initial labeled examples, a candidate pool of $1{,}000$, and query budget 50.

Under a function-preserving coordinate change, published BAIT with its fixed identity prior replaces approximately 9.6–12.8% of the query batch across MNIST, Fashion-MNIST, and CIFAR-10. The corresponding label-distribution total variation is approximately 4.8–6.8%. The Fisher-diagonal construction is invariant under the evaluated positive coordinate-rescaling family and therefore removes this representation dependence.

The downstream question is different: does using the covariant construction instead of the natural-coordinate published BAIT query degrade the resulting model? The paired overall-accuracy differences are

$$+0.05, \qquad -0.42, \qquad +0.31$$

percentage points on MNIST, Fashion-MNIST, and CIFAR-10, respectively. All three remain above the -1 percentage-point non-catastrophic margin used for this experiment. The directions are mixed, so these data provide neither evidence of systematic improvement nor evidence of systematic degradation.

A related comparison asks about the downstream effect of applying the reparameterization while retaining the published fixed identity prior. On CIFAR-10, the gauged published-BAIT condition differs from the natural published condition by

$$-0.29 \quad [-0.45, -0.15]$$

percentage points. Although the reported interval lies below zero, there are only five independent acquisition models. With this number of independent units, the smallest attainable two-sided sign-test p-value is 0.0625. Accordingly, we do not interpret this comparison as establishing a reliable negative effect.

The same caution applies more strongly to class-specific outcomes. The worst-class accuracy intervals are wide, and the experiments do not contain protected-group variables. They therefore do not support claims about fairness improvement or degradation.

Taken together, the BAIT results reinforce the separation between mechanism and consequence. The fixed identity prior provides a clear, theoretically characterized source of query non-identifiability, and covariant regularization restores the corresponding query path in the audited transformation family. Neither fact implies that replacing the query with its covariant counterpart must improve predictive performance. In the present experiments, the utility differences are small and directionally mixed.

I Scope and Boundary Cases

The invariance statements in this paper are relative to a specified transformation family, a specified selector-consumed parameter block, and the model state required to preserve the audited loss. Several boundary cases are therefore important for interpreting the results. In particular, an invariant representation need not remain informative on singular parameter strata; stateful layers enlarge the object that must be transported; exact scale-equivariance need not imply bitwise equality near numerical decision boundaries; and the LoRA arguments apply on a fixed-rank regular stratum rather than across rank-changing configurations.

Table 13 summarizes these distinctions. The purpose of these cases is not to weaken the main identifiability result, but to specify precisely which object is invariant and under which assumptions.

I.1 Zero-Parameter Strata

For positive coordinate rescaling, a useful invariant representation is the coordinatewise product

$$q_i = \theta \odot z_i,$$

where z_i is the corresponding per-example covector. If

$$\theta' = D\theta, \qquad z_i' = z_i D^{-1},$$

with positive diagonal D, then coordinatewise cancellation gives

$$q_i' = q_i.$$

This algebraic identity continues to hold when some coordinates of θ are zero. Zero coordinates therefore do *not* make the positive rescaling map locally noninvertible: the Jacobian D remains invertible.

The difficulty is instead informational. If

$$\theta_j = 0, \qquad z_{ij} \neq 0,$$

then

$$q_{ij} = \theta_j z_{ij} = 0.$$

The multiplicative representation can therefore erase a nonzero first-order gradient component even while remaining invariant to the rescaling symmetry. The corresponding log-magnitude coordinate

$$u_j = \log |\theta_j|$$

is also undefined at zero, and a diagonal geometry based on $\mathrm{Diag}(\theta^2)$ becomes singular there.

We test this boundary directly by progressively setting classifier coordinates to zero. Across ten independently generated model instances and ten rescalings per instance, the raw-gradient representation retains rank 24 across the evaluated zero fractions, whereas the multiplicative representation decreases from rank 24 to rank 0 as the classifier becomes fully zero. At the fully zero endpoint, all multiplicative representations collapse, with reported collision fraction 1.

