COMPRESSION-INDUCED REPRESENTATIONAL DISSONANCE: ANTICIPATING AND SUPPRESSING PREDICTION FLIPS

Anonymous authors
Paper under double-blind review

ABSTRACT

A compressed model can preserve benchmark accuracy while altering individual predictions, leading to potentially severe reliability failures. Across accuracy-preserving quantization, pruning, and low-rank compression of 58 vision models on 9 datasets, we find that models lose only 0.42 percentage points of accuracy on average while changing 9.4% of their predictions. We trace these changes to the interaction between compression error and the constraints it imposes, and to how they affect features of different energies. Compression disproportionately distorts low-variance feature directions, drowning them in noise. Whether these changes alter a prediction depends on the base model's decision margin, linking differential representational change to decision instability. This relationship generalizes across compression families and model classes, allowing prediction flips to be anticipated before evaluating the compressed model on the target inputs. Across 453 configurations of vision and language models spanning quantization, sparsification, and low-rank, we predict the flip rates with a median relative error of 5.7%. And as we focus more on preserving representation during compression, we substantially restore decision fidelity and reduce flips at matched accuracy by 50 to 93%.

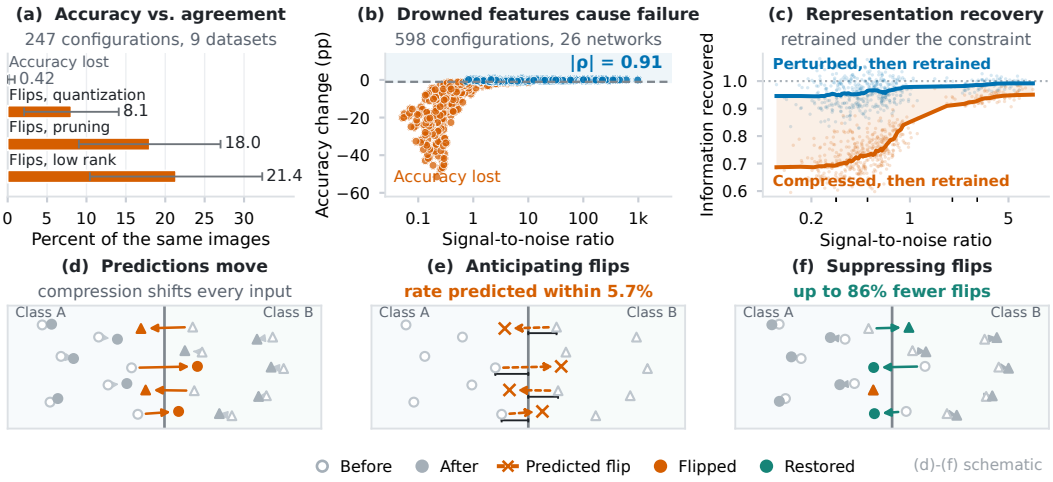


Figure 1: (a) Compression can preserve aggregate accuracy while changing many individual predictions, motivating a decision-level analysis. (b) We trace this instability to compression error overwhelming weak feature directions: feature signal-to-noise ratio provides a label-free indicator of when compression begins to damage task-relevant representations. (c) Under a persistent compression constraint, these low-SNR directions recover less than after a matched one-time perturbation, revealing a lasting representational mismatch. (d) This representation error perturbs each input's decision margin, causing some predictions to cross the decision boundary. (e) The resulting flips can therefore be anticipated from the base model's margins and compression sensitivities. (f) And preserving the base representation suppresses these shifts and restores most flipped decisions.

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.

1 INTRODUCTION

Compression reduces the cost of deploying neural networks, but preserving benchmark accuracy does not ensure that a compressed model makes the same decisions. Correct-to-incorrect and incorrect-to-correct changes can offset one another in the aggregate, leaving substantial disagreement invisible to an accuracy comparison. This discrepancy extends to generation: the evaluated language models change 28–81% of their answers to trivia and arithmetic questions while exact match changes by at most 4.5 percentage points. Figure 2 places these results alongside image classification; Table 1 reports exact match and answer disagreement together.

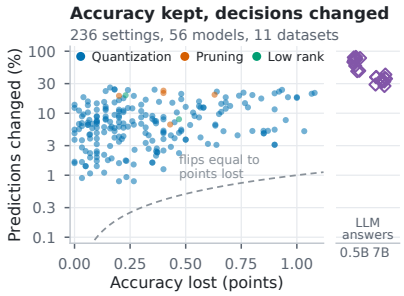


Figure 2: Accuracy lost against predictions changed, one point per accuracy-preserving compression across 9 datasets and 54 base models, hollow diamonds are language models' answers to TriviaQA and GSM8K.

We study *compression-induced representational dissonance*; changes in a model's representations and decisions that persist even when retraining compressed models (but not the perturbed ones). The trained model relies on every feature direction its classifier reads, including low-energy (specialized) ones; when compression drowns those directions in error due to reduced data precision (quantization), removed parameters (pruning) or math approximations (low-rank approximation), the model keeps relying on features that have become unreliable, and its decisions de-synchronize from the representation they were trained on, as an orchestra falls into dissonance if some instruments go out of tune. Decision fidelity to the base model is therefore a reliability property, especially in applications such as autonomous driving, as compressed models are rarely rigorously evaluated as the base model. This perspective leads to three contributions:

- **Compression causes representational dissonance.** Compression disproportionately and continuously corrupts low-energy principal directions that carry little total variance but still affect decisions. These directions recover less under a persistent compression constraint even after retraining and leave the model's decision path with unreliable features. SNR can predict accuracy retention with a Spearman correlation of 0.91 across 598 configurations.
- **Anticipating prediction flips before compressions.** We derive a signed first-order estimate of margin change from the base model's margins, layer-wise sensitivities, and a candidate compression perturbation. It anticipates flip rates with a 5.7% median relative error across 453 vision-and-language configurations spanning quantization, sparsification, and low-rank.
- **Suppressing flips and preserving decisions under compression.** We introduce spectral energy matching (SEM), a pre-distillation step that matches compression error to the energy of each layer's input, followed by feature distillation to the base representation. Together, they remove up to 86% of compression-induced disagreement at matched accuracy, averaging 83% for supervised quantization and 58% across zero-shot quantization, sparsification, and low-rank.

This work advances post-compression decision fidelity, and with it, the reliability of a compressed model, a quantity that can be measured before compressing, predicted, and optimized against.

Table 1: **Task accuracy and generated-answer fidelity.** Means ± SD across the five retained compression configurations per base model and task. EM denotes exact match; changed answers compare the normalized generated text with the base model response.

Model	Task	Configurations	Model EM (%)	Compressed EM (%)	Answers changed (%)
Qwen2.5-0.5B	TriviaQA	5	7.7	7.8 ± 1.6	76.6 ± 4.1
Qwen2.5-0.5B	GSM8K	5	1.7	2.3 ± 0.8	59.3 ± 12.1
Llama-2-7B	TriviaQA	5	41.0	43.0 ± 2.2	31.8 ± 4.0
Llama-2-7B	GSM8K	5	5.5	5.5 ± 1.1	37.6 ± 2.9

2 Related work

Prediction consistency under model changes. Churn, regression, and multiplicity study changes in individual decisions across model updates or alternative predictors (Milani Fard et al., 2016; Shen et al., 2020; Yan et al., 2021; Xie et al., 2021; Jiang et al., 2022; Bahri & Jiang, 2021; Marx et al., 2020; Black et al., 2022; Cai et al., 2022). Compression creates a related discrepancy: aggregate performance can remain nearly unchanged while individual predictions, subgroup behaviour, or retained knowledge change substantially (Hooker et al., 2019; 2020; Ogueji et al., 2022; Dutta et al., 2024; Hua et al., 2026). Recent work studies quantization-induced score shifts (Saha et al., 2026; Kiselev, 2026), output divergence in language models (Qamar et al., 2026; Zhang et al., 2026; Rababah et al., 2026), and sensitivity to inputs, scale, and training (Chang et al., 2025; Ouyang et al., 2025). These works primarily characterize instability once a compressed model is available. Closest to our problem, Wu et al. (2026) estimate flip rates from margins measured on both the original and quantized models. We instead ask whether prediction changes can be anticipated from the base model and a specified compression perturbation, without evaluating the compressed model on the target inputs.

