ACTIVATION STEERING UNDER LOW-BIT QUANTIZATION: A GEOMETRIC PERSPECTIVE ON DEGRADATION AND RECOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Activation steering enables inference-time control of language-model behavior, but whether an existing steering interface survives low-bit quantization remains unclear. We study steering compatibility with the intervention layer, direction, and strength held fixed across full-precision and quantized models. We show that ordinary output fidelity alone does not guarantee preservation of responses to the same intervention. We explain the gap through the geometry of the quantization error. At the layer level, we relate directional response error to calibration geometry, showing that weakly covered directions admit larger worst-case errors under a fixed reconstruction budget. At the output level, we characterize intervention-dependent error amplification through the relative Fisher geometry of unsteered and intervened distributions. Under suitable coverage and bounded-sensitivity conditions, we derive a local upper bound on steered output KL in terms of unsteered output KL. Motivated by this analysis, we investigate post-quantization output distillation using only ordinary, unsteered teacher outputs.

1 INTRODUCTION

Activation steering (Rimsky et al., 2024) enables inference-time control of model behavior without retraining, providing a lightweight mechanism for adapting a single model to different behaviors, safety policies, or generation styles. By adding a learned direction to intermediate activations, steering methods can modify behaviors such as refusal, sentiment, or stylistic attributes while leaving the underlying model parameters unchanged.

Quantization is a key technique for deploying large language models under limited memory and compute budgets (Frantar et al., 2022; Lin et al., 2024; Xiao et al., 2023; Shao et al., 2024). Low-bit weight quantization reduces parameter storage and can lower inference costs while seeking to preserve the behavior of the original full-precision model. Combining these techniques raises a deployment question:

> *Does a steering interface constructed on a full-precision model remain compatible with that model after quantization?*

We study this question with the intervention layer, steering vector, and intervention strength held fixed across the full-precision and quantized models. Its overview is visualized in Figure 1. This setting distinguishes *preserving an existing steering interface* from constructing a new one for a quantized model. Re-estimating a direction or retuning its strength after quantization addresses the latter problem, but does not establish that the original interface has been preserved. Our focus is therefore not merely whether a quantized model can still be steered, but whether it responds similarly to the same intervention. Recent work has empirically observed degradation of steering or related behavioral properties under quantization, but the mechanisms governing whether a fixed steering interface is preserved remain poorly understood.

Ordinary fidelity alone does not resolve this question. For example, GPTQ minimizes layer-wise reconstruction error using calibration activations collected without steering (Frantar et al., 2022). Such an objective constrains quantization errors according to the geometry of ordinary activations, rather than equally in every possible intervention direction. A steering intervention can expose 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.

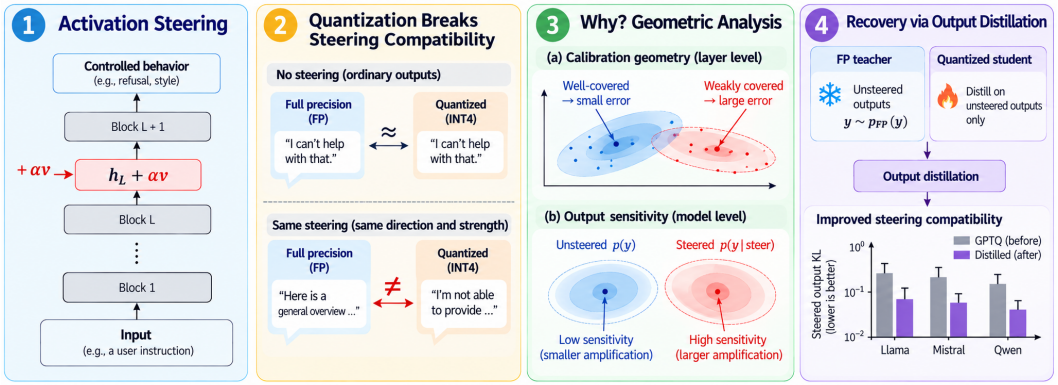

Figure 1: Overview of this research. (1) Activation steering controls model behavior by adding a learned direction to intermediate activations. (2) Low-bit quantization can preserve ordinary output fidelity while degrading responses to the same steering interface. (3) We explain the gap through the geometry of quantization error: at the layer level via calibration geometry, and at the output level via intervention-dependent error amplification. (4) We investigate output distillation using only unsteered teacher outputs, which strengthens steering compatibility.

along directions that are weakly represented by this calibration geometry. Moreover, downstream processing can amplify or suppress these perturbations, so layer-wise reconstruction quality does not by itself determine agreement between the final steered outputs.

We investigate this gap at two complementary levels. At the layer level, we derive a bound on directional response error that makes the role of calibration coverage explicit. At the model-output level, we use local quadratic expansions of unsteered and intervened output KL divergences. Their Fisher matrices describe the sensitivity of the respective output distributions to the same parameter perturbation, making intervention-dependent amplification explicit. We derive sufficient conditions under which ordinary Fisher coverage of the relevant error directions, together with bounded intervened sensitivity, yields a local upper bound on steered KL in terms of unsteered KL.

This conditional relationship motivates a practical question: can ordinary output matching recover steering compatibility without access to the target interventions? We examine a simple post-quantization distillation procedure that matches a fixed full-precision teacher using only ordinary, unsteered outputs. No steering vectors, intervened teacher outputs, or steering-specific behavioral supervision are used during training. Rather than introducing a new distillation objective, we use this procedure to investigate how far ordinary output matching can restore responses to interventions absent from training.