The conclusion is therefore

$$\text{invariance} \not\Rightarrow \text{informativeness}.$$

The multiplicative construction is an exact invariant for the positive coordinate-rescaling action, but it should not be interpreted as a globally information-preserving or maximal representation of the predictor. Zero strata are a boundary of the log-coordinate and positive-definite metric interpretations, not a failure of invertibility of the rescaling itself.

I.2 BatchNorm State Transport

Function preservation is a property of the complete model state relevant to the loss, not necessarily of the trainable parameter tensor alone. Stateful normalization provides a concrete example through BatchNorm (Ioffe & Szegedy, 2015).

For a plain positively homogeneous ReLU block, rescaling the incoming parameters and compensating the outgoing parameters is sufficient to preserve the represented function. When BatchNorm is present, however, the running statistics used at evaluation time are part of the represented state. Applying the parameter transformation while leaving this state unchanged need not produce an equivalent predictor.

In the evaluated BatchNorm construction, transporting the relevant running state together with the parameters in evaluation mode reduces the maximum function discrepancy to

$$3.24 \times 10^{-16},$$

and the corresponding multiplicative-gradient selection is unchanged. By contrast, applying the plain-ReLU compensation while leaving the BatchNorm state untreated produces a logit discrepancy of approximately

$$1.44 \times 10^{-1}.$$

The latter comparison is therefore not a valid function-preserving orbit.

Training mode introduces an additional boundary. With the standard nonzero BatchNorm stabilization constant, the evaluated rescaling leaves a residual function discrepancy of approximately

$$4.87 \times 10^{-4}.$$

Setting that stabilization term to zero restores exact scaling behavior in the same construction. Thus the train-mode BatchNorm map is not covered by the plain positive-homogeneity argument without accounting for the complete normalization rule.

These observations motivate the state condition used in Appendix A.1: the covector transport identity requires a function-preserving reparameterization of the loss in a neighborhood of the parameter point, with all relevant model state transported consistently. Parameter equality rules that ignore such state do not satisfy that assumption.

I.3 Rounding and Tie Boundaries

The exact theory distinguishes geometric invariance from numerical reproducibility. For example, if

$$K' = \alpha K, \qquad \alpha > 0,$$

then the exact max-norm/D^2 primitive, exact squared-Euclidean facility location, and unregularized positive OMP are invariant to this common scaling under fixed side information. This statement concerns mathematical comparisons before finite-precision rounding.

An implementation may introduce additional discontinuities. Two are particularly relevant here.

Rounding. The CRAIG and GradMatch implementations include fixed-decimal numerical rounding. A common rescaling can therefore move two mathematically equivalent scores onto different sides of a rounded comparison even though the exact selector is scale homogeneous.

Ties and near ties. Selection is locally constant only away from decision boundaries. For a deterministic comparison, an arbitrarily small perturbation can change the selected index when two scores are tied or nearly tied. For a randomized D^2 step, the analogous boundary is a cumulative-probability threshold relative to the fixed random tape. Small floating-point changes can therefore alter a later selection even when the predictor and the underlying exact geometry differ only at numerical precision.

This distinction also explains why exact projective invariance should not be read as a guarantee of bitwise equality across every hardware and numerical implementation. The CPU-float64 closure in Appendix F.2 recovers the analytical ModernBERT geometry to approximately 5.63×10^{-14} and yields the same selection, whereas lower-precision physical recomputation can cross sequential selection boundaries.

Accordingly, throughout the paper we separate

exact selector invariance

from

finite-precision path reproducibility.

The former is a mathematical property of the specified selection primitive; the latter additionally depends on arithmetic, rounding conventions, tie handling, and the realized decision margins.

I.4 Fixed-Rank LoRA Assumptions

The LoRA analysis concerns the non-uniqueness of a fixed-rank factorization

$$\Delta W = BA,$$

with

$$B \in \mathbb{R}^{d_{\text{out}} \times r}, \qquad A \in \mathbb{R}^{r \times d_{\text{in}}},$$

under the basis action

$$(B, A) \mapsto (BR^{-1}, RA), \qquad R \in GL(r).$$

All factor pairs on such an orbit represent the same merged update.

