MEASURE BEFORE REUSE: CHECKPOINT GEOMETRY AND POST-TRAINING REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

How do associations between checkpoint geometry and benchmark changes depend on the model settings and candidate updates being compared? We study 20 related post-training endpoints from Qwen3-8B and Llama-3.1-8B through a retrospective analysis of reported results. Each endpoint has an update-norm ratio and an MLP cosine relative to a reference checkpoint difference. We separate target-task changes from the mean and minimum changes on three shared benchmark columns. In the full collection, the pooled norm–mean correlation is -0.79 and remains -0.66 after subtracting scenario means from the pooled ranks; the corresponding alignment correlation changes from $+0.46$ to $+0.03$. Removing additive endpoints raises the adjusted alignment correlation to $+0.58$. Removing both additive and merged endpoints leaves the pooled norm correlation negative (-0.68), while its adjusted counterpart is positive ($+0.22$). Thus, even a pooled correlation that survives endpoint exclusions can give an incomplete account of the within-setting relationship. The reported additive, merged, and LoRA cases further separate angular similarity, update scale, and behavioral change. These results identify candidate composition and outcome aggregation as important checks when interpreting a checkpoint-geometry diagnostic.

1 INTRODUCTION

A model update can improve one benchmark while degrading another. In the Qwen3-8B results examined here, an SFT endpoint gains 22.9 percentage points on GSM8K and loses 18.7 points on MMLU. A merged endpoint gains 43.4 points on GSM8K while limiting the MMLU loss to 4.6 points. These differences motivate a practical question: what can be learned from the update's weights before running the evaluations that expose the trade-off?

Checkpoint differences are natural objects to inspect. Open model families such as Qwen3 and Llama provide base and post-trained checkpoints (Yang et al., 2025; Grattafiori et al., 2024); a further update can be compared with the difference between such a reference pair. The update's relative norm measures its scale. A cosine measures how closely its direction resembles the reference difference. Both quantities can be computed from compatible weights without benchmark labels. Their behavioral interpretation is less straightforward. A base-to-instruct difference can combine many changes, and a positive cosine does not specify which of them matter for a particular benchmark.

Several results in our collection expose this ambiguity. A Llama additive update has positive MLP alignment, $a_{\mathrm{MLP}} = 0.3054$, but loses 56.0 points on MMLU. A Qwen LoRA adapter has a small cosine, $a_{\mathrm{MLP}} = -0.0024$, yet gains 41.8 points on GSM8K. The small cosine therefore coexists with a substantial behavioral change. These observations motivate examining the association between each geometric coordinate and the outcomes it is meant to describe.

We study the relationship using 20 reported endpoints: nine Qwen variants, including one LoRA adapter, and eleven Llama variants. The collection contains fine-tuned, regularized, merged, and prior-guided updates. It provides several ways to change a checkpoint, but only two model settings; the endpoints are related observations, not 20 independent tasks. We use the available benchmark summaries and the additional reported geometry digits in the accompanying figure data. The resulting analysis is retrospective. It does not measure performance on a newly held-out update.

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.

The main finding is that the apparent diagnostic relationship changes with the comparison set. Relative norm is negatively associated with the mean change on ARC, HellaSwag, and MMLU in the full collection and in each model setting. MLP alignment has a positive pooled association, but little association remains after accounting for the difference between the two settings. Removing particular endpoint families can strengthen a pooled correlation while changing its within-setting interpretation. Consequently, a sensitivity analysis that only recomputes pooled correlations can miss the dependence it is meant to check.

Using reported endpoint records, we compare pooled, within-setting, and scenario-adjusted rank associations. The same pooled sign can coexist with a different adjusted sign after changing the candidate set, while mean and minimum benchmark changes expose different regressions. The individual endpoints explain which descriptive conclusions depend on these aggregation choices. The study does not test held-out prediction or the incremental predictive value of combining norm and cosine.

2 Related work

Editing and combining checkpoint differences. Task Arithmetic represents changes in behavior through differences between fine-tuned and reference models (Ilharco et al., 2023). WiSE-FT interpolates weights to improve the robustness of fine-tuned models (Wortsman et al., 2022), and TIES addresses interference when combining task vectors (Yadav et al., 2023). LoRA represents an update with low-rank factors (Hu et al., 2022). These works motivate inspecting the same quantities used to construct an endpoint. Our analysis concerns the association between an existing endpoint's geometry and its benchmark changes. It does not introduce a new merging or adaptation algorithm.

Reference directions and prior-guided adaptation. Safe LoRA already uses the relationship between base and aligned model weights to construct a safety-related subspace (Hsu et al., 2024). LoX studies low-rank extrapolation for robustness against fine-tuning attacks on safety (Perin et al., 2025). DRIFT transfers reasoning priors for MLLM fine-tuning (Huang et al., 2026). These are direct precedents for interpreting and modifying updates relative to a checkpoint difference; the use of a reference direction alone is not our novelty. The projection and additive variants in the present collection are reported local variants with DRIFT labels, not a controlled reproduction of all claims in the published MLLM method. Moreover, the present benchmark results do not measure refusal or jailbreak resistance.

