WHEN QUANTIZATION HELPS: CORRECTIVE COMPUTATION IN LOW-BIT LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-bit quantization reduces deployment costs by approximating high-precision weights, introducing changes that can improve some predictions. We investigate which activation changes support lower target-task loss than the high-precision reference and whether their useful structure can be shared across inputs. Experiments on Llama (GPTQ, AWQ, NF4) and Qwen (GPTQ) reveal systematic input-level improvements despite lower average performance over the full evaluation set. Improved inputs exhibit stronger loss-reducing contributions per unit activation displacement. Stronger first-order loss reduction can outweigh comparable or larger loss increases beyond the linear approximation. Targeted interventions link these changes to predictions: replacing selected quantized activations with their high-precision values can reverse naturally corrected answers; strengthening favorable changes or suppressing harmful ones repairs errors. We construct useful local adjustments from quantization-induced changes and find contrasting shared structure: a rank-four subspace retains 95.3% of their cross-entropy improvement on HANS, whereas MMLU retains 47.8% even at rank 128. Mechanism-guided activation adaptation raises accuracy above the matched high-precision model by 7.41 percentage points on HANS test data and 2.42 on MMLU development data. Its benefit over ordinary supervised adaptation is larger on HANS, consistent with its stronger shared structure. Together, these results connect the loss-reducing contributions of quantization-induced changes to their effects on predictions and relate their shared structure to the benefits of mechanism-guided adaptation.

1 INTRODUCTION

Low-bit quantization substantially reduces the memory and deployment cost of large language models (LLMs) by replacing high-precision weights with lower-precision approximations. Modern post-training quantization methods such as GPTQ and AWQ use second-order or activation-aware information to limit accuracy degradation, allowing large models to operate at substantially lower precision while preserving much of their original performance (Frantar et al., 2023; Lin et al., 2024). Accordingly, much of quantization research has focused on preserving the high-precision computation as faithfully as possible.

Yet closeness to the high-precision reference does not by itself determine whether a change helps or harms a prediction. A high-precision model can itself be wrong, so deviations may either increase or decrease the task loss. Existing analyses emphasize harmful deviations, including activation outliers and error propagation across layers (Dettmers et al., 2022; Xiao et al., 2023; Chang et al., 2025). We study the complementary phenomenon: individual inputs improve over the high-precision reference even when performance averaged over the full evaluation set declines. This raises three questions: what distinguishes beneficial quantization-induced changes from harmful ones, do the associated internal changes actually contribute to the resulting predictions, and when do beneficial changes contain structure that can be reused across inputs?

We first establish that beneficial quantization is a systematic, input-dependent phenomenon. On FineWeb(Penedo et al., 2024), improved examples recur under GPTQ, AWQ, and NF4 (Dettmers et al., 2023b) on both Llama and Qwen models. The same inputs improve under multiple quantizers more often than expected if improvements occurred independently. The effect also reaches actual decisions: on MMLU(Hendrycks et al., 2021), quantization corrects some high-precision errors

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.

while introducing more new errors overall. Correct-answer likelihood moves in the corresponding direction in the overwhelming majority of cases. Thus, individual improvements coexist with lower aggregate accuracy.

What distinguishes these outcomes? Improved inputs are not simply perturbed less: activation changes relative to the high-precision model have substantially overlapping magnitudes for improved and harmed inputs. A stronger distinction lies in what those changes do. Using the target-loss gradient, we form a local linear approximation to estimate each change's contribution. Improved inputs receive more loss-reducing contribution per unit activation displacement. At the final Transformer block's output, stronger first-order loss reduction can outweigh comparable or larger loss increases beyond this linear approximation. The key distinction therefore lies in how quantization-induced changes affect the target loss, beyond their magnitude alone.

We next test whether these loss-reducing changes contribute to actual predictions using activation interventions(Zhang & Nanda, 2024). For each input, we rank activation dimensions by the estimated contributions of their quantization-induced changes. Replacing values at the strongest loss-reducing dimensions with their high-precision counterparts can undo wrong-to-correct predictions. Conversely, strengthening favorable changes or removing harmful ones can repair errors. These interventions outperform random selections with the same number of dimensions and the predicted loss-change direction. Different dimension sets matched in total loss-reducing contribution recover most of the ranked intervention effect, linking prediction changes to the measured contribution.

Having established their input-level effect, we ask whether useful changes contain structure reusable across inputs. At a fixed layer, we construct useful adjustments by retaining loss-reducing components of quantization-induced changes and reversing loss-increasing ones. Projecting these adjustments onto shared directions learned from other inputs retains 95.3% of the full adjustments' cross-entropy reduction with four directions on HANS(McCoy et al., 2019), but 47.8% with 128 on MMLU. To relate this contrast to adaptation, we train small networks to predict activation adjustments from quantized representations, keeping the backbone frozen. With architecture, data, and training budget matched, mechanism guidance adds rewards for loss-reducing changes and penalties for nonlinear loss increases to task supervision. The resulting model exceeds the matched high-precision model by 7.41 percentage points in accuracy on HANS and 2.42 on MMLU, with a larger gain over ordinary supervised adaptation on HANS, consistent with its stronger shared structure.

Together, these results connect the loss-reducing contributions, causal effects on predictions, and shared structure of beneficial quantization-induced changes. They suggest looking beyond distance from high-precision computation to whether changes lower or raise task loss and how much useful structure they share across inputs.

2 RELATED WORK

Post-training quantization. Post-training quantization (PTQ) reduces LLM inference cost while attempting to preserve high-precision behavior. GPTQ uses approximate second-order information to reduce reconstruction error (Frantar et al., 2023), while AWQ protects salient weights identified from activation statistics (Lin et al., 2024). Other methods improve quantization robustness by reshaping difficult weight or activation distributions, including SmoothQuant (Xiao et al., 2023) and rotation-based methods such as QuaRot and SpinQuant (Ashkboos et al., 2024; Liu et al., 2025). These approaches primarily aim to reduce deviations from the high-precision model. We instead study what those deviations do to prediction quality, especially when they improve rather than degrade a prediction.

Quantization error and input dependence. Prior work has shown that quantization error is highly non-uniform. LLM.int8() attributes substantial difficulty to activation outliers (Dettmers et al., 2022), while SpQR identifies weights that contribute disproportionately to quantization error (Dettmers et al., 2024). Most closely related, Chang et al. (2025) show that quantization sensitivity varies strongly across inputs and analyze how harmful errors accumulate through internal representations. We study the complementary phenomenon: why some inputs improve after quantization. Rather than measuring deviation magnitude alone, we characterize whether quantization-induced changes increase or decrease the target loss, test their behavioral role through intervention, and examine whether beneficial changes contain reusable structure.

Low-rank adaptation and quantization compensation. LoRA and QLoRA use low-rank updates for parameter-efficient adaptation, including on frozen quantized backbones (Hu et al., 2022; Dettmers et al., 2023a). Methods such as LoftQ, LQ-LoRA, and RILQ further use low-rank structure to reduce or compensate for quantization-induced error (Li et al., 2024; Guo et al., 2024; Lee et al., 2025). Our study instead asks when beneficial quantization-induced changes exhibit shared low-dimensional structure and whether such cross-input regularity can support lightweight input-conditioned activation modification.