The quotient interpretation used in the paper follows the regular fixed-rank matrix-manifold setting (Mishra et al., 2014; Vandereycken, 2013; Absil et al., 2008) and assumes that the evaluated factorizations remain on the regular fixed-rank stratum. In particular, the rank-r analysis is intended for configurations in which the represented update retains rank r and the factorization does not undergo rank collapse. On this stratum, dense basis transformations change only the internal coordinates of the same low-rank update.

This restriction matters for the balanced and merged controls. The merged update remains a valid representation of the product itself, but statements about a smooth fixed-rank quotient, balanced representatives, and the corresponding factor-coordinate geometry are not claimed across rank-changing or singular configurations. Near or at rank collapse, additional stabilizers and non-uniqueness can appear, and the regular fixed-rank construction used in the present experiments need not extend without modification.

The same restriction applies to the scope of the empirical result. Our dense LoRA, MiniLM, and ModernBERT experiments use fixed-rank adapters and invertible $GL(r)$ basis changes. They establish that factor-coordinate data selection need not be invariant within those fixed-rank equivalence classes. They do not characterize rank-changing adaptation, pruning that changes the effective LoRA rank, or singular factor trajectories.

Finally, elementwise factor-times-gradient invariance should not be generalized from diagonal rescaling to the full LoRA basis group. A dense $GL(r)$ action mixes factor coordinates, and the Hadamard construction does not transform covariantly under that mixing. This is observed empirically in Appendix C.4 and Section 6.1; the merged and balanced constructions are used because they address the larger factor-basis equivalence rather than only coordinatewise scaling.

Table 13: **Scope and boundary cases for the invariance statements.** The table distinguishes failure of an invariance theorem from failure of an auxiliary representation, numerical realization, or regularity assumption. These boundaries specify the domain of the claims rather than implying that the central identifiability phenomenon disappears outside it.

Boundary	What changes	What remains valid	Implication
Zero parameter	$q = \theta \odot g$ can erase nonzero gradient information; log coordinate is undefined	positive-rescaling map can remain invertible and q remains algebraically invariant	invariance does not imply informativeness
BatchNorm state	parameters alone do not specify the full function-preserving action	equivalence is recovered in the evaluated eval-mode construction when running state is transported	model state must be included in the symmetry definition
Rounding / near tie	finite-precision comparisons can cross a discrete boundary	exact scale-homogeneity and congruence statements remain valid	analytical invariance does not guarantee bitwise path equality
LoRA rank collapse	regular fixed-rank quotient assumptions may fail	$GL(r)$ basis analysis applies on the full-rank fixed-rank stratum	no claim is made across rank-changing transformations

J Reproducibility and Provenance

This section records the experimental sources underlying the reported results and the mapping from those sources to the quantitative figures and tables. Raw experimental records are treated as the primary evidence. Derived summary files, plotting outputs, and validation reports are reproducibility aids rather than additional experimental replicates.

An important distinction is maintained throughout the release: superseded or corrected outputs are retained for provenance but are not mixed with the results used for scientific claims. In particular, all Transformer-LoRA results reported in the paper use the corrected gradient-transport experiments; earlier Transformer outputs are excluded from quantitative claims.

J.1 Canonical Experimental Sources

Table 14 lists the experimental families that support the final manuscript. The table groups files by scientific experiment rather than by every derived artifact. Where both raw records and summary files exist, the raw JSON records are the canonical numerical source and the summaries are deterministic aggregations.

The corresponding protocol documents specify the candidate pools, budgets, transformation families, selection conventions, and paired-randomness rules. For example, the faithful selector experiments are associated with the corrected BADGE selector protocol; the Transformer experiments use the corrected factor-gradient transport convention; and the held-out fidelity experiment has a separate protocol defining the exact reference and sketch dimensions.