Predicting compatibility from geometry. Zhou et al. (2026) examine interpretable properties for predicting model merging success, including geometric and functional relationships. Thus, using geometry as a diagnostic also has existing precedents. Our focus is post-training regression relative to a reported pre-update baseline, with target and retained-benchmark changes shown separately. The specific question is whether associations across a collection of related endpoints persist within its model settings and across different candidate compositions. We do not benchmark a competing mergeability predictor, and the present results do not establish superiority over one.

3 Measurements and evidence

3.1 The checkpoint pairs have distinct roles

Let $\theta_{\mathrm{ref,pre}}$ and $\theta_{\mathrm{ref,post}}$ denote the checkpoints defining a reference difference, and let θ_{start} and $\theta_{\mathrm{updated}}$ denote the pair defining the assessed update:

$$\Delta_{\mathrm{ref}} = \theta_{\mathrm{ref,post}} - \theta_{\mathrm{ref,pre}}, \qquad \Delta_{\mathrm{upd}} = \theta_{\mathrm{updated}} - \theta_{\mathrm{start}}. \tag{1}$$

The reference pair supplies a comparison direction; the update pair is the change being assessed. The reference's post-trained endpoint need not be the start of the assessed update. Calling Δ_{ref} a *reference delta* avoids assuming that this difference isolates one capability. Table 1 separates the roles needed to connect geometry to a behavioral comparison. A public reference pair supplies Δ_{ref}, but its names alone do not determine the start or evaluation baseline of a locally produced endpoint.

Table 1: Checkpoint roles. Equalities between roles are properties to be verified for an endpoint, rather than consequences of sharing a model family. The available records specify the named reference pairs and endpoint labels, but not all immutable checkpoint identities.

Symbol	Role in the comparison
$\theta_{\mathrm{ref,pre}}$	Earlier checkpoint in the named reference pair.
$\theta_{\mathrm{ref,post}}$	Later checkpoint; its difference from the earlier one defines Δ_{ref}.
θ_{start}	Origin subtracted from the assessed endpoint to define Δ_{upd}.
$\theta_{\mathrm{updated}}$	Assessed endpoint: fine-tuned model, adapter-applied model, or merge.
$\theta_{\mathrm{eval,pre}}$	Baseline whose benchmark score is subtracted in the behavioral change.

Table 2: Reported model settings. "Target probe" identifies an evaluation column, not a verified training corpus. Checkpoint names identify the reference families; exact revisions and the per-endpoint start mapping are not present in the available records.

Setting	Reference pre $\longrightarrow$ post	Target probe	Endpoints
Qwen	Qwen3-8B-Base $\longrightarrow$ Qwen3-8B	GSM8K	9
Llama	Llama-3.1-8B $\longrightarrow$ Llama-3.1-8B-Instruct	TruthfulQA	11

For compatible tensors on an index set P, the relative norm and module cosine are

$$r_P = \frac{\|\Delta_{\mathrm{upd}}^{(P)}\|_2}{\|\Delta_{\mathrm{ref}}^{(P)}\|_2}, \qquad a_m = \frac{\langle \Delta_{\mathrm{upd}}^{(m)}, \Delta_{\mathrm{ref}}^{(m)} \rangle}{\|\Delta_{\mathrm{upd}}^{(m)}\|_2 \|\Delta_{\mathrm{ref}}^{(m)}\|_2}. \tag{2}$$

The denominators must be nonzero. Equation 2 specifies concatenation of the selected tensors. The archived r and a_{MLP} values are the two features analyzed below; their exact masks and aggregation cannot be checked without the original measurement code. A LoRA update can be formed as $(\alpha/k)BA$ on adapted tensors, where k is its rank. Restricting the reference to those tensors or including the full parameter set can change both norm ratio and cosine. We retain the reported values and also analyze the collection without this endpoint.

3.2 Behavioral outcomes and their aggregation

Let $S_b(\theta)$ denote the reported benchmark score and write $d_b = S_b(\theta_{\mathrm{updated}}) - S_b(\theta_{\mathrm{eval,pre}})$. The intended evaluation baseline is $\theta_{\mathrm{eval,pre}} = \theta_{\mathrm{start}}$. For example, a reported Qwen SFT row compares its assessed update with the named Qwen base-to-post-trained reference and separately reports benchmark changes against a pre-update model. Identifying the reference pair does not establish which checkpoint supplied that pre-update model. The source tables describe their differences as post minus pre; exact checkpoint identities linking the behavioral and geometric baselines are unavailable. We therefore analyze these reported differences without claiming to have independently verified that mapping.

If the two baselines differ, the reported quantity decomposes as

$$d_b = \underbrace{S_b(\theta_{\mathrm{updated}}) - S_b(\theta_{\mathrm{start}})}_{\text{change along the assessed update}} + \underbrace{S_b(\theta_{\mathrm{start}}) - S_b(\theta_{\mathrm{eval,pre}})}_{\text{baseline difference}}. \tag{3}$$

Only the first term follows the checkpoint pair defining Δ_{upd}. The second vanishes when the intended baseline equality holds. This distinction matters even when all checkpoints have the same architecture: a regression measured against a different baseline cannot automatically be attributed to the assessed update. The available records do not determine whether such a mismatch occurred; Eq. 3 specifies the identity that an end-to-end replication needs to check.