Our experiments evaluate ordinary fidelity and steering compatibility separately across model families and steering directions. We find that improvements in ordinary fidelity need not be accompanied by improved steering preservation, while unsteered output distillation can reduce steered output mismatch in the evaluated settings. These findings motivate treating compatibility with existing steering interfaces as a distinct property to assess when deploying quantized language models.

Our contributions are to show the following:

- For activation steering, ordinary fidelity, local steered mismatch, and generation-level response curves are evaluated separately.
- Geometric analysis explains steering compatibility failures: reconstruction error depends on calibration coverage, while Fisher geometry characterizes intervention-dependent amplification and yields a local bound on steered KL.
- The analysis motivates steering-free distillation as a recovery strategy: when ordinary calibration sufficiently covers steering-relevant directions, it can reduce steered output mismatch without steering information.

2 PRELIMINARY

2.1 ACTIVATION STEERING

Activation steering (Rimsky et al., 2024) modifies a language model's behavior at inference time without updating its parameters. The basic idea is to identify a direction in the model's hidden representation space that is associated with a target behavior, such as refusal, sentiment, or style, and then move the intermediate representation along that direction.

Formally, let $p_\theta(\cdot \mid x)$ denote the next-token distribution of a pretrained autoregressive language model with parameters θ. At a chosen layer ℓ, let $h_\ell(x) \in \mathbb{R}^d$ be the residual stream representation associated with input context x. Activation steering modifies this representation at inference time as $\mathcal{I} * v, c : h * \ell(x) \mapsto h_\ell(x) + cv$, where $v \in \mathbb{R}^d$ is a unit-norm steering vector and $c \in \mathbb{R}$ is its intervention strength. We write $p_\theta^{v,c}(\cdot \mid x)$ for the distribution produced by the resulting intervened model. The case $c = 0$ recovers ordinary, unsteered inference. This formulation covers vectors estimated from contrastive examples, vectors recalibrated on held-out data, and externally supplied representation-engineering directions.

2.2 QUANTIZATION

Let Q be a post-training quantization procedure and $\theta_Q = Q(\theta)$ the quantized parameters. We write the induced parameter perturbation as $\delta = \theta_Q - \theta$. Although Q is discrete, treating δ as a perturbation provides the local parameterization used in our analysis. At an individual linear layer with full-precision weight W and quantized weight W_Q, the corresponding error is $\delta = W_Q - W$. We use $p_{\theta_Q}^{v,c}$ to denote the quantized model under the same intervention with steering direction v and strength c. Crucially, the layer, vector, and coefficient are held fixed across the full-precision and quantized models; preservation is not credited to a student-specific vector or post-hoc rescaling of c.

3 WHY ORDINARY FIDELITY DOES NOT GUARANTEE STEERING PRESERVATION

Ordinary fidelity constrains model behavior only on the unsteered input distribution and therefore does not necessarily preserve responses to internal interventions. We first characterize how interventions can amplify quantization error (Section 3.1), and then derive the coverage conditions under which small ordinary KL divergence is sufficient to control steered KL divergence (Section 3.4).

3.1 INTERVENTION-DEPENDENT ERROR AMPLIFICATION

At the layer level, the reconstruction objective does not constrain all intervention directions equally. Let δ denote the quantization error of a linear layer and let $\Gamma = \Sigma + \gamma I \succ 0$ denote the damped second-moment matrix of its calibration inputs. For an effective intervention direction d at this layer,

$$\|\delta d\|_2^2 \le D_\Gamma(\delta), d^\mathsf{T}\Gamma^{-1}d, \qquad D_\Gamma(\delta) := \mathrm{tr}(\delta\Gamma\delta^\mathsf{T}). \tag{1}$$

Proof. Since $\Gamma \succ 0$, its symmetric square root and inverse square root are well defined. Therefore,

$$\|\delta d\|_2^2 = \|(\delta\Gamma^{1/2})(\Gamma^{-1/2}d)\|_2^2 \le \|\delta\Gamma^{1/2}\|_\mathrm{F}^2\|\Gamma^{-1/2}d\|_2^2 = \mathrm{tr}(\delta\Gamma\delta^\mathsf{T}), d^\mathsf{T}\Gamma^{-1}d = D_\Gamma(\delta), d^\mathsf{T}\Gamma^{-1}d,$$

where the inequality follows from Cauchy–Schwarz applied to each row of $\delta\Gamma^{1/2}$. □

Hence, $s_\Gamma(d) := d^\mathsf{T}\Gamma^{-1}d$ quantifies the worst-case directional vulnerability permitted by a fixed reconstruction budget: directions that are weakly covered by the calibration geometry admit larger worst-case layer-level response errors. This layer-level geometry does not by itself determine the final output error, since downstream processing can further amplify or suppress the perturbation. We therefore characterize the corresponding amplification at the model-output level through the Fisher geometry of the unsteered and intervened distributions.

Let U be the subspace of feasible local quantization errors, and let F_0 and $F_{v,c}$ denote the Fisher matrices for the unsteered and intervened Kullback–Leibler divergence objectives. For sufficiently

small $\Delta \in U$,

$$L_0(\Delta) = \frac{1}{2}\Delta^\top F_0 \Delta + O(\|\Delta\|^3), \quad K_{v,c}(\Delta) = \frac{1}{2}\Delta^\top F_{v,c}\Delta + O(\|\Delta\|^3). \tag{2}$$

Define the intervention-dependent amplification factor

$$\kappa(v, c; U) = \sup_{u \in U \setminus \{0\}} \frac{u^\top F_{v,c} u}{u^\top F_0 u}, \qquad \kappa_R(U) = \sup_{\|v\|=1,\ |c| \le R} \kappa(v, c; U). \tag{3}$$

This ratio quantifies how much a parameter-error direction that is small under the ordinary objective can be amplified after intervention.