Several earlier result families are intentionally not used as scientific evidence. The initial faithful-selector lineage used a different first-center convention for the clustering primitive; earlier severity and pretrained outputs inherit that superseded convention. The initial Transformer-LoRA outputs used an incorrect transpose in the analytical transport of the B-factor gradient and were replaced by the corrected Transformer results. Likewise, the single-checkpoint merged-fidelity diagnostic is not used as the general fidelity estimate because the later five-checkpoint comparison provides the appropriate held-out evidence.

Smoke tests, timing checks, schema validators, and derived summaries are not counted as additional experimental units. The exception is the explicitly reported smaller multiround experiment in Appendix H.2, whose scientific role is the observed nonreplication of an initially negative downstream result.

Table 14: **Canonical experimental sources used in the manuscript.** Raw result records are the primary numerical evidence. Protocols define the corresponding experimental comparison; summaries and validation outputs are derived artifacts and are not counted as independent replications. Superseded or withdrawn experimental lineages are excluded from the claims reported in the paper.

Experimental family	Canonical result source(s)	Role in the paper
Faithful BADGE / CRAIG / GradMatch	`results/(mnist,fashion,cifar10)_faithful_selectors_v2.json`	Primary selected-index evidence for the ReLU experiments; source of Figure 2.
Reparameterization severity and marginal analyses	`results/(mnist,fashion,cifar10)_severity_consequence_v2.json`	Strength curves, gauge-to-gauge comparisons, marginal-inclusion analysis, and turnover-matched random controls.
BAIT geometry	`results/(mnist,fashion,cifar10)_bait_nested_audit_v2.json`	Identity-prior versus covariant-prior mechanism analysis in Figure 3.
BAIT downstream	`results/(mnist,fashion,cifar10)_bait_downstream_v1.json`	Query-composition and paired downstream-utility comparisons.
Five-round active learning	`results/(mnist,fashion)_multiround_main_v1.json`	Repeated acquisition, cumulative replacement, learning curves, and terminal utility in Figure 5.
Multiround non-replication	`results/(mnist,fashion)_multiround_smoke_v1.json`	Initial smaller-sample downstream result reported in Appendix H.2.
Pretrained representation	`results/cifar10_resnet18_frozen_audit_v2.json`	Robustness to a fixed pretrained visual representation and paired utility.
Dense LoRA bridge	`results/lora_dense_gauge_audit_v1.json`	Dense $GL(r)$ basis-change analysis with exact merged-gradient control; Appendix Figure 6.
Training-stage analysis	`results/mnist_mlp_stage_(init,early)_nested_v1.json` and final trained-stage records	Initialization, early-training, and trained-model comparison in Appendix C.2.
Zero / BatchNorm boundary analysis	`results/boundary_audit_v1.json`	Zero-stratum informativeness and model-state transport boundaries.
BADGE randomness baseline	`results/(mnist,fashion)_badge_randomness_baseline_v1.json`	Same-tape versus independent-tape set variability.
MiniLM LoRA	`results/transformer_lora_snli_v4.json`	Corrected MiniLM rank sweep, orthogonal controls, and continuation training.
ModernBERT LoRA	`results/transformer_lora_modernbert_v4.json`	Corrected end-to-end ModernBERT analytical basis-change experiment.
Physical ModernBERT writeback	`results/modernbert_balanced_writeback_confirmatory_v3.json`	Held-out physical balanced-factor checkpoint reparameterization.
ModernBERT precision analysis	`results/modernbert_precision_diagnosis_v1.json` and `results/modern_peft_final_closure_v3.json`	Analytical-versus-physical and float32-versus-float64 numerical closure.
Held-out merged fidelity	`results/modernbert_heldout_merged_fidelity_v1.json`	Five-checkpoint exact merged-gradient fidelity, runtime, and storage analysis in Table 10.

J.2 Source-to-Figure Mapping

The quantitative paper assets are generated from the canonical sources above. Table 15 provides the direct source mapping. This mapping is also useful for distinguishing conceptual panels from empirically aggregated panels and for identifying which rows use exact versus approximate merged-gradient representations.