GSM8K (Cobbe et al., 2021) is the designated Qwen target probe and TruthfulQA (Lin et al., 2022) the Llama target probe. ARC, HellaSwag, and MMLU provide shared evaluation columns (Clark et al., 2018; Zellers et al., 2019; Hendrycks et al., 2021). We keep the target change $T = d_{\mathrm{target}}$ separate and define

$$G_3 = \frac{d_{\mathrm{ARC}} + d_{\mathrm{HSwag}} + d_{\mathrm{MMLU}}}{3}, \qquad W_3 = \min\{d_{\mathrm{ARC}}, d_{\mathrm{HSwag}}, d_{\mathrm{MMLU}}\}. \tag{4}$$

G_3 is an equal-benchmark mean change. W_3 is the minimum of the three changes, not a change in the minimum absolute score and not a worst-case guarantee on unmeasured tasks. Higher values indicate better retention for both summaries. A positive mean can coexist with a large negative minimum. Tables show $100d_b$ in score points: one point is 0.01 on the reported 0–1 scale, or one percentage point for accuracy. The exact TruthfulQA metric is unspecified, so we do not assume that its score is binary accuracy.

Equal weighting makes the aggregation explicit without reconstructing unknown sample-count weights. It does not make the three benchmarks a complete measure of general capability. The source labels the Llama ARC evaluation as ARC-c, while the Qwen table says ARC; the exact Qwen subset requires confirmation. Accordingly, cross-setting G_3 comparisons refer to the reported benchmark columns rather than a verified identical evaluation configuration. TruthfulQA is a truthfulness probe here, not evidence about model safety.

3.3 What is observed, and what can be reproduced

The 20 rows comprise eight non-LoRA Qwen endpoints, one Qwen LoRA endpoint, and eleven Llama endpoints. WiSE and Task Arithmetic each contribute pre-merge and merged endpoints in both settings; Llama also contributes three projection variants labeled by β_{neg}. The LoRA endpoint is counted once. Its separate case-study presentation does not create a third independent model setting. The records use the same reference-pair names within each setting and include endpoints labeled as related construction families; exact checkpoint identities and training-run relationships are unverified.

Benchmark differences are reported to three decimal places. The archived figure data retain additional digits of r and MLP cosine compared with the rounded tables. We join these records by endpoint identifier and preserve both representations. The source package includes the anonymous inventory and exact geometry-record excerpt, their file hashes, and code checking the join before regenerating the downstream summaries. Additional digits are not independently verified measurement precision. No synthetic planning results enter this analysis.

Original checkpoint files, per-example predictions, complete training manifests, and exact evaluation configurations are absent from the available package. The records also lack complete endpoint-level global and attention cosines and raw norms. We therefore do not repeat claims that MLP is the best module, that relative norm outperforms raw norm, or that the reported accuracy changes have been independently reproduced. Appendix B describes the evidence and missing metadata.

3.4 Descriptive comparison, not held-out prediction

We compute Spearman correlations with average ranks for ties, separately for the pooled collection and each model setting. To distinguish pooled association from between-setting differences, let $R_i(x)$ be the pooled rank of feature x_i, let $s(i)$ be the setting, and define

$$\widetilde{R}_i(x) = R_i(x) - \frac{1}{n_{s(i)}} \sum_{j:s(j)=s(i)} R_j(x), \qquad \rho_{\mathrm{adj}}(x,y) = \mathrm{corr}(\widetilde{R}(x), \widetilde{R}(y)). \tag{5}$$

This is a correlation of pooled ranks after removing setting means, equivalent to residualizing those ranks on a setting indicator. It is not an average of the separate setting correlations. The adjustment does not separate model architecture, training data, and reference choice, which are bundled together in each setting.

We repeat the analysis after removing additive endpoints, merged endpoints, or both. All these choices are retrospective sensitivity analyses. They neither create independent validation data nor identify a causal effect of an update mechanism. We report no endpoint-bootstrap confidence intervals or independence-based p-values. The available records are insufficient for a reliable grouped test of predictive improvement over model identity, update type, or raw update norm.

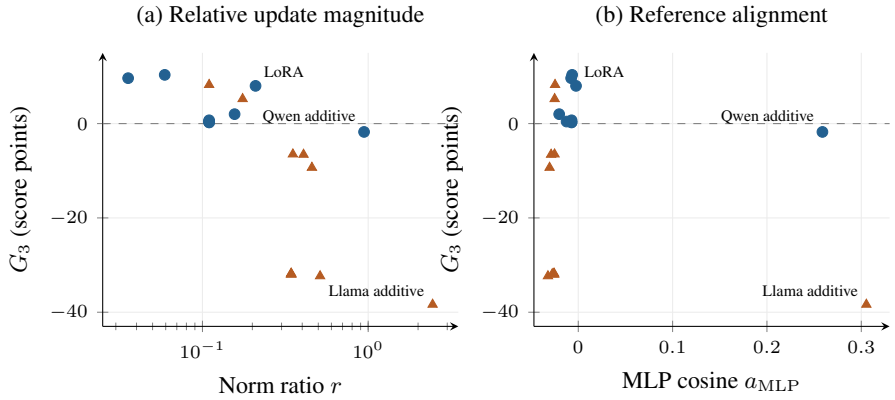


Figure 1: Reported geometry and mean benchmark change for all 20 endpoints. Blue circles are Qwen; orange triangles are Llama. The horizontal dashed line denotes no mean change. The norm axis is logarithmic; the cosine axis is linear. Additive endpoints have positive MLP alignment and negative G_3 in both settings. The narrow negative-cosine clusters contain endpoints with markedly different outcomes. No acceptance boundaries are fitted or drawn.