3 PRELIMINARIES

We compare a high-precision reference model (FP) with its quantized counterpart (Q) obtained from the same pretrained model.

Input-level effects. For an input–target pair (x, y), let $\ell_{\mathrm{FP}}(x, y)$ and $\ell_{\mathrm{Q}}(x, y)$ denote the target losses of the FP and Q models. We define

$$\Delta\ell(x, y) = \ell_{\mathrm{Q}}(x, y) - \ell_{\mathrm{FP}}(x, y). \tag{1}$$

An input is *improved* when $\Delta\ell < 0$ and *harmed* when $\Delta\ell > 0$. For multiple-choice tasks, we additionally record wrong-to-correct (W2C) and correct-to-wrong (C2W) transitions.

Representation changes. At layer l, let $\boldsymbol{h}_l^{\mathrm{FP}}(x)$ and $\boldsymbol{h}_l^{\mathrm{Q}}(x)$ denote the representations from the complete FP and Q executions. We define

$$\Delta\boldsymbol{h}_l(x) = \boldsymbol{h}_l^{\mathrm{Q}}(x) - \boldsymbol{h}_l^{\mathrm{FP}}(x), \qquad d_l(x) = \|\Delta\boldsymbol{h}_l(x)\|_2. \tag{2}$$

Here, $\Delta\boldsymbol{h}_l$ includes quantization effects accumulated through preceding layers, while d_l measures their magnitude.

Local effect on the target loss. For a fixed input and layer, we omit x and l for clarity. Let $L(\boldsymbol{h})$ denote the downstream target loss obtained by replacing the representation at this layer with $\boldsymbol{h}$ and continuing the remaining computation using the FP model. At the FP reference, we define

$$\boldsymbol{g} = \nabla_{\boldsymbol{h}} L(\boldsymbol{h})\big|_{\boldsymbol{h}=\boldsymbol{h}^{\mathrm{FP}}}, \qquad s = \boldsymbol{g}^{\top}\Delta\boldsymbol{h}. \tag{3}$$

Here, $\boldsymbol{g}$ is the local target-loss gradient at the FP reference, and s is the signed first-order loss change induced by the quantization displacement: $s < 0$ indicates locally loss-reducing movement, whereas $s > 0$ indicates locally loss-increasing movement. Thus, the displacement norm measures change size but does not determine whether loss increases or decreases, motivating the first-order analysis. Since finite changes can depart from this linear estimate, we also examine nonlinear effects.

4 EXPERIMENTAL SETUP

Models and evaluation. We compare high-precision Llama-3-8B(Grattafiori et al., 2024) with its GPTQ, AWQ, and NF4 counterparts, and high-precision Qwen2.5-7B-Instruct(Qwen et al., 2025) with its GPTQ counterpart, studying improvements from quantization alone. On FineWeb, we measure mean next-token negative log-likelihood (NLL). On MMLU, we evaluate all 14,042 questions with five-shot prompts, recording correct-answer loss and W2C/C2W transitions. For mechanism analysis on FineWeb, improved and harmed inputs are matched by the magnitude of their NLL changes for both models.

Sharing and adaptation. These experiments use a fixed Llama-3-8B GPTQ four-bit model and its matched high-precision reference. We use the pre-SiLU MLP gate output as the intervention boundary because it provides a channel-wise linear representation in which the direct local quantization change can be isolated with the gate input fixed. We modify only the final input position used for answer prediction, keeping the intervention localized. For Llama, we select zero-indexed layer 11 based on the layerwise intervention analysis reported in Section 6.3 and fix this site across tasks. For the diagnostic sharing analysis, target-loss gradients identify loss-reducing changes to retain and loss-increasing ones to reverse; the shared basis is estimated only from training-side examples and fixed for evaluation. The ordinary and mechanism-guided modifiers use the same frozen backbone, architecture, parameter budget, data protocol, and optimization budget, differing only in the added mechanism-derived supervision. At inference time, neither modifier uses the target answer, target-loss gradient, or high-precision model.

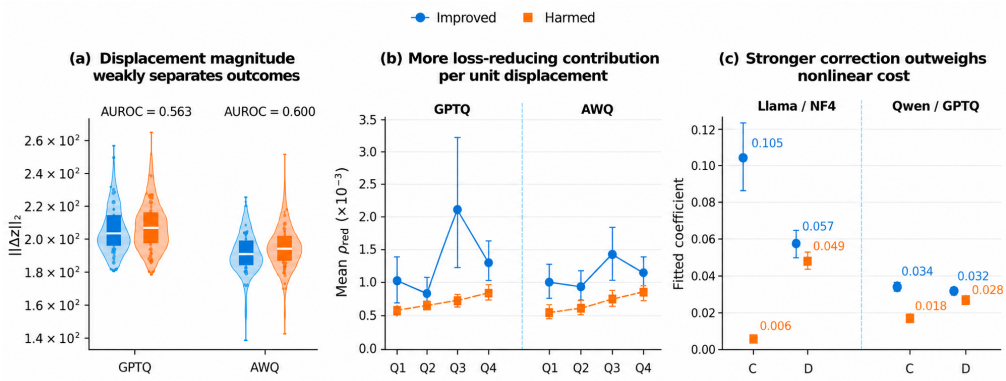


Figure 1: Mechanistic signature of beneficial quantization. **(a)** Displacement magnitude weakly separates improved and harmed inputs. **(b)** Improved inputs receive more loss-reducing contribution per unit displacement. Q1–Q4 are displacement-norm quartiles. **(c)** Improved inputs show stronger corrective movement C despite comparable or larger nonlinear term D. Error bars denote bootstrap 95% confidence intervals.

5 Why Do Some Inputs Improve After Quantization?

We begin by establishing that improvement after quantization is a reproducible input-level phenomenon, and then ask what distinguishes improved inputs internally. Following section 3, we consider three possibilities: improved inputs may undergo smaller activation changes, their changes may contain more movement that reduces the target loss, or they may incur smaller nonlinear costs.

5.1 Improvement Is Structured Across Inputs and Quantizers

We first ask whether improvement is an isolated outcome of a particular quantizer. Across separate 250,000-document FineWeb scans, improved examples consistently appear under GPTQ, AWQ, and NF4 on Llama and under GPTQ on Qwen, despite an overall increase in loss. Improvement is also structured across quantizers. On Llama, 62 documents improve under both GPTQ and AWQ, compared with only 1.495 expected from their marginal improvement rates under independence, a $41.48\times$ enrichment. The effect also reaches discrete predictions. On the full 14,042-question MMLU benchmark, Llama exhibits 602 wrong-to-correct (W2C) transitions (4.29%) and 1,004 correct-to-wrong (C2W) transitions (7.15%), while Qwen exhibits 412 W2C (2.93%) and 587 C2W (4.18%) transitions. The gold-answer likelihood moves consistently with these changes: it increases in 93.36% of Llama W2C cases and 99.52% of Qwen W2C cases, while decreasing in 97.41% and 99.83% of their respective C2W cases. Beneficial quantization therefore occurs systematically across quantizers and can alter actual predictions.