3.2 When does ordinary fidelity fail to guarantee preservation?

Ordinary fidelity fails to provide a uniform preservation guarantee when an error direction is weakly constrained before intervention but becomes sensitive after intervention. In particular, if there exists a nonzero $u \in U$ such that

$$u^\top F_0 u = 0, \qquad u^\top F_{v,c} u > 0 \tag{4}$$

then $\kappa(v, c; U) = \infty$.

The same issue arises continuously when F_0 is positive definite but has very small curvature along an intervention-sensitive direction: the amplification factor can then become arbitrarily large. This motivates requiring the ordinary objective to have non-vanishing coverage over the relevant error subspace.

3.3 Coverage as a sufficient condition

The failure mode above is avoided when the ordinary objective has non-vanishing curvature over the relevant error subspace and the intervened sensitivity is uniformly bounded.

Proposition 3.1 (Sufficient condition for bounded intervention amplification)**.** *Suppose that the ordinary Fisher has non-vanishing curvature on U,*

$$u^\top F_0 u \ge \mu_U \|u\|_2^2, \qquad \forall u \in U, \tag{5}$$

for some $\mu_U > 0$. Further, define the maximum intervened sensitivity on U as

$$\beta_R := \sup_{\|v\|=1,\ |c| \le R,\ u \in U,\ \|u\|=1} u^\top F_{v,c} u \tag{6}$$

and assume $\beta_R < \infty$. Then,

$$\kappa_R(U) \le \frac{\beta_R}{\mu_U} < \infty. \tag{7}$$

Thus, positive ordinary Fisher coverage on U, together with bounded intervened sensitivity, prevents interventions from revealing error directions that are effectively unconstrained by the ordinary objective.

3.4 Implication for ordinary KL distillation

Theorem 3.2 (Local control of steered KL by ordinary KL)**.** *Assume that F_0 is positive definite on U. Then, for any fixed intervention (v, c) and sufficiently small $\Delta \in U$,*

$$K_{v,c}(\Delta) \le \kappa(v, c; U) L_0(\Delta) + O(\|\Delta\|^3). \tag{8}$$

If, in addition, $\kappa_R(U) < \infty$ and the third-order remainder bounds hold on a common neighborhood of $\Delta = 0$ with constants independent of (v, c) for $\|v\|_2 = 1$ and $|c| \le R$, then

$$\sup_{|v|_2=1\ |c| \le R} K_{v,c}(\Delta) \le \kappa_R(U) L_0(\Delta) + O(|\Delta|_2^3), \tag{9}$$

where the remainder bound holds uniformly for all $\Delta \in U$ in that neighborhood.

The proof is shown in the Appendix. The theorem shows that ordinary fidelity controls steering fidelity only up to the intervention-dependent factor $\kappa_R(U)$. Under the coverage and bounded-sensitivity conditions above, reducing the ordinary KL also reduces a uniform local upper bound on the steered KL. This provides a theoretical rationale for why ordinary KL distillation can improve steering preservation without explicitly training on the target interventions, while not guaranteeing that such recovery occurs for arbitrary interventions.

4 Recovering Steering Compatibility without interventions

The preceding analysis (Theorem 3.2) suggests that reducing ordinary KL can also reduce steered output error when the relevant error directions are sufficiently covered. Motivated by this observation, we apply a simple post-quantization self-distillation procedure using only unsteered outputs.

$$\mathbb{E}_{x\sim D_{\text{train}}} \mathrm{KL}\left(p_{\theta,0}(\cdot \mid x) | p_{Q(u),0}(\cdot \mid x)\right) \tag{10}$$

The full-precision model is kept fixed as the teacher, and no steering vector, intervened output, or steering-specific behavioral signal is used during training. Importantly, this is not a new distillation objective. Our question is whether an ordinary output-matching objective, despite never observing the target interventions, can recover the steering responses that were degraded by quantization.

5 Related work

Activation Steering Activation steering controls model behavior at inference time by modifying intermediate representations without updating parameters. Activation Addition adds contrastive-prompt activation differences to the residual stream (Turner et al., 2023), while Representation Engineering generalizes this idea to identifying and manipulating behavior-relevant representation directions (Zou et al., 2023). Contrastive Activation Addition improves robustness by averaging differences across contrastive examples and has been applied to behaviors such as sycophancy and hallucination (Rimsky et al., 2024). Related work has identified steering directions for in-context tasks (Liu et al., 2024), function execution (Todd et al., 2024), and refusal (Arditi et al., 2024). Together, these studies show that behavior can be causally controlled through low-dimensional activation directions. We instead ask whether a steering interface defined on a full-precision model remains functionally compatible after quantization.

Quantization Post-training quantization (PTQ) reduces the memory and inference costs of large language models while preserving their original behavior. GPTQ uses approximate second-order information from calibration activations to minimize layer-wise reconstruction error (Frantar et al., 2022). AWQ protects salient weight channels identified from activation statistics via channel-wise scaling (Lin et al., 2024), while SmoothQuant shifts quantization difficulty from activation outliers to weights through an equivalent transformation (Xiao et al., 2023). OmniQuant further optimizes quantization parameters with block-wise reconstruction objectives (Shao et al., 2024). These methods are calibrated on ordinary, unsteered activations or reconstruction objectives, which do not directly constrain errors along directions that become important only under internal interventions. Our layer-level analysis formalizes this gap by relating intervention-direction response error to its coverage under the calibration geometry.