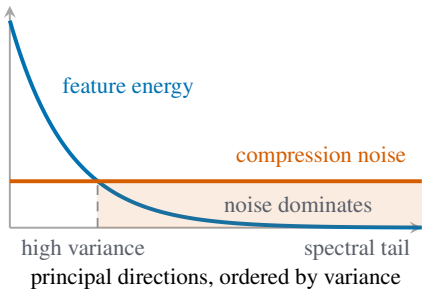


Figure 3: **Even noise does not equate to equal impact:** The SVD of the penultimate representation gives us directional variance $\lambda_k = \Sigma_{kk}^2/n$. Compression (or random noise) induces error energy ν_k along the same principal directions. In the spectral tail, small λ_k allows compression error to drown rare features ($\mathrm{SNR}_k < 1$)

Representations and precision. Representation similarity and linear probes reveal changes that task accuracy can hide (Kornblith et al., 2019; Alain & Bengio, 2017); matched accuracy can likewise conceal substantial differences in the spectrum of learned representations (Shuttleworth et al., 2025). Quantized models also differ beyond accuracy in calibration, robustness, and out-of-distribution behaviour (Bouguerra et al., 2025). Quantization methods use signal-to-noise ratios, reconstruction error, and curvature to characterize sensitivity or allocate precision (Zhao et al., 2024; Frantar et al., 2023; Nagel et al., 2020; Sakr et al., 2017; Dong et al., 2019; 2020). Our analysis instead resolves compression error in the singular-vector basis of the base representation. For $F - \bar{F} = U\Sigma V^\top$, direction v_k carries variance $\lambda_k = \Sigma_{kk}^2/n$; we compare this signal with the compression-error energy along the same direction. This exposes a regime in which many low-variance directions are overwhelmed even though they carry little total energy (Figure 3). We then ask two separate questions: whether these directions recover while the compression constraint remains active, and whether their errors project strongly enough onto the classifier to cross a decision margin. Related margin–residual decompositions certify vector-search decisions (Xiao & Cao, 2026); here the signed margin change connects directional representation error to compression-induced prediction flips.

Preserving representations under compression. Distillation from a full-precision teacher has long been combined with quantization (Mishra & Marr, 2018; Polino et al., 2018; Liu et al., 2024) and can reduce predictive churn (Jiang et al., 2022). Post-training approaches reconstruct blocks or adapt quantizer parameters (Esser et al., 2020; Li et al., 2021; Wei et al., 2022; Yuan et al., 2022; Li et al., 2023), while scaling and rotation methods reshape the coordinates in which quantization error occurs. SmoothQuant rescales channels, QuaRot, QuIP, SpinQuant, and DuQuant use rotations, and FlatQuant and OSTQuant learn richer transforms (Xiao et al., 2023; Ashkboos et al., 2024; Chee et al., 2023; Liu et al., 2025; Lin et al., 2024b; Sun et al., 2025; Hu et al., 2025). AWQ and OmniQuant further optimize quantization using calibration data (Lin et al., 2024a; Shao et al., 2024). Our objective is complementary: rather than treating reconstruction error as the endpoint, we ask whether compression preserves the base model's *decisions* at matched task performance. This motivates feature distillation and spectral energy matching in Section 6, which we evaluate against rounding, learned scaling, and our QuaRot reimplementation. Surveys provide broader taxonomies of these techniques and their composition (Wang et al., 2024; Kim et al., 2026).

3 SETUP, COMPRESSION ERROR AND DECISION FIDELITY

Let a full-precision base model produce logits $z = Wh + b$ from features $h \in \mathbb{R}^d$. A compressed model produces $z_c = (W + \Delta W)(h + \Delta h) + b + \Delta b$, giving the exact decomposition

$$\delta z = W\Delta h + \Delta W h + \Delta W \Delta h + \Delta b. \tag{1}$$

The vision experiments retain a full-precision classifier, so $\delta z = W\Delta h$. Function-preserving transformations are compared in the original output coordinates. Let $y = \arg\max_k z_k$ and $m_{yj} = z_y - z_j$. A competitor j overtakes the base model's prediction when

$$\delta z_j - \delta z_y > m_{yj}. \tag{2}$$

All-class disagreement occurs when at least one competitor replaces y, with ties resolved by the model's prediction rule. For a fixed classifier, the perturbation is $(w_j - w_y)^\top \Delta h$: decision fidelity depends jointly on representation error, its alignment with a class contrast, and the original margin. The identities of Sections 3–5 are machine-checked in Lean 4 (Appendix E).

Evaluation, populations, and reporting. We evaluate downstream tasks for prediction disagreement (which can be correct-to-incorrect, incorrect-to-correct, or incorrect-to-incorrect). Table 5 defines our experimental setup across architectures, modalities, datasets, and compression methods, as well as the compression aggressiveness and failure threshold. The two largest are distinct: the screening study (Section 4) holds 598 configurations of 26 base models, never screened, each labelled by a top-1 loss of at most 1.0 point; the anticipation population (Section 5) holds 453: 435 of 45 base models within 1.1 points of base model top-1 or 10% of base model perplexity, with at least 50 reversals, and 18 sparsified and low-rank models retrained within 3 points of base model top-1. Calibration and evaluation inputs are disjoint, and cross-model inference aggregates configurations within each base model, as the full range is too broad; please refer to (Appendix B).

4 DIRECTIONAL ERROR AND RECOVERY UNDER COMPRESSION

Compression error concentrates where the representation signal is weakest. Aggregate feature error does not reveal *where* compression changes the representation. To resolve this structure, we express the error in the principal directions of the base model. For the centred feature matrix $F - \bar{F} = U\Sigma V^\top$, define the variance of direction v_k as $\lambda_k = \Sigma_{kk}^2/n$ and the compression-induced error energy as

$$\nu_k = \mathbb{E}_x[(v_k^\top \Delta h(x))^2], \qquad \mathrm{SNR}_k = \frac{\lambda_k}{\nu_k}, \qquad \mathcal{D} = \{k : \mathrm{SNR}_k < 1\}. \tag{3}$$

A direction belongs to $\mathcal{D}$ when the compression error exceeds the variation that the base model originally represented along that direction. Because the feature spectrum decays rapidly, compression can therefore overwhelm many low-energy directions while disturbing only a small fraction of the total representation energy. This effect is pronounced for ResNet-18 on CIFAR-100. With three-bit weight-and-activation quantization, 79.5% of the 512 principal directions fall below unit SNR, yet together they account for only 12.1% of the representation variance; every direction in the lowest-variance quartile is affected. The effect also varies systematically with compression severity (Table 2). Thus, a model can retain most of its feature energy while substantially losing representations.

Low-energy directions still matter for decisions. Low variance does not imply low decision relevance. The classifier continues to read these directions, so compression-induced error in the spectral tail can alter logits even when aggregate feature energy and task accuracy change little. This distinction also appears at the accuracy-preserving compression threshold: across the cross-dataset study, the median configuration has only 9% of directions below unit SNR carrying 0.0% of the variance, while still changing 1.1–8.7% of predictions (Table 4).

The spectrum therefore describes *where* the representation is damaged, but not by itself whether a prediction changes. That depends on how the directional error projects onto the classifier and on the base model's decision margin, as formalized by Equation (2). Consistent with this separation, at matched aggregate feature-error energy, Gaussian perturbations produce approximately as many prediction changes as rounding, whereas pruning and low-rank compression produce more than their matched controls (Table 10). These comparisons concern the *rate* of changed predictions, not which individual inputs change.