5.2 Improved Inputs Receive More Corrective Movement

A simple explanation is that improved inputs are perturbed less. This is already contradicted by the Llama layer-8 gate: as shown in Figure 1(a), improved and harmed inputs have strongly overlapping displacement magnitudes under both GPTQ and AWQ. Thus, smaller activation change is not necessary for improvement. We next ask what those changes do to the target loss. Let $\boldsymbol{z}^{\mathrm{FP}}$ and $\boldsymbol{z}^{\mathrm{Q}}$ denote the pre-SiLU MLP gate activations under the high-precision and quantized executions. Following equation 3, we define

$$\Delta \boldsymbol{z} = \boldsymbol{z}^{\mathrm{Q}} - \boldsymbol{z}^{\mathrm{FP}}, \qquad a_{tj} = g_{tj}\Delta z_{tj}, \qquad A_j = \sum_t a_{tj},$$
$$M_{\mathrm{red}} = \sum_j [-A_j]_+, \quad M_{\mathrm{inc}} = \sum_j [A_j]_+, \quad \rho_{\mathrm{red}} = \frac{M_{\mathrm{red}}}{\|\Delta \boldsymbol{z}\|_2}. \tag{4}$$

Here $[u]_+ = \max(u, 0)$, so that

$$\boldsymbol{g}^\top \Delta \boldsymbol{z} = M_{\mathrm{inc}} - M_{\mathrm{red}}, \tag{5}$$

while ρ_{red} measures loss-reducing contribution per unit displacement.

Figure 1(b) shows one representative example at the Llama layer-8 gate, where mean ρ_{red} is $1.95\times$ higher for improved inputs under GPTQ and $1.71\times$ higher under AWQ. Thus, improved inputs receive more loss-reducing contribution per unit displacement. Similar patterns hold across most layers (Appendix B).

5.3 Stronger Correction Outweighs Nonlinear Cost

The first-order analysis shows that improved inputs receive stronger loss-reducing movement, but finite displacements may incur nonlinear costs. To capture finite-displacement effects after quantization has propagated through the network, we analyze the output of the final Transformer block. For Llama-3-8B, this is zero-indexed layer 31; for Qwen2.5-7B, it is zero-indexed layer 27. Let $\boldsymbol{h}^{\mathrm{FP}}$ and $\boldsymbol{h}^{\mathrm{Q}}$ denote the FP and quantized representations at this boundary, respectively, and define

$$\Delta \boldsymbol{h} = \boldsymbol{h}^{\mathrm{Q}} - \boldsymbol{h}^{\mathrm{FP}}. \tag{6}$$

Using the FP-reference target-loss gradient $\boldsymbol{g}$ defined in Section 3, we decompose this displacement into components parallel and orthogonal to the local loss gradient:

$$\Delta \boldsymbol{h}_{\parallel} = \frac{\boldsymbol{g}^{\top}\Delta \boldsymbol{h}}{\|\boldsymbol{g}\|_2^2}\boldsymbol{g}, \qquad \Delta \boldsymbol{h}_{\perp} = \Delta \boldsymbol{h} - \Delta \boldsymbol{h}_{\parallel}. \tag{7}$$

The parallel component carries the complete first-order effect, $\boldsymbol{g}^{\top}\Delta \boldsymbol{h}$, whereas the orthogonal component has zero first-order contribution.

To measure finite-displacement effects without explicitly forming the Hessian, for each $\boldsymbol{v} \in \{\Delta \boldsymbol{h}_{\parallel}, \Delta \boldsymbol{h}_{\perp}\}$, we evaluate the downstream loss along the corresponding direction:

$$\Delta L_{\boldsymbol{v}}(\alpha) = L\big(\boldsymbol{h}^{\mathrm{FP}} + \alpha \boldsymbol{v}\big) - L\big(\boldsymbol{h}^{\mathrm{FP}}\big)\,. \tag{8}$$

Here, α scales the measured displacement along direction $\boldsymbol{v}$. We evaluate the downstream loss at $\alpha \in \{0, 0.25, 0.5, 0.75, 1.0\}$ and fit these five measured points with

$$\Delta L_{\boldsymbol{v}}(\alpha) \simeq a_{\boldsymbol{v}}\alpha + b_{\boldsymbol{v}}\alpha^2. \tag{9}$$

The median R^2 exceeds 0.998 across all three quantizers. We summarize the fitted loss-reducing movement along the gradient-parallel component and the orthogonal nonlinear cost as

$$C = -a_{\parallel}, \qquad D = b_{\perp}. \tag{10}$$

Here, $a_{\parallel}$ denotes the fitted linear coefficient for $\Delta \boldsymbol{h}_{\parallel}$, and $b_{\perp}$ denotes the fitted quadratic coefficient for $\Delta \boldsymbol{h}_{\perp}$. The gradient defines the decomposition axis, but the observed quantization displacement determines its signed projection onto that axis; C is therefore not constrained to be positive or large. Figure 1(c) shows that improved inputs are not characterized by smaller nonlinear cost. Under both Llama/NF4 and Qwen/GPTQ, improved inputs exhibit substantially stronger corrective movement C, while D is comparable or slightly larger than for harmed inputs. Thus, improvement does not require a smaller absolute nonlinear cost. Improved inputs instead receive stronger loss-reducing movement, which can outweigh a comparable or larger orthogonal nonlinear cost. Analysis of the full nonlinear residual reveals a similar pattern, with detailed results reported in Appendix A.

6 Do Loss-Reducing Changes Affect Predictions?

The previous section shows that improved inputs contain stronger loss-reducing activation changes. We next ask whether these changes are merely descriptive signatures of improvement or whether they help sustain the resulting predictions. We test this by intervening on the measured quantization-induced displacement and continuing the remaining network from the modified activation.

6.1 Targeted Activation Interventions

Let $\Delta \boldsymbol{z} = \boldsymbol{z}^{Q} - \boldsymbol{z}^{FP}$ be the activation displacement, and let $\boldsymbol{m}$ be a binary mask over channels. We define

$$\boldsymbol{z}_{\mathrm{restore}} = \boldsymbol{z}^{Q} - \boldsymbol{m} \odot \Delta \boldsymbol{z}, \qquad \boldsymbol{z}_{\mathrm{amplify}} = \boldsymbol{z}^{Q} + \boldsymbol{m} \odot \Delta \boldsymbol{z}. \tag{11}$$

Restoration removes the selected quantization-induced changes, whereas amplification strengthens them. For each input, we rank channels by the signed first-order contribution A_j from equation 4. We select the 128 strongest loss-reducing channels ($A_j < 0$) or loss-increasing channels ($A_j > 0$), with random same-sign channel sets as controls. We test three interventions: restoring loss-reducing changes, amplifying loss-reducing changes, and restoring loss-increasing changes.