Behavioral robustness and activation steering under quantization. Recent work shows that preserving conventional language-model utility after quantization does not guarantee preservation of other behavioral properties. Quantization has been studied in terms of safety and reliability (Kharinaev et al., 2026), with evidence that fairness or safety can degrade despite relatively preserved utility (Al Hakim et al., 2026). While recent studies also report degradation of steering-related behavior under quantization, what determines whether a fixed steering interface is preserved remains poorly understood.

6 Experiments

We consider the following research questions:

- Is ordinary output fidelity alone sufficient to determine whether responses to a fixed steering interface are preserved? (Section 6.1)
- To what extent can calibration geometry characterize direction-dependent quantization error at the layer level? (Section 6.2)
- To what extent can Fisher geometry locally characterize intervention-dependent error amplification at the model-output level? (Section 6.3)
- To what extent can distillation on ordinary, unsteered outputs recover responses to a fixed steering interface without access to intervention-specific information? (Section 6.4)

6.1 STEERING PRESERVATION

We first test whether ordinary output fidelity is sufficient to identify models that preserve steering responses. Starting from the same 3-bit GPTQ formulation, we optimize checkpoints with next-token CE, final hidden-state MSE, or ordinary output KL. For each model, we select one checkpoint per objective to minimize the relative spread of development unsteered KL, then evaluate all selected checkpoints on disjoint prompts and steering directions.

Table 1: Difference between ordinary KL and steered KL. In each row, one checkpoint per objective is selected to minimize relative spread in development unsteered KL. Slash-separated entries are CE / hidden-MSE / KL objectives. L_0^{dev} is unsteered KL on development prompts; K^{held} is mean intervened next-token KL on held-out prompts over the fixed steering-direction evaluation set.

Model	Relative L_0^{dev} spread	L_0^{dev} (CE / MSE / KL)	K^{held} (CE / MSE / KL)
Llama-3.1-8B-Instruct	0.7%	$.759/.764/.764$	$1.130/1.342/1.270$
Mistral-7B-Instruct-v0.3	12.3%	$.476/.471/.532$	$.743/.465/.868$
Qwen3-8B	6.7%	$.790/.760/.812$	$1.865/1.545/1.700$

The held-out protocol uses, throughout, the same FP-extracted Safety and Persona directions and the fixed seeded random-direction set for every selected checkpoint. Safety is the unit-normalized difference between mean middle-layer residual activations for harmful and harmless instruction prompts; Persona is analogously extracted from cheerful and terse prompts. We use each direction and the FP16-derived intervention scale unchanged for the teacher and every quantized checkpoint, and evaluate them only on disjoint neutral prompts. Thus, a lower K^{held} cannot arise from retuning the steering vector or its strength after quantization. Table 1 shows that development ordinary fidelity alone does not determine held-out steering preservation. For example, the Llama checkpoints have nearly matched development L_0 (a 0.7% spread), but their held-out K ranges from 1.130 to 1.342 across objectives, an absolute increase of 18.8% relative to the lowest value. The effect is configuration dependent rather than a uniform ranking of objectives.

6.2 CALIBRATION GEOMETRY AND DIRECTIONAL QUANTIZATION ERROR

Our first experiment directly evaluates the quantities in the layer-level bound in section 3.1. We use Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3, and Qwen3-8B, evaluating the input-side matrix at decoder layers $4, 12, 20, 28$ for the $\mathrm{mlp.gate_proj}$ module. GPTQ and RTN use the same WikiText-2 calibration stream (16,384 tokens, chunked into sequences of length 512), group size 128, and damping fraction 0.01; only the quantizer changes. At each layer, the direction set contains eight refusal-cone directions, seven one-dimensional concept directions, and eight seeded random controls. The reported real-direction correlations exclude the random controls. For each effective intervention direction d and quantizer Q, we measure $\Delta_Q(d) = \|(W_Q - W)d\|_2^2$ and compare it with $s_\Gamma(d) = d^\mathsf{T}\Gamma^{-1}d$ as defined in Section 3.1. The bound does not by itself guarantee a monotone ordering because it is an upper bound for each direction. We therefore use Spearman's ρ to test empirically whether its direction-dependent factor ranks the realized errors. RTN serves as a negative control: unlike GPTQ, its quantization error is not optimized against Γ.

Across three model families, four decoder layers, and 3- and 4-bit GPTQ, the correlation remains high (0.902 ± 0.087 and 0.894 ± 0.105, respectively) over 12 model-layer pairs. With calibration and direction sets fixed, RTN has substantially weaker correlations on average, providing a negative control for the GPTQ-specific geometry. The result supports the ranking interpretation of the bound, but not a claim that $s_\Gamma(d)$ exactly predicts the final model-output KL for every layer.

We compare the full inverse-covariance score with inverse-variance, diagonal, norm-only, and cosine controls. The full score is consistently strongest (see Appendix A), indicating that the result depends on calibration geometry rather than direction norm alone.