For the released plotting pipeline, the derived publication assets are constructed from the raw records rather than by reading previously generated summary tables. The publication bundle additionally

Table 15: **Source-to-asset mapping.** Each quantitative figure or table is mapped to the raw experimental family from which its values are derived. Reparameterizations and repeated tapes are aggregated within the independently trained model or checkpoint before cross-model summaries are formed. Conceptual or theorem-only panels do not introduce additional empirical observations.

Paper asset	Primary source(s)	Provenance note
Figure 1	Faithful-selector v2; corrected ModernBERT v4; five-round active-learning main	Conceptual panels are schematic; quantitative panels use the same aggregates as the corresponding detailed figures.
Figure 2	`results/(mnist,fashion,cifar10)_faithful_selectors_v2.json`	Five independent models per dataset; ten transformations are aggregated within model.
Figure 3	Theory constructions plus `results/(mnist,fashion,cifar10)_bait_nested_audit_v2.json`	Panels (a)–(b) are analytical; panel (c) is the exact BAIT experiment.
Figure 4	`results/transformer_lora_snli_v4.json`; `results/transformer_lora_modernbert_v4.json`	Only corrected Transformer gradient transport is used. Transformer merged values refer to the invariant merged sketch, not exact merged-gradient selection.
Figure 5	Multiround main; BAIT downstream; pretrained ResNet audit; corrected MiniLM continuation	Every plotted turnover–utility point is paired within one experimental protocol; heterogeneous settings are not pooled statistically.
Appendix Figure 6	`results/lora_dense_gauge_audit_v1.json`	Uses exact merged-gradient geometry rather than the Transformer merged sketch.
Table 1	Theory statements and selector definitions	Taxonomy table; no empirical averaging.
Table 6	Dense-LoRA v1; MiniLM v4; ModernBERT v4	Dense LoRA uses exact merged gradients; Transformer rows use the invariant merged sketch.
Table 10	`results/modernbert_heldout_merged_fidelity_v1.json`	Five held-out checkpoints; exact streamed merged Gram is the reference.
Table 8	`results/modernbert_balanced_writeback_confirmatory_v3.json`	Physical checkpoint rewrite rather than analytical gradient transport.
Table 11	Multiround main; BAIT downstream; pretrained ResNet; MiniLM v4 continuation	Repeated downstream fits and annotation tapes remain nested within the corresponding independent model.

stores machine-readable derived quantities and source hashes, allowing each aggregate to be traced back to its input records.

The query-replacement metric used throughout the asset pipeline is

$$1 - \frac{|S \cap S'|}{k},$$

for equal-cardinality batches. In particular, the plotted quantity is not $1 - J(S, S')$; its relation to Jaccard is given in Equation equation 18. The same aggregation convention is used across figures: gauges and other nested repetitions are summarized within a model or checkpoint before the independent model-level observations are used for uncertainty summaries.

The figure-generation pipeline also distinguishes sources that are permitted to contribute to publication assets from superseded experimental lineages. This prevents an earlier selector convention or withdrawn Transformer transport result from being silently mixed with the final quantitative figures.

J.3 Software and Model Versions

The experiments use PyTorch-based implementations for model training, automatic differentiation, and per-example gradient construction, together with standard numerical and machine-learning libraries for matrix operations, selection routines, and summary statistics. Exact arithmetic claims in the paper refer to the corresponding mathematical transformations; implementation checks additionally distinguish float64 computation from float32 physical checkpoint execution as described in Appendix F.2.

The pretrained/model variants that materially define the experimental settings are:

- **ResNet-18:** ImageNet-1K-v1 pretrained weights, evaluated as a fixed feature extractor for the pretrained visual experiments. Inputs are evaluated at the resolution specified in Appendix C.3.
- **MiniLM-L12-H384:** the Transformer encoder used for the SNLI LoRA experiments, with LoRA modules inserted into the query and value projections of all attention layers. Ranks 4, 8, and 16 are evaluated.
- **ModernBERT-base:** the Transformer encoder used for the rank-8 PEFT experiments. LoRA is applied to 44 fused-QKV and MLP-input projections across the 22 Transformer layers.