Table 1: MMLU answer-level interventions on selected W2C and C2W questions. Top-128 selects channels by signed first-order contribution for each input; Random-128 samples the same number of channels with the same contribution sign. Reversal and repair rates are computed within the corresponding selected transition cohort, rather than over the full MMLU benchmark.

Model	Cohort	Intervention / outcome	Top-128 (%)	Random-128 (%)
Llama	W2C	Restore loss-reducing / reversal	58.18	3.64
Llama	C2W	Amplify loss-reducing / repair	33.33	0.00
Llama	C2W	Restore loss-increasing / repair	38.60	0.00
Qwen	W2C	Restore loss-reducing / reversal	74.60	4.76
Qwen	C2W	Amplify loss-reducing / repair	60.66	3.28
Qwen	C2W	Restore loss-increasing / repair	65.57	1.64

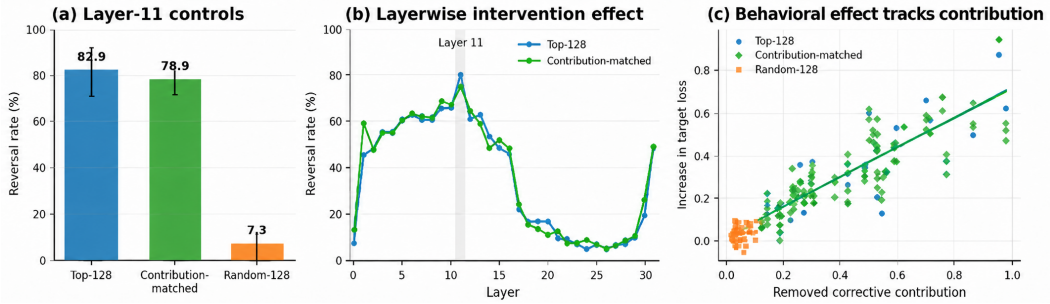


Figure 2: Intervention effect tracks functional contribution. **(a)** Contribution matching recovers most of the Top-128 reversal effect on mmlu, unlike Random-128. **(b)** The effect varies across depth, with matched controls tracking the top-ranked intervention. **(c)** Larger removed corrective contribution is associated with larger target-NLL increases.

6.2 Removing Favorable Changes Reverses Improvements

The interventions produce the predicted behavioral effects. On naturally improved FineWeb inputs, restoring the strongest loss-reducing channels turns approximately 85.7% of GPTQ cases and 87.3% of AWQ cases back into harmed cases. The same pattern reaches discrete MMLU predictions as shown in Table 1. Restoring loss-reducing changes reverses wrong-to-correct predictions, while amplifying loss-reducing changes or restoring loss-increasing changes repairs correct-to-wrong errors. For example, on Qwen, these three interventions achieve reversal or repair rates of 74.60%, 60.66%, and 65.57%, respectively, compared with 4.76%, 3.28%, and 1.64% for the corresponding random controls. Thus, the signed activation changes identified in Section 5.2 do more than correlate with the observed loss difference: directly modifying them changes the resulting prediction.

6.3 Intervention Strength Tracks Loss Contribution

The strong effect of the top-ranked interventions raises a further question: are these channels uniquely important, or are they effective because they concentrate a large signed first-order contribution? We compare the top-128 set with same-sign random channels and alternative same-sign sets matched to its total first-order contribution by including more channels when needed.

As shown in Figure 2, matching first-order contribution with a different channel set recovers most of the ranked intervention effect, while same-sign random sets remain much weaker. Across different channel selections, larger removed corrective contribution is also associated with a larger downstream loss effect. Thus, the ranked channels are effective primarily because they concentrate functional contribution, rather than because their identities are uniquely important. The layerwise results further show that this functional effect is distributed nonuniformly across network depth.

7 When Can Useful Changes Be Shared Across Inputs?

The intervention results in Section 6 show that loss-reducing quantization-induced changes can affect the prediction on the input where they are observed. We now ask whether these useful changes are entirely input-specific, or whether they contain structure that can be reused across inputs.

Table 2: Functional concentration of rank-eight shared structure at the Llama layer-11 gate. Shared effect is the fraction of the full cross-entropy improvement retained by the shared component under a common intervention scale. Residual effect is the standalone cross-entropy improvement of the orthogonal residual relative to the full change. Shared and residual effects are evaluated separately and need not sum to 100%.

Task	Shared energy (%)	Shared effect (%)	Residual effect (%)
HANS	73.70	96.06	7.21
MMLU	10.21	24.77	82.01

7.1 Isolating Local Quantization Effects

The complete-execution displacement $\Delta \boldsymbol{h}_l = \boldsymbol{h}_l^Q - \boldsymbol{h}_l^{FP}$ includes quantization effects accumulated through preceding layers. To compare useful changes across inputs more directly, we instead isolate the direct effect of a single quantized module while holding its local input fixed. We use the pre-SiLU MLP gate at zero-indexed layer 11 of Llama-3-8B and consider only the final input position. For input i, let $\boldsymbol{x}_i^Q$ denote the input to this gate in the quantized model, and let $\boldsymbol{W}^Q$ and $\boldsymbol{W}^{FP}$ denote the quantized and high-precision gate weights. The unmodified quantized gate output is

$$\boldsymbol{u}_i^Q = \boldsymbol{W}^Q \boldsymbol{x}_i^Q. \tag{12}$$

Holding $\boldsymbol{x}_i^Q$ fixed, the direct local change induced by weight quantization is

$$\boldsymbol{e}_i = \left(\boldsymbol{W}^Q - \boldsymbol{W}^{FP}\right) \boldsymbol{x}_i^Q. \tag{13}$$

Let $\ell_i(\boldsymbol{u})$ denote the target cross-entropy loss obtained by replacing the gate output at this position with $\boldsymbol{u}$ and continuing the downstream computation through the quantized model. We define

$$\boldsymbol{g}_i = \nabla_{\boldsymbol{u}} \ell_i(\boldsymbol{u})\big|_{\boldsymbol{u}=\boldsymbol{u}_i^Q}. \tag{14}$$

We then separate coordinates of the local quantization change according to their signed first-order contribution:

$$\boldsymbol{c}_i = \boldsymbol{e}_i \odot \mathbf{1}\left[\boldsymbol{g}_i \odot \boldsymbol{e}_i < 0\right], \qquad \boldsymbol{r}_i = -\boldsymbol{e}_i \odot \mathbf{1}\left[\boldsymbol{g}_i \odot \boldsymbol{e}_i > 0\right]. \tag{15}$$

Here, $\odot$ denotes elementwise multiplication and $\mathbf{1}[\cdot]$ the elementwise indicator. The component $\boldsymbol{c}_i$ retains coordinates whose quantization-induced change already reduces the target loss, whereas $\boldsymbol{r}_i$ reverses coordinates whose change increases it. Their sum,

$$\boldsymbol{d}_i = \boldsymbol{c}_i + \boldsymbol{r}_i, \tag{16}$$

defines an input-specific useful local change. Because constructing $\boldsymbol{d}_i$ uses the target loss, it serves as a diagnostic of the useful change available for that input rather than as an inference-time procedure.