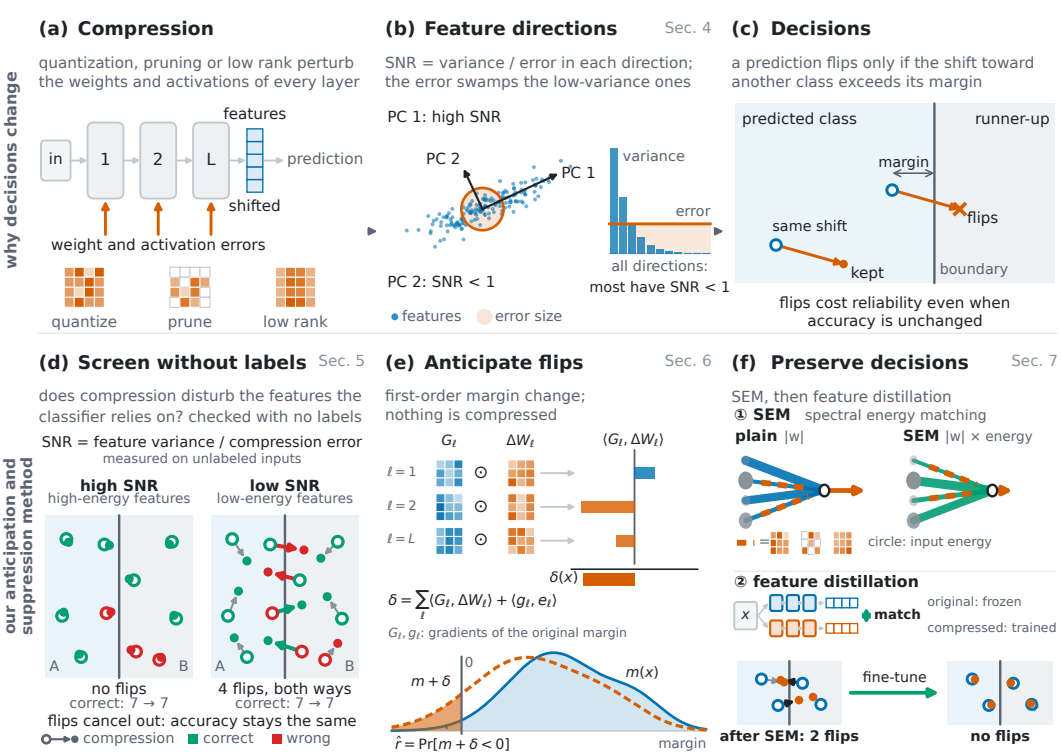


Figure 4: (a,b) Compression perturbs representations, and its error can overwhelm many low-variance feature directions despite small loss of total feature energy (Section 4). (c) These changes alter a prediction only when their signed effect exceeds the base model's decision margin. (d) Feature SNR therefore provides a label-free screen for damaging compression, even when opposite-direction flips cancel in aggregate accuracy (Section 4). (e) Base-model margin gradients combined with a candidate compression perturbation predict the resulting margin changes and flip rate before evaluating the compressed model (Section 5). (f) SEM and feature distillation instead reduce the representation error that causes these flips, restoring decision fidelity (Section 6).

Persistent constraints hinder recovery. We next ask whether adaptation restores the weak directions damaged by compression. Recoverability is measured with ridge decoders trained on adapted representations, using disjoint data for direction estimation, decoder fitting, and evaluation. Each compressed model is compared with an unconstrained control that receives a one-time perturbation matched to the same initial aggregate feature error and is trained for the same budget.

Directions starting below unit SNR recover less under persistent quantization, pruning, and low-rank constraints than under the matched perturbation (Figure 5c,d). The effect exceeds seed variation for three-bit quantization and both low-rank settings, and repeats on Tiny-ImageNet, with 77% versus 97% recovery. The gap persists after 100 epochs, so compression slows recovery of already weakened directions at a fixed adaptation budget.

The classifier determines which directional errors matter. With a fixed classification head, the change in the logit gap between the base prediction y and a competitor j can be written in the base-model principal basis as

$$\delta z_j - \delta z_y = \sum_{k=1}^{d} \underbrace{(w_j - w_y)^\top v_k}_{\text{decision sensitivity}} \underbrace{v_k^\top \Delta h}_{\text{directional error}} . \tag{4}$$

Thus, a direction can have little variance and still matter if the classifier places substantial weight on it. On cached ImageNet models, directions carrying only 0.8% of the feature variance explain a median 28% of the gap change on flipped inputs and 30% on unchanged inputs. Their effect on the prediction therefore depends not on variance alone, but on their alignment with the class contrast and on whether the resulting gap change is large enough to cross the margin in Equation (2).

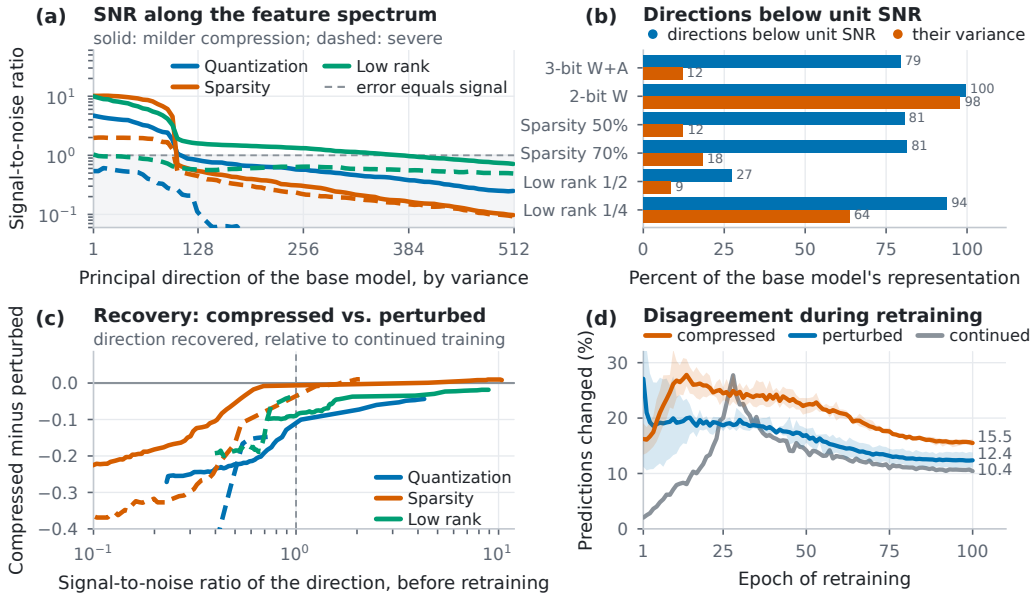


Figure 5: (a) Compression error increasingly dominates toward low-variance principal directions, driving their SNR below one. (b) Many directions are therefore lost while accounting for little representation energy. (c) These low-SNR directions recover less under a persistent compression constraint than after a matched one-time perturbation. (d) The same persistence appears at the decision level: retraining reduces disagreement, but compressed models remain farther from the base model than matched perturbations or continued training.

Table 2: **Compression preserves accuracy while leaving low-energy directions damaged.** Many principal directions fall below unit SNR, and those directions recover less under the compression constraint than after a matched perturbation. ResNet-18 on CIFAR-100, mean ± SD over 3 seeds.

	Δtop-1 (pp) Perturbed / Compressed		*directions lost*			*median SNR by quartile*		*Lost directions recovered (%)*	
Configuration	**Before**	**After**	**Count**	**Share (%)**	**Variance (%)**	**Top quartile**	**Bottom quartile**	**Compressed**	**Perturbed**
3-bit W+A	−12.9 / −8.0	+0.1 / −1.3	407	79.5	12.1	3.20	0.29	84.9 ± 0.6	99.0 ± 1.8
2-bit W	−26.9 / −76.1	+0.1 / −1.8	510	99.6	97.9	0.38	0.01	85.3 ± 0.4	98.9 ± 0.1
Sparsity 70%	−76.1 / −3.7	−0.8 / −0.2	417	81.4	18.3	1.81	0.12	93.5 ± 0.7	95.2 ± 2.0
Low-rank 1/2	−5.9 / −3.3	+0.2 / −0.2	140	27.3	8.6	6.06	0.83	94.8 ± 0.7	100.3 ± 1.5
Low-rank 1/4	−75.9 / −47.4	−0.7 / −1.4	480	93.8	63.7	0.85	0.51	91.0 ± 0.0	96.4 ± 0.9

Screening task retention without labels: Directional SNR measures representation fidelity. Projecting signal and error onto the decision margin yields a label-free statistic for task retention. On 512 calibration images, let j be the base model's runner-up and define

$$\mathrm{SNR}_{\mathrm{marg}} = \frac{\mathrm{med}_x[m_{yj}(x)^2]}{\mathbb{E}_x[(m_{yj}(x) - m^c_{yj}(x))^2]}, \qquad \mathrm{SNR}_{\mathrm{feat}} = \frac{\|F - \bar{F}\|_F^2}{\|F_c - F\|_F^2}. \tag{5}$$

Both require only unlabeled calibration inputs. The numerator measures the base-model signal and the denominator the distortion introduced by compression.