Manipulating calibration coverage. Correlation alone does not establish that calibration coverage controls directional error. We therefore directly manipulate the calibration geometry as

$$\Gamma_\alpha = \Gamma_0 + \alpha\ \mathrm{mean}(\mathrm{diag}\,\Gamma_0)\, dd^\mathsf{T}, \tag{11}$$

Table 2: Primary layer-level directional-error experiment. Each entry is the Spearman correlation between $E_Q(d) = \|(W_Q - W)d\|_2^2$ and $s_\Gamma(d) = d^\mathsf{T}\Gamma^{-1}d$ over 15 real directions, excluding eight random controls. We evaluate three model families at decoder layers $4, 12, 20, 28$ with the same calibration stream.

Model	Layer	ρ_{GPTQ} (3-bit)	ρ_{RTN} (3-bit)	ρ_{GPTQ} (4-bit)	ρ_{RTN} (4-bit)
Llama-3.1-8B-Instruct	4	.968	.036	.961	.461
	12	.814	−.093	.796	.004
	20	.786	−.046	.939	−.218
	28	.964	.582	.957	−.146
Mistral-7B-Instruct-v0.3	4	.996	.039	.964	.182
	12	.914	.286	.875	.293
	20	.914	−.300	.911	−.111
	28	.886	−.350	.782	.243
Qwen3-8B	4	.714	−.721	.639	−.625
	12	.939	−.057	.936	−.218
	20	.961	.136	.979	.154
	28	.971	−.232	.993	.171

while keeping the weights, bit width, group size, and damping fixed. Let d be the normalized first principal direction of the positive-minus-negative prompt activation means at the affected layer. We use Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3, and Qwen3-8B at layers $8, 16, 24$ in `mlp.gate_proj`, with 3-bit GPTQ, group size 128, damping fraction 0.01, and 16,384 WikiText-2 calibration tokens in sequences of length 512; the reported coverage-behavior runs use $\alpha \in \{10, 30\}$. Table 3 shows that it dramatically decreases the error in the target direction. At $\alpha = 10$, target coverage reduces the directional error by at least 92% in every model–layer pair, and increasing to $\alpha = 30$ reduces it by at least 98%. For example, at layer 16 the error changes from .295 to .014/.006 for Llama and from .650 to .037/.014 for Qwen. The orthogonal control remains near its original value, whereas covering an orthogonal direction produces the complementary pattern. Thus the effect is direction-specific and strengthens monotonically with the added target coverage.

Table 3: Directional error before and after increasing calibration coverage of the target direction. Each entry reports GPTQ result under target coverage with $\alpha = 10/30$.

Model	Layer 8 (↓)	Layer 16 (↓)	Layer 24 (↓)
Llama-3.1-8B-Instruct	.227 → .012/.004	.295 → .014/.006	.287 → .013/.005
Mistral-7B-Instruct-v0.3	.013 → .001/.000	.014 → .001/.000	.014 → .001/.000
Qwen3-8B	.598 → .045/.017	.650 → .037/.014	.794 → .034/.013

The orthogonal directional component remains close to its control value, while increasing coverage of an orthogonal direction produces the complementary pattern. These results provide a mechanism-level validation of the layer-level bound: calibration coverage selectively controls the directional quantization error permitted along an intervention direction.

6.3 Intervention-Dependent Output Amplification

From directional error to model-level steering. The layer-level geometry does not, however, fully determine the final behavioral response. We therefore repeat the target-coverage intervention while measuring the signed steering-response curve under the same affected layer. For each coefficient c, we generate up to 48 new tokens for every prompt in the fixed evaluation set and apply a direction-specific binary scorer to each completion. The Safety scorer detects refusal behavior using the shared refusal-keyword rule, whereas the Persona scorer detects cheerfulness using a fixed lexicon of positive or warm cues. The behavioral rate is the fraction of completions receiving a positive score. Let $r(c)$ denote this steered behavioral rate minus the rate from unsteered generations on the same prompts, and let $r_{\mathrm{FP}}(c)$ be the corresponding full-precision teacher curve. For the sweep $\mathcal{C} = \{\pm 1, \pm 0.5\}c^*$, we define the response-curve error as $E_{\mathrm{resp}} := \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} |r(c) - r_{\mathrm{FP}}(c)|$.

Mean E_{resp} changes from .0083 to .0083 for Llama, .0146 to .0208 for Mistral, and .0375 to .0365 for Qwen. Thus, reducing local directional error is not by itself sufficient to guarantee preservation of the final steering response: it is unchanged for Llama, worsens for Mistral, and improves only modestly for Qwen. This separation is expected because the layer-level intervention controls one local component of the quantization error, whereas E_{resp} also reflects downstream transformations and their sensitivity to the perturbation.

This observation motivates the output-level Fisher analysis in Section 3.2. For the 3-bit GPTQ comparison shown in Figure 2, we fix the steering layer to decoder layer 16 and evaluate 64 seeded random directions together with safety and persona directions ; the quantization perturbation is interpolated from the shared FP teacher state, and the finite-difference Fisher estimate uses $\varepsilon = 0.05$. For each quantized model, let Δ denote its parameter difference from the shared FP teacher and consider the one-dimensional experimental slice $\text{span}\{\Delta\} \subseteq U$. To directly test its local nature, we interpolate the quantization perturbation as $\theta + t\Delta$ on the dense grid $t \in \{0.05, 0.10, \ldots, 1.00\}$.