7.2 A Shared Low-Dimensional Structure

We next ask whether these input-specific useful changes contain structure shared across inputs. Using N training-side examples only, we form the uncentered matrix

$$\boldsymbol{S} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_N, \boldsymbol{r}_1, \ldots, \boldsymbol{r}_N] \tag{17}$$

without per-example normalization. We keep $\boldsymbol{c}_i$ and $\boldsymbol{r}_i$ as separate columns so that the shared basis can capture reusable structure in the loss-reducing and harm-removing components separately, rather than only after they have been coupled through the input-specific sum $\boldsymbol{d}_i$.

We take the leading r left singular vectors of $\boldsymbol{S}$ as an orthonormal basis $\boldsymbol{U}_r$. We use $r = 8$ in the main experiments.

For an evaluation input, we define

$$\boldsymbol{d}_i^{\text{shared}} = \boldsymbol{U}_r \boldsymbol{U}_r^\top \boldsymbol{d}_i, \qquad \boldsymbol{d}_i^{\text{res}} = \boldsymbol{d}_i - \boldsymbol{d}_i^{\text{shared}}. \tag{18}$$

The shared component is the part of the useful change represented by a subspace learned from other inputs, while the residual contains the remaining directions outside this subspace.

7.3 The Degree of Sharing Differs Across Tasks

We compare HANS and MMLU using the same Llama-3-8B model, layer-11 gate boundary, subspace construction procedure, and intervention protocol. A separate rank-eight basis is fitted for each task using only its training-side changes.

For each evaluation input, we determine the intervention scale from the full useful change:

$$\alpha_i = \min\left(1,\ \frac{0.05\|\boldsymbol{u}_i^Q\|_2}{\|\boldsymbol{d}_i\|_2 + \epsilon}\right), \tag{19}$$

where the second term imposes a 5% relative-norm cap and ϵ is a small constant for numerical stability. Starting from the same unmodified quantized state, we evaluate

$$\boldsymbol{u}_i^Q + \alpha_i \boldsymbol{d}_i, \qquad \boldsymbol{u}_i^Q + \alpha_i \boldsymbol{d}_i^{\text{shared}}, \qquad \boldsymbol{u}_i^Q + \alpha_i \boldsymbol{d}_i^{\text{res}}. \tag{20}$$

Thus, the shared and residual components use the scale determined by the full change rather than being independently rescaled.

We quantify geometric concentration by the mean fraction of useful-change energy captured by the shared subspace:

$$E_{\text{shared}} = \frac{1}{M}\sum_{i=1}^{M} \frac{\|\boldsymbol{d}_i^{\text{shared}}\|_2^2}{\|\boldsymbol{d}_i\|_2^2}, \tag{21}$$

where M is the number of evaluation examples. To measure how much of the functional effect is retained, for $z \in \{\text{shared}, \text{res}\}$ we define

$$R_z = \frac{\sum_i \left[\ell_i(\boldsymbol{u}_i^Q) - \ell_i\left(\boldsymbol{u}_i^Q + \alpha_i \boldsymbol{d}_i^z\right)\right]}{\sum_i \left[\ell_i(\boldsymbol{u}_i^Q) - \ell_i\left(\boldsymbol{u}_i^Q + \alpha_i \boldsymbol{d}_i\right)\right]}. \tag{22}$$

Thus, E_{shared} measures geometric concentration, whereas R_z measures the retained functional effect relative to the full useful change.

As shown in Table 2, the two tasks exhibit sharply different structure. On HANS, the rank-eight shared component contains 73.70% of the useful-change energy and retains 96.06% of the full cross-entropy improvement, whereas the residual alone retains only 7.21%. On MMLU, the shared component contains only 10.21% of the useful-change energy and retains 24.77% of the full improvement, while the residual alone retains 82.01%. Thus, useful input-specific changes can be individually effective without being equally reusable across inputs: most of the useful effect is concentrated in a compact shared component on HANS, but remains largely outside that component on MMLU. This contrast persists at higher ranks (Appendix C).

This contrast motivates a practical question: when useful changes contain substantial reusable structure, can a lightweight model learn beneficial input-dependent adjustments directly from the quantized representation? We test this in Section 8 using a small input-conditioned modifier at the same layer-11 gate, with the quantized backbone frozen and no access to the target answer or target-loss gradient at inference time.

8 Relating Shared Structure to Mechanism-Guided Adaptation

Section 7 shows that useful local changes are much more concentrated in a rank-eight shared component on HANS than on MMLU. We now test whether this contrast is reflected in the additional benefit of mechanism-derived supervision over ordinary supervised adaptation.

8.1 Matched Input-Conditioned Modifiers

We insert a small FP32 modifier at the layer-11 pre-SiLU gate of a frozen GPTQ-quantized Llama-3-8B, modifying only the final input position. For input i, let $\boldsymbol{x}_i^Q$ and $\boldsymbol{u}_i^Q$ denote the gate input and

its unmodified quantized output, respectively. A network f_θ with trainable parameters θ produces the bounded adjustment

$$\boldsymbol{\delta}_{\theta,i} = 0.05\|\boldsymbol{u}_i^Q\|_2 \frac{f_\theta(\boldsymbol{x}_i^Q)}{\sqrt{1 + \|f_\theta(\boldsymbol{x}_i^Q)\|_2^2}}. \tag{23}$$

The FP32 sum $\boldsymbol{u}_i^Q + \boldsymbol{\delta}_{\theta,i}$ is cast back to BF16 before the remaining backbone computation.

We compare BASE and C/D modifiers with the same architecture, data protocol, and 800-step optimization budget. Each contains 604,192 trainable parameters and uses neither the target answer nor a target-loss gradient at inference time. BASE minimizes an objective $\mathcal{J}_0$ comprising target-answer cross-entropy, an additional penalty for loss increases relative to the unmodified quantized model, and a relative adjustment-norm penalty (Appendix D.1). The mechanism-guided variant adds

$$\mathcal{J}_{C/D} = \mathcal{J}_0 + \mathbb{E}_i \left[-0.1 \left(C_{\theta,i} - C_{Q,i}\right) + 0.1 \left(D_{\theta,i} - D_{Q,i}\right)\right]. \tag{24}$$

Here, $C_{\theta,i}$ and $D_{\theta,i}$ are computed from the modified final-block representation relative to the matched FP reference, using the parallel–orthogonal quadratic construction in Section 5.3 with target-answer cross-entropy as the task loss. $C_{Q,i}$ and $D_{Q,i}$ are the corresponding fixed quantities for the unmodified quantized execution. The added terms reward stronger corrective movement and penalize greater orthogonal nonlinear cost.

Neither modifier directly uses the rank-eight basis from Section 7; BASE and C/D differ only in their training objective.

8.2 Generic Adaptation and Mechanism-Guided Gain

Both tasks benefit from ordinary supervised activation modification, but the additional effect of mechanism guidance differs substantially. On HANS, accuracy increases from 50.00% for the unmodified four-bit backbone (Q4) to 56.40% with BASE and 57.68% with C/D, giving an additional 1.28-point gain with a paired bootstrap 95% confidence interval of $[+0.45, +2.13]$ points. On MMLU, BASE improves accuracy from 55.49% to 58.56%, while C/D reaches 58.65%, a 0.10-point difference with a paired bootstrap 95% confidence interval of $[-0.23, +0.42]$ points. The full FP/Q4/BASE/C/D comparison is reported in Appendix D.2.