Across 598 configurations of 26 base models, margin SNR correlates with top-1 change at $|\rho| = 0.91$, feature SNR at 0.93, and variance-weighted directional retention at 0.94, compared with only 0.58 for relative weight error. The difference is structural: activation-only quantization can degrade accuracy with zero weight error, while GPTQ can improve accuracy even as weight error increases. Thresholds selected on development models classify accuracy-preserving configurations correctly in 91–97% of cases when a compression family is held out and 90–94% when an architecture family is held out. Representation-space SNR, therefore, can provide a label-free signal to compression damage.

5 ANTICIPATING PREDICTION FLIPS FROM THE BASE MODEL

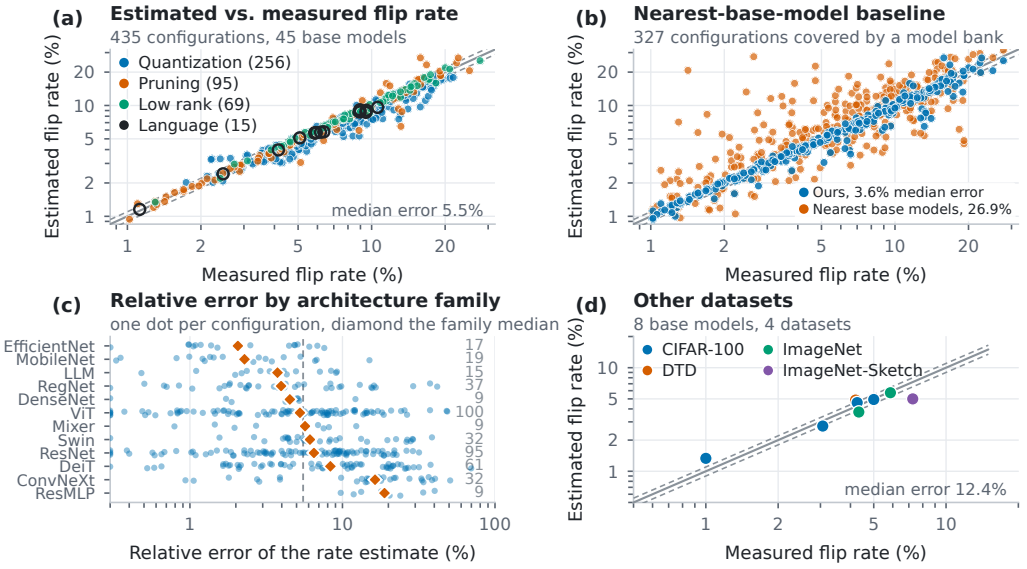


Figure 6: **Compression-induced flips can be anticipated from the base model.** Predicted rates track measured flips across configurations (a), improve over base-model-only alternatives (b), remain consistent across architecture families (c), and transfer across datasets (d).

Table 3: **Signed margin changes are more predictive of flips than error magnitude alone.** The comparison includes methods that use only the base model and a paired-margin reference that evaluates the compressed model, on the 327 configurations that every estimator covers.

Estimator	Evaluates candidate	Error (%) mean ± SD	Median (%)	p vs. anticipation
Nearest-base model transfer	No	50.7 ± 107.0	26.9	$< 10^{-4}$
Paired-margin CDF, measured shift (Wu et al., 2026)	Yes	7.9 ± 7.5	5.8	0.075
Variance form $\Phi(-m/\sqrt{V})$	No	17.0 ± 17.0	10.2	$< 10^{-4}$
Signed anticipation	No	7.0 ± 9.8	3.6	–

Predicting margin crossings. The previous section identifies where compression damages the representation; a prediction changes only when that damage closes the base model's margin. Let $G_\ell(x)$ and $g_\ell(x)$ denote the gradients of the winner–runner-up margin with respect to layer weights and activations. For candidate weight changes ΔW_ℓ and activation errors $e_\ell(x)$,

$$\delta_{\text{lin}}(x) = \sum_\ell \langle G_\ell(x), \Delta W_\ell \rangle + \sum_\ell \langle g_\ell(x), e_\ell(x) \rangle. \tag{6}$$

We anticipate a flip when the predicted margin crosses zero,

$$\widehat{I}(x) = \mathbf{1}\{m(x) + \delta_{\text{lin}}(x) < 0\}, \qquad \widehat{r} = \frac{1}{n} \sum_{i=1}^{n} \widehat{I}(x_i). \tag{7}$$

The sensitivities are computed once from the base model and reused across compression candidates.

The sign of compression error matters. Two perturbations of equal magnitude can have opposite effects if one closes the margin and the other widens it. Retaining this sign yields a 5.7% median relative error over 453 configurations spanning quantization, sparsification, and low rank, while the sign-free variance approximation is substantially less accurate. The effect persists across architectures and datasets and is comparable to methods that require observing the compressed model (Table 3; Figure 6).

Table 4: **Spectral damage and flip predictability across datasets.** Even when little representation variance is lost, compression changes decisions; The signed anticipation remains predictive across architectures and datasets.

Model	Dataset	Bits	Changed (%)	SNR < 1 (%)	Their variance (%)	Rate error (%)	Anticipation AUC	Margin AUC
MobileNetV2	CIFAR-100	W6A6	5.40	25.4	0.0	1.3	0.972	0.958
ResNet-56	CIFAR-100	W6A6	4.87	0.0	0.0	7.8	0.983	0.961
ViT-B/16 (in21k)	CIFAR-100	W5A6	1.13	14.7	0.2	33.3	0.994	0.978
ViT-B/16 (in21k, b)	CIFAR-100	W6A6	3.40	50.1	0.6	10.9	0.978	0.965
ViT-B/16 (in21k, in1k)	DTD	W6A6	4.67	12.2	0.1	15.9	0.987	0.948
DeiT-Ti/16	ImageNet	W6A6	7.13	0.5	0.0	2.3	0.920	0.924
ViT-S/16	ImageNet	W8A6	4.87	0.3	0.0	13.8	0.947	0.947
ViT-B/16	ImageNet-Sketch	W7A8	8.73	6.4	0.0	31.2	0.937	0.912

How to identify fragile inputs. The score $-(m + \delta_{\text{lin}})$ also ranks which inputs are likely to flip, outperforming the base margin alone in 6 of the 8 pairs (Table 4). Thus, fragility is not determined only by proximity to the decision boundary, but by whether the candidate compression pushes an input toward it. The same behavior holds across CIFAR-100, DTD, ImageNet, and ImageNet-Sketch (Table 4).

6 Preserving decisions under compression

Energy-aware compression with SEM. Consider a linear layer $y = Wx$ and let $d_i = \sqrt{\mathbb{E}_x[x_i^2]}$ denote the RMS energy of input dimension i, with $D = \text{diag}(d_1, \ldots, d_p)$. The output error produced by a compressed weight $\widehat{W}$ is

$$\mathbb{E}_x \|(\widehat{W} - W)x\|_2^2 = \text{tr}\left[(\widehat{W} - W)\Sigma_x(\widehat{W} - W)^\top\right].$$

Approximating the input second-moment matrix by its diagonal gives the SEM objective

$$\mathcal{E}_{\text{SEM}}(\widehat{W}) = \|(\widehat{W} - W)D\|_F^2, \tag{8}$$

so errors on dimensions carrying more activation energy are penalized more. It recovers the Wanda score $|w_{ji}|d_i$ (Sun et al., 2024) for sparsification and activation-aware SVD $(WD)_r D^{-1}$ (Yuan et al., 2023) for rank r. For quantization, SEM uses the function-preserving transform $T = \text{diag}(s)H$, where H mixes the input dimensions and s is fitted to minimize quantized layer-output error. Algorithm 1 and Appendix D give the family-specific procedures.