The Fisher prediction is 1.375, 2.096, and 2.060 for Llama, Mistral, and Qwen, respectively. At the first nonzero interpolation point $t = 0.05$, the empirical ratios are 1.383, 2.130, and 2.071, showing close local agreement. As the perturbation grows, the empirical curves depart from this local prediction in model-specific ways: at $t = 1$ the ratios are 1.312 for Llama, 1.460 for Mistral, and 2.562 for Qwen. Thus the results support Fisher geometry as a local characterization of intervention-dependent amplification, while the larger-t behavior cautions against treating the local ratio as a universal model ranking rule.

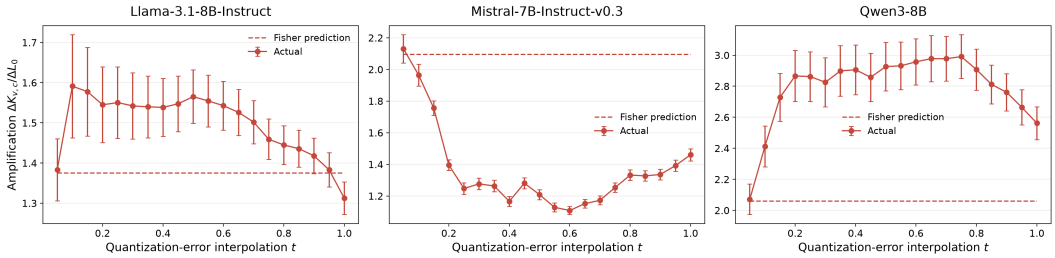

Figure 2: Fisher amplification under quantization-error interpolation for the three model families at 3 bits. The horizontal axis is the interpolation scale t in $\theta + t\Delta$, and the vertical axis is the ratio of steered to unsteered KL increases, $\Delta K_{v,c}/\Delta L_0$. Agreement at small t and divergence at larger t illustrate that the Fisher characterization is local.

Overall, the result supports the ranking and mechanism interpretation of $s_\Gamma(d)$ at the layer level, while also showing why calibration geometry alone cannot exactly predict the final model-output KL.

6.4 Steering Recovery with Unsteered Distillation

Finally, we isolate the effect of standard KL matching on unsteered teacher outputs, which was motivated by Theorem 3.2. Starting from the same GPTQ model, we compare unsteered KL distillation against the unmodified GPTQ baseline. All recovery results use 3-bit GPTQ with group size 128. The GPTQ initialization uses 96 calibration sequences of length 2048. Distillation optimizes the scale factor based on GPTQ format with Adam at learning rate 10^{-3} for 300 iterations, using a cache of 64 calibration sequences of length 512 and the top 500 teacher logits per position. Table 4 shows lower mean steered KL after distillation for all three model families, with the largest reductions for Llama and Mistral.

We also evaluate preservation across a signed strength sweep $c/c^* \in \{\pm 1, \pm 0.5\}$, rather than at only a single intervention coefficient. c^* represents the FP16 reference strength selected for each direction, and the same signed relative-strength grid is applied to every quantized method. Distillation lowers mean of response-curve error E_{resp} for Safety from $.229 \pm .144$ to $.183 \pm .061$. The lower value means that the behavior of the model is closer to that of the original full-precision model. Figure 3 shows the response curve, in which distillation follows more closely than GPTQ. Unsteered

Table 4: Recovery with unsteered scale optimization before a final GPTQ pass. Safety, Persona, and Random denote the refusal, cheerful-versus-terse, and eight-seed random directions, respectively.

Model	Method	$L_0(\downarrow)$	Safety $K(\downarrow)$	Persona $K(\downarrow)$	Random $K(\downarrow)$
Llama-3.1-8B-Instruct	GPTQ	1.16 ± 0.34	1.20 ± 0.41	0.89 ± 0.14	1.52 ± 0.22
	Distillation	0.76 ± 0.16	0.73 ± 0.16	0.53 ± 0.05	0.95 ± 0.07
Mistral-7B-Instruct-v0.3	GPTQ	0.71 ± 0.11	0.82 ± 0.05	1.50 ± 0.36	0.94 ± 0.14
	Distillation	0.69 ± 0.17	0.52 ± 0.09	0.80 ± 0.17	0.64 ± 0.10
Qwen3-8B	GPTQ	0.76 ± 0.12	1.08 ± 0.10	1.72 ± 0.11	1.75 ± 0.09
	Distillation	0.61 ± 0.13	0.81 ± 0.15	1.39 ± 0.13	1.40 ± 0.11

distillation reduces both L_0 and steered KL in many of the evaluated settings despite never observing steering vectors. However, the broader recovery comparison shows that this effect should not be interpreted as a universal superiority of a particular distillation parameterization.

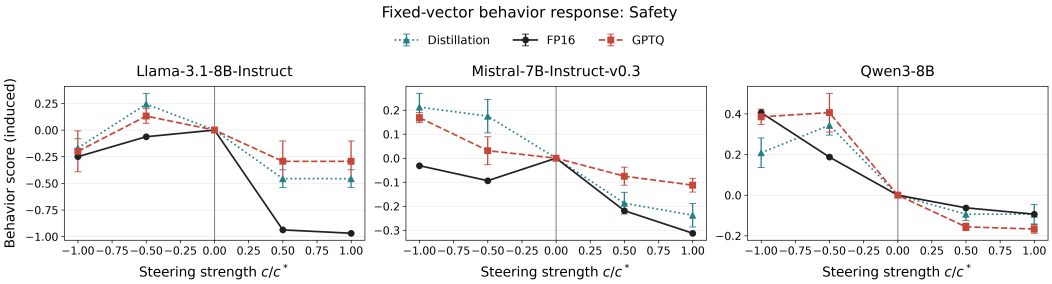

Figure 3: Behavior score for the Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3, and Qwen3-8B model families under Safety and Persona directions. The horizontal axis is the signed intervention strength relative to the FP16 reference strength.