The Q4-to-BASE gain measures ordinary supervised adaptation, whereas the C/D-minus-BASE difference measures the added effect of mechanism guidance. This increment is clear on HANS and small on MMLU, consistent with the stronger shared structure identified on HANS in Section 7.

The two modifiers produce strongly aligned adjustments: their median cosine similarity is 0.948 on HANS and 0.978 on MMLU, with median C/D-to-BASE norm ratios close to one on both tasks. This suggests that mechanism guidance primarily refines the adjustment already learned from task supervision. The full comparison is reported in Appendix D.3.

9 Conclusion

We study why low-bit quantization can improve some inputs even when aggregate performance declines. Across multiple quantizers and model families, these improvements are distinguished not simply by smaller activation changes, but by stronger loss-reducing movement that can outweigh comparable or larger nonlinear costs. Targeted interventions further show that these favorable changes participate in the resulting predictions, while contribution-matched controls show that intervention strength closely tracks functional contribution.

We further find that the reusable structure of these changes varies substantially across tasks. On HANS, a compact shared representation preserves most of their functional effect, whereas on MMLU the same low-rank representation retains much less. Consistent with this contrast, mechanism-derived supervision provides additional gain beyond ordinary supervised modification on HANS but little additional gain on MMLU.

Together, these results connect the functional signature, behavioral role, and cross-input reuse of beneficial quantization-induced changes. More broadly, they suggest that quantization should be analyzed not only by how closely it preserves high-precision computation, but also by what its induced changes do to predictions and when their useful structure can be reused.

REFERENCES

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L. Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. Quarot: Outlier-free 4-bit inference in rotated llms. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 100213–100240. Curran Associates, Inc., 2024. doi: 10.52202/079017-3180. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/b5b939436789f76f08b9d0da5e81af7c-Paper-Conference.pdf.

Ting-Yun Chang, Muru Zhang, Jesse Thomason, and Robin Jia. Why do some inputs break low-bit LLM quantization? In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3410–3429, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.168. URL https://aclanthology.org/2025.emnlp-main.168/.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3.int8(): 8-bit matrix multiplication for transformers at scale. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30318–30332. Curran Associates, Inc., 2022. doi: 10.52202/068431-2198. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c3ba4962c05c49636d4c6206a97e9c8a-Paper-Conference.pdf.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 10088–10115. Curran Associates, Inc., 2023a. doi: 10.52202/075280-0441. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1feb87871436031bdc0f2beaa62a049b-Paper-Conference.pdf.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 10088–10115. Curran Associates, Inc., 2023b. doi: 10.52202/075280-0441. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1feb87871436031bdc0f2beaa62a049b-Paper-Conference.pdf.

Tim Dettmers, Ruslan Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, and Dan Alistarh. Spqr: A sparse-quantized representation for near-lossless llm weight compression. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 5733–5761, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/1787533e171dcc8549cc2eb5a4840eec-Paper-Conference.pdf.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. OPTQ: Accurate quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tcbBPnfwxS.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco

Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L.

Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Han Guo, Philip Greengard, Eric Xing, and Yoon Kim. Lq-lora: Low-rank plus quantized matrix decomposition for efficient language model finetuning. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 50167–50184, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/d97a16cb83d74195b76e0bf1e85bf072-Paper-Conference.pdf.

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://openreview.net/forum?id=d7KBjmI3GmQ.

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.

Geonho Lee, Janghwan Lee, Sukjin Hong, Minsoo Kim, Euijai Ahn, Du-Seong Chang, and Jungwook Choi. Rilq: rank-insensitive lora-based quantization error compensation for boosting 2-bit large language model accuracy. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i17.33990. URL https://doi.org/10.1609/aaai.v39i17.33990.

Yixiao Li, Yifan Yu, Chen Liang, Nikos Karampatziakis, Pengcheng He, Weizhu Chen, and Tuo Zhao. Loftq: Lora-fine-tuning-aware quantization for large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 13409–13424, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/39ec972afab01e0d8ddc6834a9d12ac1-Paper-Conference.pdf.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. In P. Gibbons, G. Pekhimenko,

and C. De Sa (eds.), *Proceedings of Machine Learning and Systems*, volume 6, pp. 87–100, 2024. URL https://proceedings.mlsys.org/paper_files/paper/2024/file/42a452cbafa9dd64e9ba4aa95cc1ef21-Paper-Conference.pdf.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. Spinquant: Llm quantization with learned rotations. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 92009–92032, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/e5b1c0d4866f72393c522c8a00eed4eb-Paper-Conference.pdf.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334/.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 30811–30849. Curran Associates, Inc., 2024. doi: 10.52202/079017-0970. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/370df50ccfdf8bde18f8f9c2d9151bda-Paper-Datasets_and_Benchmarks_Track.pdf.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods, 2024. URL https://arxiv.org/abs/2309.16042.

AI USE STATEMENT

We used AI tools to polish the language and improve the clarity of the manuscript. The authors reviewed all AI-assisted revisions for accuracy and consistency with the intended meaning and take full responsibility for the final content.

REPRODUCIBILITY STATEMENT

The main text describes the models, quantization methods, datasets, sampling procedures, and intervention protocols, and explicitly defines the quantities used in our analysis. Appendices A and B provide the full nonlinear residual analysis and detailed layerwise comparisons. Appendix C reports sharing results across ranks using fixed evaluation inputs and a common intervention scale. Appendix D provides the BASE training objective and additional adaptation comparisons. Shared bases are estimated from training examples and fixed for evaluation. BASE and C/D use the same frozen backbone, architecture, parameter budget, data protocol, and optimization budget, differing only in the added mechanism-derived supervision.

Table 3: Raw full nonlinear residual at the final-block boundary. Differences are improved minus harmed, computed before rounding.

Setting	Improved mean	Harmed mean	Difference
Llama/GPTQ	0.0728	0.0755	-0.0027
Llama/AWQ	0.0631	0.0662	-0.0030
Llama/NF4	0.0977	0.0748	0.0229
Qwen/GPTQ	0.0377	0.0355	0.0021

Table 4: Full nonlinear residual normalized by squared displacement magnitude, $R/(\|\Delta \boldsymbol{h}\|_2^2 + 10^{-12})$. Differences are improved minus harmed, computed before rounding.