Table 3: Descriptive rank associations using the archived figure's reported geometry digits. n counts related endpoints. Scenario adjustment follows Eq. 5; unadjusted setting rows use ranks computed within that setting. All outcomes use the same reported benchmark differences.

		$\rho(\cdot, G_3)$		$\rho(\cdot, W_3)$	
Subset	n	r	a_{MLP}	r	a_{MLP}
All	20	-0.79	+0.46	-0.56	+0.16
Qwen (incl. LoRA)	9	-0.53	-0.11	-0.55	-0.22
Qwen (excl. LoRA)	8	-0.68	-0.25	-0.71	-0.41
Llama	11	-0.60	+0.21	-0.54	+0.17
All, scenario-adjusted	20	-0.66	+0.03	-0.52	-0.05

4 How candidate composition changes the signal

4.1 Pooled alignment largely reflects the two settings

Table 3 reports the main associations. Relative norm has a pooled correlation of -0.79 with G_3 and -0.56 with W_3. Its G_3 correlation is also negative within Qwen (-0.53, including LoRA) and Llama (-0.60). These results support a descriptive association between larger relative updates and worse mean retention in the full collection.

MLP alignment has a pooled G_3 correlation of $+0.46$, but its within-setting correlations are -0.11 in Qwen and $+0.21$ in Llama. The scenario-adjusted correlation is $+0.03$. Its association with W_3 is weaker still: $+0.16$ pooled and -0.05 adjusted. Thus, the positive pooled alignment result is insufficient evidence that higher alignment orders updates by retention within a setting. Excluding LoRA does not restore such an interpretation for Qwen: the G_3 correlation becomes -0.25.

Relative norm retains a negative scenario-adjusted association in the full collection (-0.66 with G_3 and -0.52 with W_3). This difference from alignment is informative. The geometry plot in Figure 1 shows both the broad pattern and the endpoint families that influence it.

4.2 Pooled robustness can coexist with within-setting reversal

Table 4 varies the candidate set while keeping the outcome definition fixed. Removing the two additive endpoints strengthens the pooled alignment correlation from $+0.46$ to $+0.77$; the adjusted correlation is then $+0.58$. These positive-alignment, poor-outcome endpoints are consequential

Table 4: Candidate-composition sensitivity of correlations with G_3. Removing a family changes both the ranks and the population being described. The lower block repeats every comparison without LoRA, whose parameter-mask convention is unresolved. A stable pooled sign need not imply a stable adjusted sign.

		Pooled		Scenario-adjusted	
Candidate subset	n	r	a_{MLP}	r	a_{MLP}
All	20	-0.79	+0.46	-0.66	+0.03
Without additive	18	-0.79	+0.77	-0.61	+0.58
Without merged	16	-0.68	+0.45	-0.33	-0.34
Without additive or merged	14	-0.68	+0.80	+0.22	+0.23
Sensitivity: excluding the LoRA endpoint throughout					
All	19	-0.80	+0.45	-0.67	+0.02
Without additive	17	-0.80	+0.76	-0.66	+0.58
Without merged	15	-0.67	+0.41	-0.38	-0.40
Without additive or merged	13	-0.69	+0.75	+0.11	+0.12

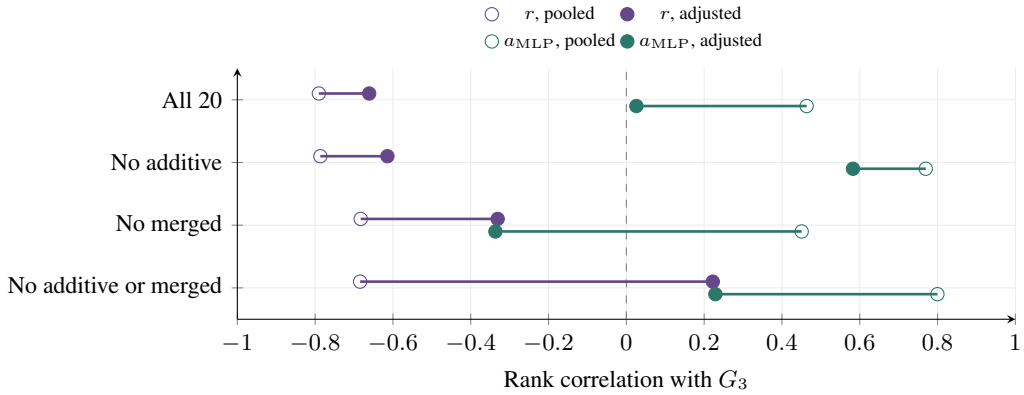


Figure 2: Pooled and scenario-adjusted correlations with G_3 for the four subsets in the upper block of Table 4. Hollow markers show pooled values; filled markers show adjusted values. Purple denotes norm ratio and green denotes MLP cosine. Lines connect two descriptions of the same subset, not training trajectories or confidence intervals.

counterexamples, not data errors to discard. The stronger association in their absence applies to a narrower set of update mechanisms.

Removing the four merged endpoints has a different effect. The pooled alignment correlation remains positive ($+0.45$), while the adjusted correlation becomes negative (-0.34). Relative norm remains negatively associated with G_3, but its adjusted magnitude falls from 0.66 to 0.33. The low-norm merged endpoints with favorable outcomes account for part of the full-collection pattern.