Feature distillation to recover representations. After compression, we adapt the model on unlabeled inputs by matching its representation to that of the frozen base model,

$$\mathcal{L}_{\text{feat}} = \mathbb{E}_x \|h_c(x) - h(x)\|_2^2. \tag{9}$$

The original classifier is retained, so reducing feature error also bounds the change in each class contrast, $|(w_j - w_y)^\top \Delta h| \leq \|w_j - w_y\|_2 \|\Delta h\|_2$. SEM is thus a pre-distillation step, motivated by the poor recovery of weak directions under a persistent constraint (Section 4): it reduces the error distillation must undo and suppresses flips that distillation alone leaves (Figure 7).

Preserving representation restores decisions. The two stages are complementary. Across quantization, sparsification, and low rank, SEM with feature distillation removes up to 86% of the prediction changes introduced by one-shot compression while preserving task performance (Figure 7; examples in Figure 8). The mean reduction is 83% for supervised quantization and 56% across the zero-shot evaluations. The same trend extends to zero-shot CLIP and language models (Tables 9 and 15), indicating that the benefit is not specific to one architecture or compression family.

Decision sensitivity helps information allocation. The anticipation of Section 5 provides a complementary use of the same analysis. Under a fixed bit budget, we remove precision from layers where the predicted increase in flip rate is smallest. This reduces held-out disagreement relative to uniform, energy-based, and Hessian-based allocation, using predicted decision sensitivity to guide where compression is applied (Table 14; Appendix D).

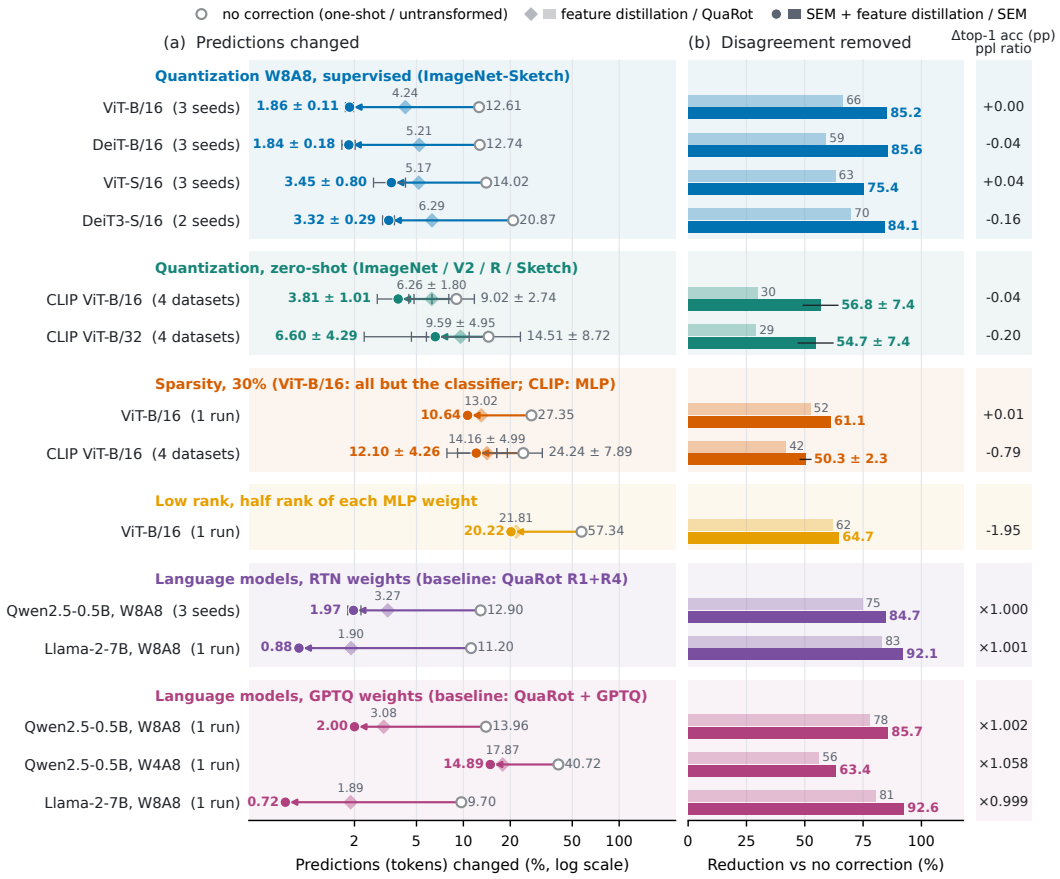


Figure 7: **Results:** Focusing on preserving representation reduces prediction disagreement; (a) Prediction changes after one-shot compression, feature distillation, and SEM with feature distillation. (b) Fraction of one-shot disagreement removed; right: corresponding Δtop-1.

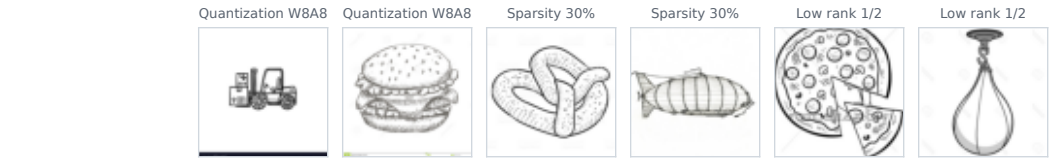


Figure 8: Examples of compression-induced flips restored by SEM while all models have similar accuracy. Feature distillation (FD) alone preserves the wrong compressed prediction, whereas SEM with feature distillation restores the base-model decision.

7 Conclusion

Advances in model compression can preserve aggregate performance while quietly changing individual decisions. We show that these changes have a coherent origin: compression disproportionately damages weak representation directions, their recovery can remain constrained, and prediction flips occur when this damage exceeds the original decision margin. This structure lets us diagnose degradation without labels, anticipate flips before evaluating the compressed model, and suppress them by preserving the representation itself. Across quantization, sparsification, and low-rank compression, our work helps preserve decision fidelity for a reliable, efficient, and accurate compression.

AI USE STATEMENT

In this work, we used generative AI tools to assist with writing and language polishing, including reformulation of text originally drafted in bad English; to assist with code and computer-use tasks; to support literature retrieval and discovery; and to assist with formalizing and checking mathematical arguments (Lean 4). We have not used generative AI tools in any other undisclosed category. We reviewed all AI-assisted content, verified the technical and scientific statements, and revised the resulting text and code as necessary. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The supplementary material contains the complete Lean 4 development and code that reproduces the key results; the full experiments were run on an Nvidia H100/H200 cluster. Appendix A specifies the statistical units underlying the reported results. The algebraic identities used in the analysis have machine-checked proofs (Appendix E); these verify the identities themselves, not the empirical error of any estimate.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *International Conference on Learning Representations, Workshop Track*, 2017.

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. *Advances in Neural Information Processing Systems*, 2024.

Dara Bahri and Heinrich Jiang. Locally adaptive label smoothing improves predictive churn. *International Conference on Machine Learning*, 2021.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. *ACM Conference on Fairness, Accountability, and Transparency*, 2022.

Aymen Bouguerra, Daniel Montoya, Alexandra Gomez-Villa, Chokri Mraidha, and Fabio Arnez. Less precise can be more reliable: A systematic evaluation of quantization's impact on VLMs beyond accuracy. arXiv preprint arXiv:2509.21173, 2025.

Deng Cai, Elman Mansimov, Yi-An Lai, Yixuan Su, Lei Shu, and Yi Zhang. Measuring and reducing model update regression in structured prediction for NLP. *Advances in Neural Information Processing Systems*, 2022.

Ting-Yun Chang, Muru Zhang, Jesse Thomason, and Robin Jia. Why do some inputs break low-bit LLM quantization? *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3410–3429, 2025.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa. QuIP: 2-bit quantization of large language models with guarantees. *Advances in Neural Information Processing Systems*, 2023.