7 Conclusion

We study whether an existing activation-steering interface is preserved after low-bit quantization. Our layer-level calibration and output-level Fisher analyses explain why ordinary fidelity alone need not control responses to the same intervention, while providing a local steered-KL guarantee under explicit conditions. Distillation using only ordinary teacher outputs can improve steered next-token agreement in many settings, but the direction-dependent response curves show that behavioral recovery is not universal and should be evaluated separately.

Limitations. Empirically, our study is limited to a small number of model families, low-bit quantization settings, and steering constructions, and the observed recovery from unsteered output matching is not universal across models, bit widths, or optimization parameterizations.

AI USE STATEMENT

We used generative AI assistants as tools to support this work. Specifically, AI tools assisted with finding related work, implementation and debugging, maintaining documentation, and drafting or polishing the manuscript. All AI-assisted code and text were carefully reviewed and, where necessary, revised by the authors; all reported measurements were obtained from the experimental artifacts and verified against logs and result files. The authors remain responsible for the final proofs, provenance, and claims.

ETHICS STATEMENT

This work uses existing models and public benchmarks and introduces no new collection of personal data. Activation steering is a dual-use technique: it can improve model controllability and help preserve desired behaviors, but it can also be used to suppress safety-related behaviors or induce undesirable model responses. Our work studies how quantization affects existing steering interfaces, including refusal-related directions, and is intended to characterize these deployment risks rather than to facilitate bypassing model safeguards. Researchers and practitioners should therefore evaluate both intended and unintended behavioral changes when applying steering or quantization methods, particularly in safety-sensitive settings.

REPRODUCIBILITY STATEMENT

Proofs are provided in Section 4, Appendix B, and Appendix C. Implementation details and experimental settings are provided in the main text. Source code required to reproduce the experiments will also be provided as supplementary material.

REFERENCES

Muhammad Alif Al Hakim, Alfan Farizki Wicaksono, and Fajri Koto. Preserving fairness and safety in quantized llms through critical weight protection. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 19831–19855, 2026.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Artem Kharinaev, Viktor Moskvoretskii, Egor Shvetsov, Kseniia Studenikina, Mikhail Bykov, and Evgeny Burnaev. Investigating the impact of quantization methods on the safety and reliability of large language models. *IEEE Access*, 2026.

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. *Proceedings of machine learning and systems*, 6:87–100, 2024.

Sheng Liu, Haotian Ye, Lei Xing, and James Y. Zou. In-context vectors: Making in context learning more effective and controllable through latent space steering. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32287–32307. PMLR, 21–27 Jul 2024.

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

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Gao Peng, Yu Qiao, and Ping Luo. Omniquant: Omnidirectionally calibrated quantization for

large language models. In *International Conference on Learning Representations*, volume 2024, pp. 45472–45496, 2024.

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

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

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 *International conference on machine learning*, pp. 38087–38099. PMLR, 2023.

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

A Simple indicator is not sufficient

To test whether the layer-level predictor is reducible to a simpler direction statistic, we compare the full inverse-covariance score $d^\mathsf{T}\Gamma^{-1}d$ with inverse-variance, diagonal-inverse, norm-only, and inverse-cosine predictors. (see Section 3.1) The configurations use Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3, and Qwen3-8B at decoder layers $4, 12, 20, 28$, targeting the `mlp.gate_proj` module. For each bit width, we use 3- or 4-bit GPTQ with group size 128 and 16,384 WikiText-2 calibration tokens split into sequences of length 512. Each layer uses 15 real directions (eight refusal-cone directions and seven concept directions).

Table 5: Predictor ablation for GPTQ directional error. The full inverse-covariance score substantially outperforms simpler direction-only indicators.

Predictor	3-bit GPTQ	4-bit GPTQ
Full inverse covariance $d^\mathsf{T}\Gamma^{-1}d$	0.902 ± 0.087	0.894 ± 0.105
Inverse variance	0.406 ± 0.298	0.372 ± 0.273
Diagonal inverse covariance	0.354 ± 0.230	0.345 ± 0.259
Norm squared $\|d\|_2^2$	0.184 ± 0.296	0.068 ± 0.458
Inverse cosine squared	0.474 ± 0.220	0.441 ± 0.289

Results are shown in Table 5. The full covariance geometry is therefore necessary for the strong ranking result: simple norm, variance, diagonal, and cosine indicators do not provide a comparably stable predictor across model families and bit widths.

B Proof of Proposition 4.1

Proposition B.1 (Sufficient condition for bounded intervention amplification)**.** *Suppose that the ordinary Fisher has non-vanishing curvature on U,*

$$u^\mathsf{T} F_0 u \geq \mu_U \|u\|_2^2, \qquad \forall u \in U, \tag{5}$$

for some $\mu_U > 0$. Further, define the maximum intervened sensitivity on U as

$$\beta_R := \sup_{\|v\|=1,\ |c|\leq R,\ u\in U,\ \|u\|=1} u^\mathsf{T} F_{v,c} u \tag{6}$$

and assume $\beta_R < \infty$. Then,

$$\kappa_R(U) \leq \frac{\beta_R}{\mu_U} < \infty. \tag{7}$$

Proof. Fix an intervention (v, c) with $\|v\|_2 = 1$ and $|c| \le R$, and let $u \in U \setminus \{0\}$. By the definition of β_R,

$$u^\mathsf{T} F_{v,c} u = \|u\|_2^2 \left(\frac{u}{\|u\|_2}\right)^\mathsf{T} F_{v,c} \left(\frac{u}{\|u\|_2}\right) \le \beta_R \|u\|_2^2.$$