When both additive and merged endpoints are removed, the pooled norm correlation is still -0.68. Considered alone, this would suggest that the norm signal survives the removal of these families. Yet the adjusted correlation is $+0.22$, and the separate G_3 correlations are $+0.60$ for the six remaining Qwen endpoints and $+0.07$ for the eight remaining Llama endpoints. The negative pooled association in this subset does not describe a negative within-setting ordering. Alignment similarly has a strong pooled correlation ($+0.80$) but a much smaller adjusted one ($+0.23$). These coefficients describe the ordering of the available reported values. Small within-setting subsets are sensitive to rounding and ties; the positive adjusted values do not establish that larger updates improve retention.

These results locate an empirical limitation more precisely than saying that geometry sometimes fails. The apparent relationship depends on which model settings and construction families are compared. A collection can retain a strong pooled correlation after excluding unusual endpoints because the remaining settings occupy different regions of both feature and outcome space. Checking only pooled subset correlations does not detect this effect.

The remaining settings occupy separate ranges. The subset without additive or merged endpoints makes the source of the pooled pattern concrete. Its six Qwen records have norm ratios between 0.1094 and 0.2091, whereas the eight Llama records range from 0.3429 to 0.5126. Every Qwen G_3 is positive ($+0.3$ to $+8.0$ points), and every Llama G_3 is negative (-32.3 to -6.5 points). Consequently, the between-setting comparisons all associate the smaller norm with the better mean outcome, even though the separate within-setting correlations are nonnegative. The MLP cosine ranges are also separated: Qwen spans -0.020231 to -0.002380 and Llama spans -0.032148 to -0.025123. These ranges explain why knowing the model setting can account for a strong rank pattern in this subset. They do not identify which bundled difference between settings produces it.

4.3 The pattern persists when LoRA is excluded

The lower block of Table 4 repeats the analysis after removing LoRA throughout. With 19 endpoints, the norm–G_3 correlations are -0.80 pooled and -0.67 adjusted; the corresponding alignment values are $+0.45$ and $+0.02$. Thus, the near-zero adjusted alignment association in the full collection is not attributable to the one adapter record. Removing additive endpoints raises adjusted alignment to $+0.58$, while removing merged endpoints makes it negative (-0.40), following the same qualitative pattern as the original set.

Excluding additive and merged endpoints as well leaves 13 records. Their norm–G_3 correlation is -0.69 pooled and $+0.11$ adjusted; alignment changes from $+0.75$ pooled to $+0.12$ adjusted. The difference between pooled and adjusted descriptions therefore persists without the unresolved LoRA tensor mask. Its magnitude is smaller for this subset, and the positive adjusted coefficient should not be interpreted as evidence that increasing update size improves retention. The remaining records still share model settings and method families.

4.4 The outcome definition also matters

An update with a positive G_3 can still have a negative W_3. The Qwen SFT endpoint has $G_3 = +0.3$ points after rounding, despite an MMLU loss of 18.7 points. Gains on ARC and HellaSwag offset that loss in the mean. Even the Qwen merged endpoints with G_3 near $+10$ points retain an MMLU loss of 4.6 points. Reporting the target and the mean alone would hide that residual regression.

The full-collection norm association is weaker for W_3 than for G_3 (-0.56 versus -0.79). After removing additive and merged endpoints, its pooled correlation with W_3 is only -0.15. Association with mean benchmark change does not establish an equally strong association with the magnitude of the worst observed change. The individual changes in Tables 5 and 6 remain the relevant evidence when the use case places special value on one benchmark.

Across the reported Qwen records, all nine GSM8K changes are positive and all nine MMLU changes are negative. Eight of the nine G_3 values are positive, so positive average change is common despite the shared MMLU regression. Llama has two positive G_3 records, both labeled as merged endpoints, but only TaskArith merged has nonnegative changes on all three shared benchmarks. Overall, nine of the ten records with positive G_3 still have a negative W_3. This is a property of the reported benchmark columns; it does not depend on fitting a geometry threshold or choosing a particular weighting of evaluation items.

5 What the individual endpoints establish

Tables 5 and 6 report all 20 endpoints. Target and shared-benchmark entries are score changes in points ($100d_b$). G_3 and W_3 are recomputed from the three shared columns; geometry retains the additional reported digits for analysis and is rounded only for display. The tables make the individual changes visible alongside the coordinates used in the pooled correlations.

Positive alignment does not imply preservation. The additive Qwen and Llama endpoints have the only positive MLP cosines in this collection, approximately $+0.259$ and $+0.305$. Their minimum shared-benchmark changes are -20.3 and -56.0 points. Their relative norms are also the largest within each setting. These records show that positive cosine alone does not guarantee preservation on the reported benchmarks. They do not identify whether magnitude, direction, optimization details,

Table 5: All nine Qwen endpoints, including LoRA once. Target is GSM8K. All changes are in score points ($100d_b$); the source labels its ARC evaluation as ARC.