Zhen Dong, Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. HAWQ: Hessian aware quantization of neural networks with mixed-precision. *IEEE/CVF International Conference on Computer Vision*, 2019.

Zhen Dong, Zhewei Yao, Yaohui Cai, Daiyaan Arfeen, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. HAWQ-V2: Hessian aware trace-weighted quantization of neural networks. *Advances in Neural Information Processing Systems*, 2020.

Abhinav Dutta, Sanjeev Krishnan, Nipun Kwatra, and Ramachandran Ramjee. Accuracy is not all you need. *Advances in Neural Information Processing Systems*, 2024.

Steven K. Esser, Jeffrey L. McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S. Modha. Learned step size quantization. *International Conference on Learning Representations*, 2020.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. *International Conference on Learning Representations*, 2023.

Mitchell A. Gordon, Kevin Duh, and Nicholas Andrews. Compressing BERT: Studying the effects of weight pruning on transfer learning. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, 2020.

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2019.

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058*, 2020.

Xing Hu, Yuan Cheng, Dawei Yang, Zukang Xu, Sifan Zhou, Jiangyong Yu, Chen Xu, Zhixuan Chen, Zhe Jiang, and Zhihang Yuan. OSTQuant: Refining large language model quantization with orthogonal and scaling transformations for better distribution fitting. *International Conference on Learning Representations*, 2025.

Stanley Z. Hua, Sanae Lotfi, and Irene Y. Chen. Investigating social bias changes in quantized large language models. *Conference on Language Modeling*, 2026.

Heinrich Jiang, Harikrishna Narasimhan, Dara Bahri, Andrew Cotter, and Afshin Rostamizadeh. Churn reduction via distillation. *International Conference on Learning Representations*, 2022.

Minjun Kim, Jaehyeon Choi, Hyunwoo Yang, Jongjin Kim, Jinho Song, and U Kang. Prune-then-quantize or quantize-then-prune? Understanding the impact of compression order in joint model compression. *International Conference on Learning Representations*, 2026.

O. M. Kiselev. Boundary-aware quantization: Finite-scale decision geometry of neural classifiers. *arXiv preprint arXiv:2607.01478*, 2026.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. *International Conference on Machine Learning*, 2019.

Tanishq Kumar, Zachary Ankner, Benjamin F. Spector, Blake Bordelon, Niklas Muennighoff, Mansheej Paul, Cengiz Pehlevan, Christopher Ré, and Aditi Raghunathan. Scaling laws for precision. In *International Conference on Learning Representations*, 2025.

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. BRECQ: Pushing the limit of post-training quantization by block reconstruction. *International Conference on Learning Representations*, 2021.

Zhikai Li, Junrui Xiao, Lianwei Yang, and Qingyi Gu. RepQ-ViT: Scale reparameterization for post-training quantization of vision transformers. *IEEE/CVF International Conference on Computer Vision*, 2023.

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*, 2024a.

Haokun Lin, Haobo Xu, Yichen Wu, Jingzhi Cui, Yingtao Zhang, Linzhan Mou, Linqi Song, Zhenan Sun, and Ying Wei. DuQuant: Distributing outliers via dual transformation makes stronger quantized LLMs. *Advances in Neural Information Processing Systems*, 2024b.

Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. LLM-QAT: Data-free quantization aware training for large language models. *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 467–484, 2024.

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. *International Conference on Learning Representations*, 2025.

Charles Marx, Flavio P. Calmon, and Berk Ustun. Predictive multiplicity in classification. *International Conference on Machine Learning*, 2020.

Mahdi Milani Fard, Quentin Cormier, Kevin Canini, and Maya Gupta. Launch and iterate: Reducing prediction churn. *Advances in Neural Information Processing Systems*, 2016.

Asit Mishra and Debbie Marr. Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy. *International Conference on Learning Representations*, 2018.

Markus Nagel, Rana Ali Amjad, Mart van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? Adaptive rounding for post-training quantization. *International Conference on Machine Learning*, 2020.

Kelechi Ogueji, Orevaoghene Ahia, Gbemileke Onilude, Sebastian Gehrmann, Sara Hooker, and Julia Kreutzer. Intriguing properties of compression on multilingual models. *Conference on Empirical Methods in Natural Language Processing*, 2022.

Xu Ouyang, Tao Ge, Thomas Hartvigsen, Zhisong Zhang, Haitao Mi, and Dong Yu. Low-bit quantization favors undertrained LLMs. *Annual Meeting of the Association for Computational Linguistics*, pp. 32338–32348, 2025.

Antonio Polino, Razvan Pascanu, and Dan Alistarh. Model compression via distillation and quantization. *International Conference on Learning Representations*, 2018.

Shahzeb Qamar, Lorenz Sparrenberg, Christian Bauckhage, Baha Rababah, Carson K. Leung, Murat Kantarcioglu, Cuneyt Gurcan Akcora, and Rafet Sifa. Accuracy is not enough: A divergence-based approach to evaluate fidelity loss in quantized LLMs. *arXiv preprint arXiv:2609.07664*, 2026.

Baha Rababah, Shahzeb Qamar, Lorenz Sparrenberg, Rafet Sifa, Murat Kantarcioglu, Cuneyt Gurcan Akcora, and Carson K. Leung. The illusion of equivalency: Statistical characterization of quantization effects in LLMs. *arXiv preprint arXiv:2607.08734*, 2026.

Sourajit Saha, Shubhashis Roy Dipta, Shaswati Saha, Nobin Sarwar, and Yuxuan Jiang. Before the flip: Measuring hidden score shifts in quantized vision language models before the answer changes for visual question answering. *arXiv preprint arXiv:2609.06922*, 2026.

Charbel Sakr, Yongjune Kim, and Naresh Shanbhag. Analytical guarantees on numerical precision of deep neural networks. *Proceedings of the 34th International Conference on Machine Learning*, 70:3007–3016, 2017.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. OmniQuant: Omnidirectionally calibrated quantization for large language models. *International Conference on Learning Representations*, 2024.

Yantao Shen, Yuanjun Xiong, Wei Xia, and Stefano Soatto. Towards backward-compatible representation learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Reece Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. LoRA vs full fine-tuning: An illusion of equivalence. *Advances in Neural Information Processing Systems*, 2025.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. A simple and effective pruning approach for large language models. *International Conference on Learning Representations*, 2024.

Yuxuan Sun, Ruikang Liu, Haoli Bai, Han Bao, Kang Zhao, Yuening Li, Jiaxin Hu, Xianzhi Yu, Lu Hou, Chun Yuan, Xin Jiang, Wulong Liu, and Jun Yao. FlatQuant: Flatness matters for LLM quantization. *International Conference on Machine Learning*, 2025.

Hidenori Tanaka, Daniel Kunin, Daniel L. K. Yamins, and Surya Ganguli. Pruning neural networks without any data by iteratively conserving synaptic flow. In *Advances in Neural Information Processing Systems*, 2020.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. *Advances in Neural Information Processing Systems*, 2018.

Wenxiao Wang, Wei Chen, Yicong Luo, Yongliu Long, Zhengkai Lin, Liye Zhang, Binbin Lin, Deng Cai, and Xiaofei He. Model compression and efficient inference for large language models: A survey. *arXiv preprint arXiv:2402.09748*, 2024.

Xiuying Wei, Ruihao Gong, Yuhang Li, Xianglong Liu, and Fengwei Yu. QDrop: Randomly dropping quantization for extremely low-bit post-training quantization. *International Conference on Learning Representations*, 2022.

Zekun Wu, Swati Dhiman, and Adriano Koshiyama. Which decisions low-bit quantization breaks, and how to predict them. *arXiv preprint arXiv:2608.06564*, 2026.

Wenxuan Xiao and Xu Cao. When does low-bit quantization preserve the decisions of vector search? *arXiv preprint arXiv:2609.09854*, 2026.

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

Yuqing Xie, Yi-An Lai, Yuanjun Xiong, Yi Zhang, and Stefano Soatto. Regression bugs are in your model! Measuring, reducing and analyzing regressions in NLP model updates. *Annual Meeting of the Association for Computational Linguistics*, 2021.