Together with the assumed lower bound $u^\mathsf{T} F_0 u \ge \mu_U \|u\|_2^2$, this yields

$$\frac{u^\mathsf{T} F_{v,c} u}{u^\mathsf{T} F_0 u} \le \frac{\beta_R \|u\|_2^2}{\mu_U \|u\|_2^2} = \frac{\beta_R}{\mu_U}.$$

Taking the supremum first over $u \in U \setminus \{0\}$ and then over all interventions with $\|v\|_2 = 1$ and $|c| \le R$, we obtain

$$\kappa_R(U) \le \frac{\beta_R}{\mu_U} < \infty.$$

□

C Proof of Theorem 4.2

Theorem 3.2 (Local control of steered KL by ordinary KL)**.** *Assume that F_0 is positive definite on U. Then, for any fixed intervention (v, c) and sufficiently small $\Delta \in U$,*

$$K_{v,c}(\Delta) \le \kappa(v, c; U) L_0(\Delta) + O(\|\Delta\|^3). \tag{8}$$

If, in addition, $\kappa_R(U) < \infty$ and the third-order remainder bounds hold on a common neighborhood of $\Delta = 0$ with constants independent of (v, c) for $\|v\|_2 = 1$ and $|c| \le R$, then

$$\sup_{|v|_2=1 \ |c|\le R} K_{v,c}(\Delta) \le \kappa_R(U) L_0(\Delta) + O(|\Delta|_2^3), \tag{9}$$

where the remainder bound holds uniformly for all $\Delta \in U$ in that neighborhood.

Proof. Fix an intervention (v, c) and write $\kappa = \kappa(v, c; U)$. Since U is finite dimensional and F_0 is positive definite on U, the generalized Rayleigh quotient defining κ has a finite supremum on the unit sphere of U. Thus $0 \le \kappa < \infty$, and

$$\Delta^\mathsf{T} F_{v,c} \Delta \le \kappa \, \Delta^\mathsf{T} F_0 \Delta, \qquad \forall \Delta \in U.$$

By the local expansions stated above, write

$$L_0(\Delta) = \frac{1}{2} \Delta^\mathsf{T} F_0 \Delta + r_0(\Delta),$$
$$K_{v,c}(\Delta) = \frac{1}{2} \Delta^\mathsf{T} F_{v,c} \Delta + r_{v,c}(\Delta).$$

There exist constants $\rho > 0$ and $C_0, C_{v,c} < \infty$ such that, for every $\Delta \in U$ with $\|\Delta\|_2 \le \rho$,

$$|r_0(\Delta)| \le C_0 \|\Delta\|_2^3, \qquad |r_{v,c}(\Delta)| \le C_{v,c} \|\Delta\|_2^3.$$

Consequently,

$$\begin{aligned} K_{v,c}(\Delta) &= \frac{1}{2} \Delta^\mathsf{T} F_{v,c} \Delta + r_{v,c}(\Delta) \\ &\le \frac{\kappa}{2} \Delta^\mathsf{T} F_0 \Delta + r_{v,c}(\Delta) \\ &= \kappa L_0(\Delta) + r_{v,c}(\Delta) - \kappa r_0(\Delta) \\ &\le \kappa L_0(\Delta) + \big(C_{v,c} + \kappa C_0\big) \|\Delta\|_2^3. \end{aligned}$$

This proves the bound for a fixed intervention.

For the uniform statement, assume $\kappa_R(U) < \infty$. By the uniform remainder assumption, we may choose $\rho > 0$ and $C_R < \infty$ such that

$$|r_{v,c}(\Delta)| \le C_R \|\Delta\|_2^3$$

for every intervention with $\|v\|_2 = 1$ and $|c| \le R$, and every $\Delta \in U$ with $\|\Delta\|_2 \le \rho$. Shrinking ρ if necessary, the same neighborhood also satisfies $|r_0(\Delta)| \le C_0\|\Delta\|_2^3$.

By the definition of $\kappa_R(U)$,

$$\Delta^\mathsf{T} F_{v,c}\Delta \le \kappa_R(U)\,\Delta^\mathsf{T} F_0\Delta$$

for every admissible intervention and every $\Delta \in U$. Hence,

$$\begin{aligned} K_{v,c}(\Delta) &\le \frac{\kappa_R(U)}{2}\Delta^\mathsf{T} F_0\Delta + C_R\|\Delta\|_2^3 \\ &= \kappa_R(U)L_0(\Delta) - \kappa_R(U)r_0(\Delta) + C_R\|\Delta\|_2^3 \\ &\le \kappa_R(U)L_0(\Delta) + \big(C_R + \kappa_R(U)C_0\big)\|\Delta\|_2^3. \end{aligned}$$

The right-hand side is independent of (v, c). Taking the supremum over $\|v\|_2 = 1$ and $|c| \le R$ yields

$$\sup_{\substack{\|v\|_2=1 \\ |c|\le R}} K_{v,c}(\Delta) \le \kappa_R(U)L_0(\Delta) + \big(C_R + \kappa_R(U)C_0\big)\|\Delta\|_2^3,$$

which proves the claimed uniform bound. □