Endpoint	Target	ARC	HSwag	MMLU	G_3	W_3	r	a_{MLP}
SFT	+22.9	+15.9	+3.6	-18.7	+0.3	-18.7	0.110	-0.0071
L2-regularized	+25.8	+15.6	+3.8	-18.1	+0.4	-18.1	0.109	-0.0124
LoRA	+41.8	+18.6	+10.1	-4.7	+8.0	-4.7	0.209	-0.0024
WiSE pre	+23.3	+16.9	+3.6	-18.4	+0.7	-18.4	0.110	-0.0071
WiSE merged	+43.4	+23.7	+11.9	-4.6	+10.3	-4.6	0.059	-0.0065
TaskArith pre	+27.0	+14.9	+3.7	-17.8	+0.3	-17.8	0.110	-0.0071
TaskArith merged	+43.6	+22.0	+11.5	-4.6	+9.6	-4.6	0.036	-0.0076
Projection DRIFT	+24.3	+16.6	+4.1	-14.7	+2.0	-14.7	0.156	-0.0202
Additive DRIFT	+35.0	+12.9	+2.1	-20.3	-1.8	-20.3	0.944	+0.2588

Table 6: All eleven Llama endpoints. Target is TruthfulQA; the source labels the ARC column ARC-c. Changes are in score points ($100d_b$).

Endpoint	Target	ARC	HSwag	MMLU	G_3	W_3	r	a_{MLP}
SFT	-6.2	-15.3	-43.0	-37.6	-32.0	-43.0	0.343	-0.0257
L2-regularized	-9.5	-14.6	-42.7	-38.4	-31.9	-42.7	0.343	-0.0276
WiSE pre	-6.5	-14.2	-44.4	-36.7	-31.8	-44.4	0.343	-0.0257
WiSE merged	+3.1	+11.2	+5.3	-0.8	+5.2	-0.8	0.175	-0.0250
TaskArith pre	-4.4	-13.9	-44.7	-37.3	-32.0	-44.7	0.343	-0.0257
TaskArith merged	+3.3	+17.6	+6.4	+0.7	+8.2	+0.7	0.110	-0.0246
Projection DRIFT	-12.2	-14.2	-44.5	-38.3	-32.3	-44.5	0.513	-0.0321
Projection, $\beta_{\mathrm{neg}} = 0.3$	-26.3	-2.7	-7.3	-18.0	-9.3	-18.0	0.458	-0.0304
Projection, $\beta_{\mathrm{neg}} = 0.5$	-24.8	-1.0	-7.6	-11.1	-6.6	-11.1	0.407	-0.0288
Projection, $\beta_{\mathrm{neg}} = 1.0$	-17.4	-2.0	-6.6	-10.9	-6.5	-10.9	0.351	-0.0251
Additive DRIFT	-49.6	-32.2	-26.9	-56.0	-38.4	-56.0	2.448	+0.3054

or evaluation configuration caused the losses. A norm-matched directional intervention would be needed to isolate those alternatives.

Small cosine does not imply small behavioral change. The Qwen LoRA endpoint has $a_{\mathrm{MLP}} = -0.00238$, GSM8K gain $+41.8$, and MMLU change -4.7 points. The small reported cosine coexists with large benchmark changes. The unresolved reference support affects both norm ratio and cosine; this value does not establish weak alignment within the adapted tensors or allow a like-for-like comparison with full updates.

Reported merged endpoints have more favorable outcomes. For Qwen, the WiSE pre and merged endpoints have target changes of $+23.3$ and $+43.4$ points, and MMLU changes of -18.4 and -4.6 points. Their reported norm ratios are 0.1097 and 0.0592. For Llama, the corresponding G_3 changes are -31.8 and $+5.2$ points, with norm ratios 0.3430 and 0.1747. Their MLP cosines are close (-0.025678 and -0.025041), while their outcomes differ substantially. The Task Arithmetic pairs show a similar observed improvement.

These are comparisons of recorded endpoints. Their parent checkpoints, mixing coefficients, and selection histories are not recoverable from the available records. In particular, near-equal cosine with one reference does not establish that two update vectors are parallel. We therefore do not treat these pairs as controlled experiments that hold direction fixed, or attribute their benefits to a geometric routing rule.

Prior-guided variants do not yield a single retention rule. Qwen projection DRIFT has a smaller MMLU loss than SFT (14.7 versus 18.7 points), with negative MLP alignment and a larger relative norm. For Llama, the default projection endpoint loses 44.5 points on its worst shared benchmark. The β_{neg} records have less negative W_3 values, but TruthfulQA changes of -26.3, -24.8, and -17.4 points, versus -12.2 for the default projection endpoint. Better retention on the shared probes can coexist with a worse target outcome.

6 INTERPRETING A REFERENCE-RELATIVE SIGNAL

A norm is a scale, not a signed loss change. For a twice-differentiable loss L_b near a fixed starting checkpoint, a local expansion gives

$$L_b(\theta + \delta) - L_b(\theta) = \nabla L_b(\theta)^\top \delta + \tfrac{1}{2}\delta^\top H_b(\theta)\delta + o(\|\delta\|^2). \tag{6}$$

The norm constrains scale under additional derivative bounds, but does not determine the sign of this expression. A mechanistic interpretation of cosine would require a demonstrated relationship between the reference direction and the behavior under study. The expansion is an explanatory identity, not a fitted model of these endpoints: we measure neither gradients nor curvature, the benchmark score is not this smooth loss, and large updates need not remain in a local regime.