Sijie Yan, Yuanjun Xiong, Kaustav Kundu, Shuo Yang, Siqi Deng, Meng Wang, Wei Xia, and Stefano Soatto. Positive-congruent training: Towards regression-free model updates. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Zhihang Yuan, Chenhao Xue, Yiqi Chen, Qiang Wu, and Guangyu Sun. PTQ4ViT: Post-training quantization for vision transformers with twin uniform quantization. *European Conference on Computer Vision*, 2022.

Zhihang Yuan, Yuzhang Shang, Yue Song, Qiang Wu, Yan Yan, and Guangyu Sun. ASVD: Activation-aware singular value decomposition for compressing large language models. arXiv preprint arXiv:2312.05821, 2023.

Yanzhe Zhang, Sanmi Koyejo, and Diyi Yang. The illusion of robustness: Aggregate accuracy hides prediction flips under task-irrelevant context. *arXiv preprint arXiv:2607.12963*, 2026.

Tianchen Zhao, Xuefei Ning, Tongcheng Fang, Enshu Liu, Guyue Huang, Zinan Lin, Shengen Yan, Guohao Dai, and Yu Wang. MixDQ: Memory-efficient few-step text-to-image diffusion models with metric-decoupled mixed precision quantization. *European Conference on Computer Vision*, 2024.

APPENDIX

A EXPERIMENTAL DATA AND STATISTICAL ANALYSIS

The spectral and recovery analyses use one ResNet-18 base model on CIFAR-100 and Tiny-ImageNet; anticipation uses per-configuration reversal rates and the base models' sensitivities; suppression uses the adapted models' predictions on held-out inputs. The cross-dataset rows of Table 4 come from a separate experiment with its own precision and evaluation set.

Table 5 defines every population the paper quotes.

Table 5: **Every population in the paper**, with its screen and where it appears.

Population	Configurations	Models	Data	Screen or selection	Used in
Threshold population (vision)	247	58	9 datasets	each base model–dataset pair at its own compression threshold: strongest configuration within 1.1 pp, one step below it, a failing step above; base models below 93% top-1	Fig. 1a
Question answering (language)	20	2	TriviaQA, GSM8K	perplexity within 10% of the base model; W8A8, W7A8, W6A8, A8, and 10% sparsity with A8	Fig. 2, Table 1
Screening study	598	26	ImageNet-Sketch	label: top-1 loss at most 1.0 pp (199 kept, 399 not); no screen on the population	Sec. 4, Fig. 1b
Candidate set	602	45	ImageNet-Sketch, WikiText	every candidate before screening	Table 8
Anticipation population	435 (420 vision, 15 language)	45	ImageNet-Sketch, WikiText	vision within 1.1 pp of base model top-1, language within 10% of base model perplexity, at least 50 reversals	Sec. 5, Fig. 6a
Retrained compressed models	18 (10 ViT-B/16, 8 CLIP ViT-B/16 dataset pairs)	2	ImageNet-Sketch; ImageNet, V2, R, Sketch	sparsified and low-rank models retrained with feature distillation or SEM, within 3 pp of base model top-1; with the row above, the 453 configurations of Sec. 5	Sec. 5
Candidate-set subset	327	43	ImageNet-Sketch	anticipation population restricted to configurations the candidate set covers (weight-only quantization, pruning, low rank)	Table 3, Fig. 6b
Cross-dataset spectral study	8 pairs	8	4 datasets	each pair at its own compression threshold, base models below 93%, own evaluation set	Table 4
Recovery study	5 configurations, 3 seeds	1	CIFAR-100, Tiny-ImageNet	destructive configurations chosen to expose the tail (Table 2)	Sec. 4
Supervised suppression	4 base models	4	ImageNet-Sketch	W8A8, adapted models within 0.2 pp of base model top-1	Fig. 7, Table 11
Sparsity and low-rank suppression	3 (ViT-B/16 sparsity and low rank; CLIP ViT-B/16 sparsity on 4 datasets)	2	ImageNet-Sketch; ImageNet, V2, R, Sketch	30% sparsity, half rank of the MLP weights; SEM within 2 pp of base model top-1	Fig. 7, Table 11
Zero-shot suppression	8 pairs	2	ImageNet, V2, R, Sketch	image encoder at the pair's compression threshold, text classifier fixed	Fig. 7, Table 9
Language-model suppression	13	2	held-out text	W8A8 to W4A4, round-to-nearest and GPTQ weights; rows within 10% of base model perplexity shown in Fig. 7	Fig. 7, Table 15
Restored examples	6 images	1	ImageNet-Sketch	base model correct, one-shot and feature distillation wrong, SEM with feature distillation correct; selected; the quantization examples come from an additional W8A8 run under the protocol of App. D	Fig. 8

Cross-model comparisons use the base model as the replication unit. The paired Wilcoxon tests in Table 3 compare each base model's median relative error across matched configurations. All tests are two-sided and unadjusted. Standard deviations describe sample variation across the seeds or datasets specified in each table; they are not standard errors. These conventions preserve the distinction between repeated compression configurations, training seeds, and distinct models.

B EVALUATION CONVENTIONS

Protocol. Vision results use ImageNet-pretrained base models evaluated on ImageNet-Sketch unless stated, language results Qwen2.5-0.5B and Llama-2-7B; calibration inputs are always disjoint from evaluation inputs. Anticipated reversals compare the original winner with its original runner-up; disagreement compares the full argmax predictions. Language anticipation uses teacher-forced next-token decisions, while question-answering results compare generated responses. Comparisons are made at matched accuracy with added arithmetic stated. For cross-model summaries, the independent unit is the *base model*: configurations on one base model are not independent replications. Recovery results report variation over seeds.

Language-model protocol. The question-answering study (Table 1, Table 12) uses 200 TriviaQA (rc.nocontext) and 200 GSM8K questions, a zero-shot prompt of the form "Question: … Answer:", greedy decoding with a single beam and 20 new tokens, so that base model and compressed model are deterministic and every difference is the compression's; exact match uses the alias set for TriviaQA and the final number for GSM8K. The compressions are round-to-nearest per-channel weights and per-token activations with no calibration and no GPTQ, which is why their W8A8 perplexity ratios lie at 1.05–1.09; the transform comparisons of Table 15 report the perplexity ratio beside every rate. The low exact match of Qwen2.5-0.5B means that a share of its answer changes is movement between wrong answers; the Llama-2-7B rows, at 41% exact match on TriviaQA, carry the same pattern.

C RECOVERY PROTOCOL

Training conditions. The base model is a ResNet-18 trained from scratch on CIFAR-100 and, separately, on Tiny-ImageNet at 64 pixels. Every condition starts from that checkpoint and trains 100 further epochs on the task loss only, one-cycle SGD with Nesterov momentum, three seeds; the base model's features are never a training target. Constraints act on every convolution but the stem, classifier in full precision: fake quantization with free latent weights, per-channel rounding and a straight-through gradient, joint configurations adding a per-layer affine activation quantizer set from running minima and maxima; a global magnitude mask applied on every forward pass; a trainable factorisation at the stated rank fraction. The matched perturbation adds one Gaussian draw to the same weights, scaled by bisection so its initial change of the penultimate features equals the constraint's, then trains free; for the weight-only rows (2-bit W in Table 2) the draw is instead scaled to the per-layer energy of the rounding error. Continued training of the base model is the reference. The perturbation is matched on aggregate damage, not per direction: in the weight-only runs its initial per-direction ratio has Spearman correlation 0.92 with the constraint's at three bits and 0.61 at two, and at three bits it leaves 219 directions below unit ratio against the constraint's 355, so the controls match total damage rather than the complete directional error profile.

Decoding. Three disjoint splits of the $10{,}000$ held-out images: $4{,}000$ define the base model's principal directions, per-direction median thresholds and initial signal-to-noise ratios; $3{,}000$ fit one ridge decoder per adapted network for all 512 binary targets; $3{,}000$ score. Accuracy is chance-adjusted, divided by the base model's own under the identical protocol, and reported as the variance-weighted mean over a set of directions relative to continued training.