Setting	Improved mean	Harmed mean	Difference
Llama/GPTQ	5.31×10^{-7}	6.99×10^{-7}	-1.68×10^{-7}
Llama/AWQ	5.69×10^{-7}	7.48×10^{-7}	-1.79×10^{-7}
Llama/NF4	6.64×10^{-7}	7.48×10^{-7}	-8.44×10^{-8}
Qwen/GPTQ	5.43×10^{-9}	5.55×10^{-9}	-1.23×10^{-10}

A Analysis of the Full Nonlinear Residual

We complement the orthogonal quadratic coefficient D with the full residual beyond the first-order approximation. At the final Transformer block output, before the final RMSNorm, we compute

$$R = L(\boldsymbol{h}^{\mathrm{Q}}) - L(\boldsymbol{h}^{\mathrm{FP}}) - \boldsymbol{g}^{\top} \Delta \boldsymbol{h}, \qquad \Delta \boldsymbol{h} = \boldsymbol{h}^{\mathrm{Q}} - \boldsymbol{h}^{\mathrm{FP}}. \tag{25}$$

Here, L is the mean next-token NLL under the shared downstream readout, and $\boldsymbol{g}$ is its gradient at the FP reference. This residual includes interactions between displacement components and higher-order effects beyond the linear approximation. We retain the original FineWeb cohorts, with sampling weights for Llama and absolute-NLL-change matching for Qwen.

Table 3 and Figure 3(a) show that the mean raw residual is not consistently smaller for improved inputs. Under Llama/GPTQ and Llama/AWQ, the two groups have similar mean raw residuals, while under Llama/NF4 and Qwen/GPTQ, improved inputs have larger mean residuals. These results extend the main-text finding to the full displacement: stronger first-order loss reduction can outweigh comparable or larger nonlinear loss increases.

Table 4 and Figure 3(b) additionally report

$$\frac{R}{\|\Delta \boldsymbol{h}\|_2^2 + 10^{-12}}. \tag{26}$$

The mean normalized residual is lower for improved inputs in all four settings. Thus, smaller nonlinear effects relative to squared displacement magnitude can coexist with comparable or larger absolute nonlinear costs.

B Layerwise Displacement Magnitude and Corrective Efficiency

We extend the pre-SiLU MLP-gate analysis to all 32 layers of Llama-3-8B under GPTQ and AWQ and all 28 layers of Qwen2.5-7B-Instruct under GPTQ. At each layer, we measure the displacement between complete FP and quantized executions and compute ρ_{red} following equation 4, using FP-reference gradients of mean next-token NLL. All layers use the original FineWeb cohorts, sampling weights or matching, and fixed group assignments.[1]

Figure 4 shows that the displacement distributions of improved and harmed inputs overlap throughout the network. Their interquartile ranges overlap at every analyzed layer in all three settings, extending the displacement overlap observed in the main text beyond the representative layer.

Figure 5 shows higher mean ρ_{red} for improved inputs at 26 of 32 Llama layers under both GPTQ and AWQ and at all 28 Qwen layers under GPTQ. For Llama, this ordering holds at layers 0–25

[1]Some NLL-change signs differ under numerical replay; these inputs retain their original improved/harmed labels.

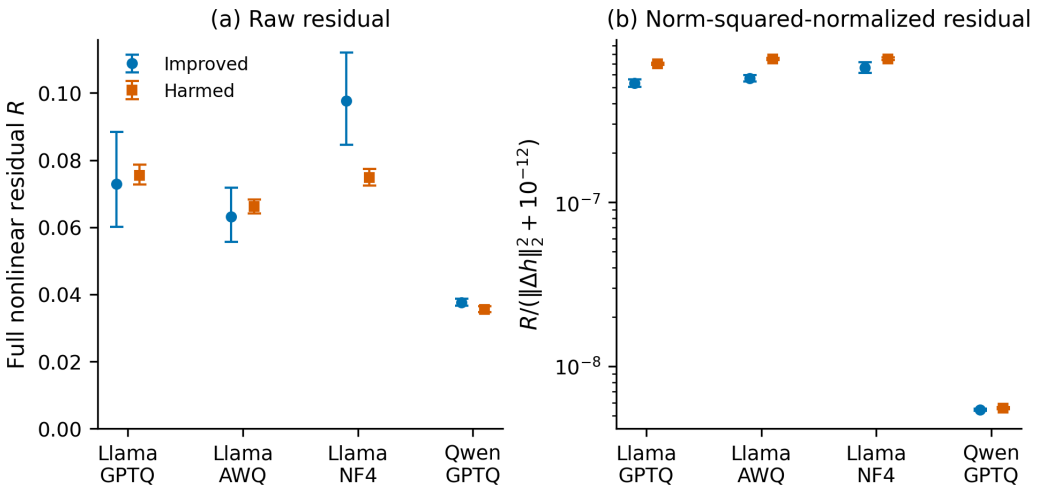


Figure 3: Full nonlinear residuals at the final-block boundary. (a) Raw residual R. (b) Residual normalized by squared displacement magnitude. Points show weighted group means, and error bars indicate 95% bootstrap confidence intervals. The normalized-residual axis is logarithmic.

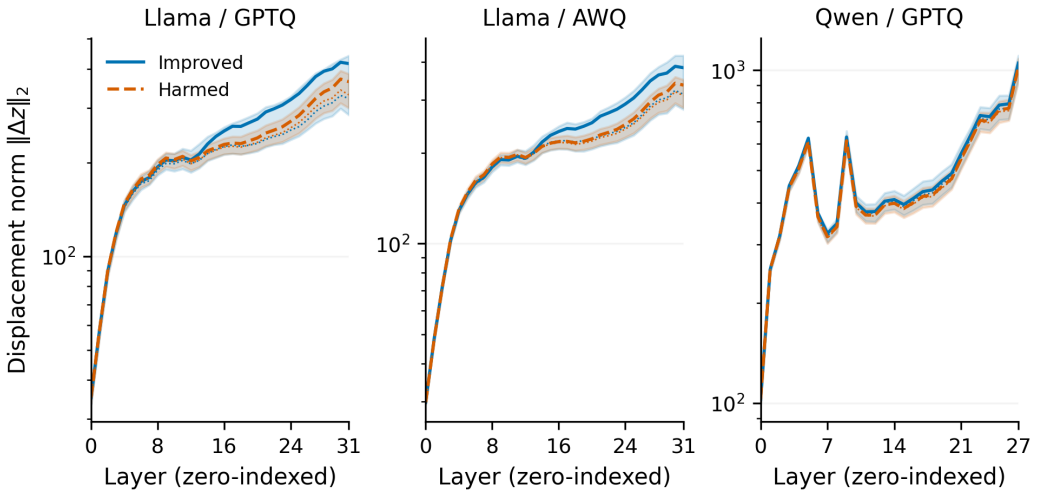


Figure 4: Gate displacement magnitudes across layers for Llama-3-8B under GPTQ and AWQ and Qwen2.5-7B-Instruct under GPTQ. Solid and dashed curves denote weighted means for improved and harmed inputs, respectively; dotted curves denote weighted medians. Shaded bands span the weighted 25th–75th percentiles. Vertical axes are logarithmic, and layer indices are zero-based.

and reverses slightly at layers 26–31. These results show that stronger loss-reducing contributions per unit displacement occur across most Llama layers and throughout Qwen, supporting the broader relevance of the main-text observation.