Reference magnitude and reference direction need different tests. For a fixed nonzero reference and a fixed tensor set P, $r_P = \|\Delta^{(P)}_{\text{upd}}\|/c$ with constant $c = \|\Delta^{(P)}_{\text{ref}}\| > 0$. This scaling leaves within-setting update-norm ranks unchanged. A within-setting correlation of r_P therefore cannot demonstrate a benefit from the reference's capability semantics. A normalization benefit would need a cross-setting comparison against raw norms under consistent tensor masks. A directional benefit would need comparisons against alternative compatible reference directions, with diagnostic performance as the outcome. Neither comparison is available in this collection.

Why pooled correlations can change interpretation. For centered pooled rank vectors X_i, Y_i, the cross-product decomposes as

$$\sum_i X_i Y_i = \sum_s \sum_{i \in s} (X_i - \bar{X}_s)(Y_i - \bar{Y}_s) + \sum_s n_s \bar{X}_s \bar{Y}_s. \tag{7}$$

The second term reflects between-setting differences. Keeping a large pooled cross-product after excluding an endpoint family does not ensure that the within-setting term has the same sign. This elementary decomposition explains what the adjusted statistic checks; it does not establish causality or remove confounding by update type.

Reported precision. The main analyses use the additional geometry digits embedded in the archived figure. Using the rounded inventory instead yields pooled G_3 correlations of approximately -0.79 for norm ratio and $+0.45$ for MLP alignment. Small subsets can be more sensitive to ties: after excluding LoRA, additive, and merged Qwen endpoints, only five rows remain, many with norm ratio rounded to 0.11. Their norm–G_3 correlation changes from about $+0.29$ at figure precision to $+0.73$ at table precision. Neither value supports a precise effect-size claim. The source package reports both precision conventions for all subsets.

7 LIMITATIONS AND CONCLUSION

The main limitation is the scope and provenance of the observations. There are two model settings, not a sample of independent downstream tasks. Model setting, named reference pair, and method labels are not crossed in a controlled design. The analysis uses rounded reported benchmark changes and archived geometry, with incomplete training and evaluation metadata. It cannot verify baseline identities, contamination controls, individual-example uncertainty, or end-to-end reproducibility. Alternative parameter masks, references, and independent seeds could change the observed associations. These limitations apply to both favorable and unfavorable results.

Within those limits, the analysis changes the interpretation of a checkpoint diagnostic. Relative norm has a strong pooled association with mean retention in the full collection, while MLP alignment's pooled association nearly disappears after removing setting differences. Removing endpoint families can preserve a pooled norm association while reversing its adjusted sign. The reported additive, merged, and LoRA endpoints place these correlations in context. In these records, checking only pooled subset correlations misses changes in the within-setting association. Candidate composition and outcome aggregation materially affect the descriptive interpretation of checkpoint geometry.

AI USE STATEMENT

Generative AI tools assisted with restructuring the manuscript, refining the hypotheses and retrospective analysis, interpreting results, implementing summary-statistic and figure-generation code, checking references, and editing prose. The mathematical discussion and proposed validation controls were developed with AI assistance. No AI-generated synthetic endpoint outcomes are included in the empirical results. The analysis scripts recompute the reported summaries from the supplied aggregate records; this does not substitute for checking the original training and evaluation artifacts.

REPRODUCIBILITY STATEMENT

The accompanying source package contains the endpoint summaries, their mapping to the archived geometry records, a source-verification script, and a Python script that regenerates the tables and vector-figure sources. The script requires only the Python standard library. The evidence available for this study, and the records needed for an end-to-end rerun, are described in Appendix B. Original model weights and per-example predictions are not included.

ETHICS STATEMENT

The results concern measured benchmark changes and do not certify deployment safety. TruthfulQA outcomes should not be read as evidence about refusal behavior or resistance to adversarial misuse. A geometry diagnostic used without task-specific evaluation could miss regressions; the findings here do not justify replacing those evaluations.

REFERENCES

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2009.03300.

Chia-Yi Hsu, Yu-Lin Tsai, Chih-Hsun Lin, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Safe LoRA: The silver lining of reducing safety risks when finetuning large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 65072–65094. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/77baa7c2a3a675823e89131698fd6e19-Paper-Conference.pdf.

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.

Chao Huang, Zeliang Zhang, Jiang Liu, Ximeng Sun, Jialian Wu, Xiaodong Yu, Ze Wang, Chenliang Xu, Emad Barsoum, and Zicheng Liu. DRIFT: Transferring reasoning priors for efficient MLLM fine-tuning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 20294–20309, San Diego, California, United States, jul 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1015. URL https://aclanthology.org/2026.findings-acl.1015/.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2212.04089.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, may 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Gabriel J. Perin, Runjin Chen, Xuxi Chen, Nina S. T. Hirata, Zhangyang Wang, and Junyuan Hong. LoX: Low-rank extrapolation robustifies LLM safety against fine-tuning, 2025. URL https://arxiv.org/abs/2506.15606.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7959–7971, June 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wortsman_Robust_Fine-Tuning_of_Zero-Shot_Models_CVPR_2022_paper.html.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 7093–7115. Curran Associates, Inc., 2023. doi: 10.52202/075280-0310. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1644c9af28ab7916874f6fd6228a9bcf-Paper-Conference.pdf.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, jul 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

Luca Zhou, Bo Zhao, Rose Yu, and Emanuele Rodolà. Demystifying mergeability: Interpretable properties to predict model merging success, 2026. URL https://arxiv.org/abs/2601.22285.