C Shared Structure Across Ranks

We vary the shared-subspace rank from 1 to 128 using the same Llama-3-8B GPTQ model and layer-11 gate as in Section 7.3. For each task, all ranks use the same training-side changes and evaluation inputs. The intervention scale is determined by the full useful change and held fixed across ranks for both shared and residual components.

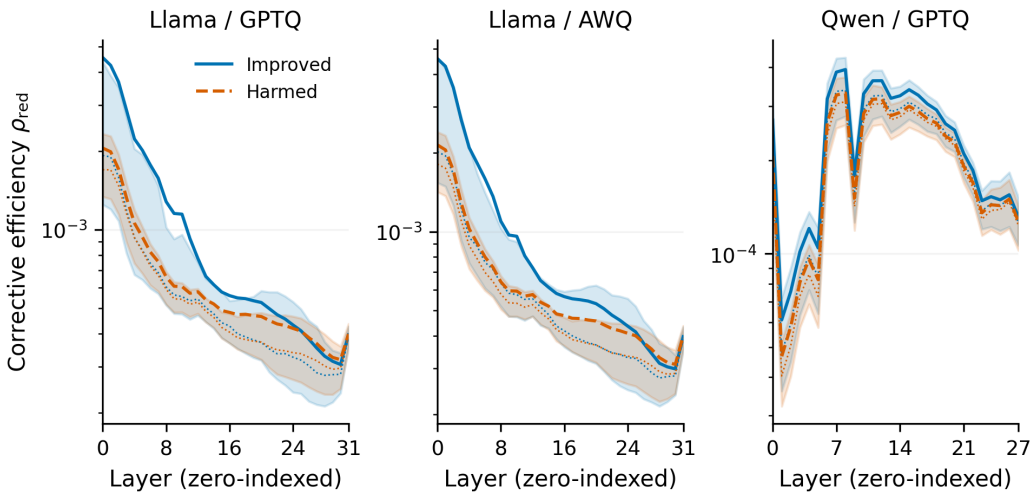


Figure 5: Loss-reducing contribution per unit displacement, ρ_{red}, across layers for the same model–quantizer settings as Figure 4. Solid and dashed curves denote weighted means for improved and harmed inputs, respectively; dotted curves denote weighted medians. Shaded bands span the weighted 25th–75th percentiles. Vertical axes are logarithmic, and layer indices are zero-based.

Table 5: Shared energy and functional effects across ranks at the Llama-3-8B layer-11 gate. All values are percentages. Energy denotes the mean per-view fraction of useful-change energy captured by the shared subspace. Shared and Residual denote ratios of summed cross-entropy reductions relative to the full useful change. Both components use the full-change intervention scale and are evaluated separately; their effects need not sum to 100%.

	HANS			MMLU		
Rank r	Energy	Shared	Residual	Energy	Shared	Residual
1	0.03	0.01	99.95	0.02	0.05	99.95
2	52.09	79.55	33.49	5.10	12.17	91.50
4	71.81	95.33	8.61	7.04	17.52	87.53
8	73.70	96.06	7.21	10.21	24.77	82.01
16	76.70	97.02	5.55	14.04	32.02	76.35
32	79.16	97.66	4.41	16.91	38.62	70.76
64	81.04	98.16	3.43	19.74	43.71	66.31
128	82.79	98.46	3.00	22.77	47.76	62.65

Table 5 reports shared energy, shared effect, and residual effect. On HANS, four directions already retain 95.33% of the full cross-entropy improvement, increasing to 98.46% at rank 128. On MMLU, shared retention increases more gradually, from 17.52% at rank four to 47.76% at rank 128. At rank 128, the residual alone retains 3.00% of the full improvement on HANS and 62.65% on MMLU. The contrast between tasks therefore persists beyond the rank-eight comparison in the main text: HANS concentrates most of the useful effect in a compact shared subspace, whereas MMLU retains less than half even with substantially more directions.

D Additional Details on Mechanism-Guided Adaptation

D.1 BASE Training Objective

The BASE modifier uses the bounded adjustment defined in equation 23 and minimizes

$$\mathcal{J}_0 = \mathbb{E}_i \left[\mathrm{CE}_{\theta,i} + 0.1 \left[\mathrm{CE}_{\theta,i} - \mathrm{CE}_{Q,i} \right]_+ + 0.01 \frac{\|\boldsymbol{\delta}_{\theta,i}\|_2^2}{\|\boldsymbol{u}_i^Q\|_2^2 + 10^{-12}} \right], \tag{27}$$

where $\mathrm{CE}_{\theta,i}$ and $\mathrm{CE}_{Q,i}$ are the target-answer cross-entropies of the modified and unmodified quantized executions, respectively, and $[z]_+ = \max(z, 0)$. The hinge term adds a penalty when the

Table 6: Accuracy (%) of the matched FP reference, Q4 backbone, and BASE and C/D modifiers. HANS uses the test split and MMLU uses the development split. Δ_{FP} and Δ_{BASE} denote C/D minus FP and BASE, respectively, in percentage points. Differences are computed before rounding.

Task	FP	Q4	BASE	C/D	Δ_{FP}	Δ_{BASE}
HANS	50.27	50.00	56.40	57.68	+7.41	+1.28
MMLU	56.24	55.49	58.56	58.65	+2.42	+0.10

modification increases the target loss relative to the unmodified quantized model. The final term penalizes the squared adjustment norm relative to the unmodified gate output. The C/D variant retains all three terms and adds the mechanism objective in equation 24.

D.2 Accuracy Relative to the High-Precision Reference

Table 6 compares the matched high-precision reference, the unmodified quantized backbone, and the two activation modifiers on the same evaluation inputs for each task. C/D exceeds the FP reference by 7.41 percentage points on HANS and 2.42 on MMLU, while its additional gain over BASE is larger on HANS.

D.3 Mechanism Guidance Refines the Learned Adjustment

We compare the activation adjustments produced by the two training objectives. For each input i, let $\boldsymbol{\delta}_i^{\mathrm{BASE}}$ and $\boldsymbol{\delta}_i^{C/D}$ denote the bounded FP32 adjustments before BF16 writeback. Their cosine similarity is

$$s_i = \frac{\left\langle \boldsymbol{\delta}_i^{\mathrm{BASE}}, \boldsymbol{\delta}_i^{C/D} \right\rangle}{\|\boldsymbol{\delta}_i^{\mathrm{BASE}}\|_2 \|\boldsymbol{\delta}_i^{C/D}\|_2}. \tag{28}$$

The median cosine similarity is 0.948 on HANS and 0.978 on MMLU. The median ratios

$$\frac{\|\boldsymbol{\delta}_i^{C/D}\|_2}{\|\boldsymbol{\delta}_i^{\mathrm{BASE}}\|_2}$$

are approximately one on both tasks. The two objectives therefore learn strongly aligned adjustments with similar magnitudes.

These directional differences can nevertheless affect behavior. On HANS, C/D guidance improves over BASE despite the high similarity, whereas on MMLU the adjustments are even more strongly aligned and their accuracy difference is small. Together, these results suggest that mechanism guidance primarily refines an adjustment already learned from task supervision, with a larger incremental benefit on HANS, where useful changes are more concentrated in a compact shared component.