A RANK CALCULATIONS AND PRECISION

For n endpoints, average ranks assign tied observations the mean of their occupied rank positions. Spearman's coefficient is the Pearson correlation of those rank vectors. Reported benchmark-score entries have increments of 0.001. To avoid creating artificial differences through floating-point addition, the reproduction script uses decimal arithmetic and ranks $3G_3 = d_{\mathrm{ARC}} + d_{\mathrm{HSwag}} + d_{\mathrm{MMLU}}$, which has exactly the same ordering as G_3. Exact decimal arithmetic preserves ties in the reported summaries; it does not recover the ordering before the source benchmark values were rounded.

Scenario adjustment first ranks the full selected subset, then subtracts each scenario's mean rank separately from the feature and outcome. For example, the full-collection MLP–G_3 association is $+0.4635$ before and $+0.0258$ after this operation. Re-ranking within each scenario before centering defines a different descriptive statistic; it need not agree with the adjusted statistic in Eq. 5. Neither operation is a held-out prediction experiment.

The precision sensitivity and the within/between-setting decomposition are discussed in Section 6, including Eq. 7. The accompanying CSV retains both geometry-digit conventions for every analyzed subset.

B EVIDENCE PROVENANCE AND UNRESOLVED METADATA

The source is the anonymous endpoint record set archived with this submission. Its behavioral inputs are the reported S1/S2 scenario-table rows and the matching endpoint inventory. Copies appear in `data/sources/original_inventory.csv` and `data/sources/reported_table_rows.txt`. The geometric inputs are the 20 identifier–cosine–norm records in `data/sources/geometry_records.txt`, excerpted verbatim from the archived geometry figure. The manifest records SHA-256 hashes of these source files and excerpts; these are artifact hashes, not checkpoint hashes. Aliases such as WiSEFT_TRAIN and WISE_PRE are mapped explicitly in the accompanying data. The inventory and figure are two presentations of the same study, not independent replications.

Available records. The package contains target and shared-benchmark changes, rounded Overall changes, norm ratios, MLP cosines, endpoint labels, and limited case-study transition and null-control summaries. The latter are copied in `data/sources/auxiliary_reported_summaries.txt`; they are reported aggregates rather than raw draws or predictions. The original Overall statistic is described as pooling evaluation items. Because full sample counts and weighting metadata are unavailable, the present paper uses G_3 and W_3 rather than treating that pooled number as its primary outcome. It does not infer exact sample counts from rounded differences.

Records required to reproduce the original experiment. A complete release would identify the immutable revisions or hashes of both reference checkpoints; each assessed update's start, training initialization, and evaluation baseline; the actual training data and split; optimization settings and seeds; LoRA rank, scaling and target modules; merge parents, coefficients and candidate-selection procedure; geometry tensor masks and dtype; and benchmark configurations and per-example predictions. These details are not established by the available summary records. Configuration files describing planned extensions cannot supply missing historical settings.

Accuracy changes and sample transitions. If per-example correctness is available, define $y_i^{\mathrm{pre}}, y_i^{\mathrm{post}} \in \{0, 1\}$ using identical items and a fixed scoring rule. The lost-correct count is $F = \#\{i : y_i^{\mathrm{pre}} = 1, y_i^{\mathrm{post}} = 0\}$ and the newly-correct count is $B = \#\{i : y_i^{\mathrm{pre}} = 0, y_i^{\mathrm{post}} = 1\}$. For unweighted binary accuracy on the same N items, the change is $(B - F)/N$. The fraction F/N is not the fraction of initially correct items that are lost; the latter has a different denominator. The limited reported transition summaries are insufficient to reconstruct these quantities for all 20 endpoints, so we do not report a new forgetting-fraction correlation.

Nonrandom alignment is not validated diagnostic utility. The source reports no equally extreme draw among 500 random-same-norm and 500 layer-shuffled controls for each of three representative alignments. Those summaries address whether the observed angles resemble those particular null constructions. They do not show that the real reference predicts retention better than another reference,

or that reusing it improves an endpoint. The draws and tensor-level inputs are unavailable, so these null summaries are not used as independent statistical support for the new analyses.

An evaluation policy needs its own comparison. Computing geometry without labels is distinct from designing and validating a decision rule without access to the test outcomes. A retrospective partition followed by selection of the best evaluated candidate does not isolate the contribution of the partition. A useful test would freeze a rule on development lineages and compare it with fixed-probe evaluation and unconditional merge search on new lineages, using the same candidate sets, evaluation budgets, and target-retention criteria. No such prospective test is established by the available data.

C CONTROLS THAT WOULD DISTINGUISH THE REMAINING EXPLANATIONS

The following are proposed experiments, not completed results.

Fixed-direction scaling. For a known checkpoint pair, evaluate $\theta(\lambda) = \theta_{\mathrm{start}} + \lambda\Delta_{\mathrm{upd}}$ at several preselected positive λ. With fixed tensor support and reference, the norm ratio scales by λ and the cosine stays constant. This would measure how target and retention outcomes vary along one actual update direction. Existing merged endpoints should count as samples of this path only after their exact construction is verified.

Norm-matched direction controls. Holding the update norm fixed while changing its component along the reference would help distinguish angular effects from scale. Compatible alternative references and norm-matched random controls should be evaluated through their association with behavior, not merely whether their cosines differ. Such interventions still need identical baseline and evaluation configurations.