D METHOD DETAILS

Feature distillation. The objective is the squared error between the compressed model's penultimate features and the base model's on the same inputs, averaged over the batch. No labels and no task loss are used, and the base model's head is kept. For ViT-B/16 at W8A8: $20{,}000$ ImageNet-Sketch training images and a disjoint set of $10{,}000$ evaluation images, $4{,}000$ steps at batch 128, learning rate 2×10^{-6} on the weights and 2×10^{-5} on the quantizer steps, which are trained jointly. The proposed configurations add less than five percent arithmetic. Table 13 also reports dense-rotation references and their larger cost.

Sparsity and low rank. The sparsity and low-rank rows of Figure 7 use unlabeled training images, $1{,}500$ steps at batch 32, learning rate 3×10^{-5} with a one-cycle schedule, and every arm (one-shot,

Algorithm 1 Spectral energy matching (SEM), applied before feature distillation

Require: base model with linear layers $W_\ell \in \mathbb{R}^{o\times d}$; unlabeled calibration inputs $\mathcal{X}$; a compression (sparsification, low rank, or quantization)

for each linear layer ℓ with input activations $a(x) \in \mathbb{R}^d$ **do**
  $e_j \leftarrow \big(\mathbb{E}_{x\in\mathcal{X}}[a_j(x)^2]\big)^{1/2}$ for every input dimension j ▷ energy of the layer input
  if sparsification **then**
    keep the weights with the largest $|W_{ij}|\, e_j^{\alpha}$ and set the others to zero; the mask stays fixed
  else if low rank **then**
    $U\Sigma V^\top \leftarrow \mathrm{SVD}\big(W\,\mathrm{diag}(e^{\alpha})\big)$; $W \leftarrow U_r\Sigma_r V_r^\top\,\mathrm{diag}(e^{-\alpha})$
  else if quantization **then**
    fit s to minimise the expected output error of the quantized layer on $\mathcal{X}$
    $a \leftarrow H(a\odot s)$, $W \leftarrow W\,\mathrm{diag}(s)^{-1}H$ ▷ same function; s folds into the previous layer
    quantize W and a in these coordinates
  end if
end for
fine-tune the compressed model on $\mathcal{X}$ so that its features match the base model's (feature distillation)

feature distillation, SEM with feature distillation) the same images and budget. Sparsity is per-layer magnitude sparsity, ramped to 30% over the first half of training: of every linear and convolutional layer except the classifier for ViT-B/16, and of the MLP layers of the image encoder for CLIP ViT-B/16, which adapts on CC3M image–caption pairs. Low rank replaces each MLP weight of ViT-B/16 by its two rank-r SVD factors, r half the smaller dimension, and retrains the factors. The quantization examples of Figure 8 come from a W8A8 run of ViT-B/16 under this protocol, used only for those examples.

Quantization-aware training for zero-shot encoders. The compressed model of Table 9 keeps a full-precision latent weight, quantizes it in the forward pass and passes the gradient through the rounding unchanged (Bengio et al., 2013), on both weights and activations; the weight scale is fixed at the base model's weights rather than recomputed from the latent weight, which otherwise lets a few weights widen a channel's scale and coarsen the rest of it. The objective is the squared error to the base model's unit-norm image features on the proxy images, and no label of any evaluation set is used.

Cost of anticipation. The sensitivities cost 3.5 forward-equivalents per input, paid once (a forward pass is 7.4 ms per image), and each further candidate costs 8 ms to construct its weight change and contract it, with no compressed model built or run. A compressed model evaluation that matches the anticipation's accuracy needs 2,080 images through base model and compressed model so the anticipation is cheaper from the first candidate on a set of 512 inputs and from the 5th on 5,000. Memory is set by the per-sample sensitivities of one batch, which are contracted with every candidate as they are produced; what is kept per input is one scalar per layer and candidate.

Spectral energy matching. For row-vector inputs and an invertible T,

$$(xT)(WT^{-\top})^\top = xW^\top, \tag{10}$$

which preserves the full-precision layer function while changing its quantization coordinates. We use $T = \mathrm{diag}(s)H$, H a fast flat transform and $s = \sigma \odot \exp(m)$: the sign pattern σ is the best of four random draws and the magnitude m is fitted by Adam for 300 steps at learning rate 0.05 on the layer's calibration activations, under the expected squared output error of the quantized layer. The objective accounts for both quantizers: the activation step is chosen per token from that token's maximum and is fixed by the quantizer, giving the term $\mathbb{E}_t[(\max_c |x_{tc}|/Q_p^a)^2]\|W\|_F^2$; the weight step is learned, giving a term proportional to the per-row variance. Both σ and m fold into one per-channel scale, so the additional inference arithmetic is due to H; mixing only keeps σ and drops m.

Algorithm 1 summarizes SEM for sparsification, low rank, and quantization. For sparsification and low rank, the strength $\alpha \in \{0, 0.25, \ldots, 1.5\}$ is chosen per layer by that layer's output error on $\mathcal{X}$; $\alpha = 1$ is the objective of Equation (8). H is a fast orthogonal mixing transform.

Precision allocation. Each layer takes an activation width from $\{8, 7, 6, 5\}$ under a fixed average width weighted by the layer's activation count, at budgets 7, 6.5 and 6: from the widest grid, the layer whose predicted flip rate rises least per bit removed is stepped down until the budget is met. The rounding-energy rule is the same greedy on rounding energy; uniform allocation exists only at integer budgets. Widths are chosen on one half of the input set and every rule is scored on the other.

The variance form. With A_ℓ a layer's calibration activations and $g_{\ell i}$ the token vector of margin gradients for output channel i, the variance in $\Phi(-m/\sqrt{V})$ is $V(x) = \sum_\ell \sum_i s_{\ell i}^2\, g_{\ell i}^\top A_\ell A_\ell^\top g_{\ell i} + V_{\mathrm{a}}(x)$, with $s_{\ell i}^2$ the channel's step variance and V_{a} the activation term. It assumes zero-mean independent entries of ΔW with per-channel variance, so layers add and no cross-layer covariance is carried; that suits rounding and not the deterministic, structured changes of pruning and low rank, where the signed form, which keeps the actual ΔW, is markedly better (Table 3).

E Machine-checked statements

The supplementary Lean 4 development contains 53 machine-checked statements covering logit identities, margin crossings, spectral decompositions, function-preserving transforms, and first-order remainders. They are algebraic: they fix what each quantity in the paper is and how the quantities relate, and they do not bound the size of the first-order remainder for a nonlinear network, which Appendix F measures instead. Table 6 summarizes their relationship to the analysis. The accompanying axiom audit and proof sources provide the complete formal development.

F Further checks on anticipation

The remainder. Equation (6) is the differential of the margin in every perturbed weight and activation; `forecast_is_differential` gives its remainder as $o(\|\Delta\theta\|)$ and `two_layer_residual` makes it the product of the layers' perturbations for composed linear layers. Neither fixes its size for a nonlinear network, and a zero-mean remainder does not by itself make a crossing rate accurate. Table 7 measures it per input on the saved population: for rounding the residual is no larger in the smallest-margin quartile than elsewhere (0.95 against 0.80), false positives and false negatives are each a sizeable share of the measured flips (39% and 48%) and cancel to a net count error of 6.5%, and anticipation and measurement agree on 73% of the smallest-margin decile. The small net rate error is consistent with cancellation between these two types of input-level error; for pruning and low rank the remainder is small and per-input agreement higher.

Candidate populations and event coverage. The main rate analysis uses a task-retention screen. Table 8 scores the anticipated rate on every candidate with the screen reported separately; the label-free screen of Section 4 is what a practitioner applies instead. Per base model, 34 of the 45 base models sit under 10% median error and the worst is at 39%. The strict event covers a median 87% of all-class disagreement, never under 57%, per-model medians 71–93%; dividing by the other base models' coverage predicts the all-class rate to 10.0% error, and one backward pass per further competitor extends the anticipation to the top five, which cover 99% and track the all-class rate to 13%.

Directional attribution and harmful changes. With the head fixed, Equation (4) attributes each input's gap change to the base model's principal directions: on cached ImageNet base models the directions below unit ratio carry a median 28% of the gap change on inputs that flipped and 30% on inputs that did not, while holding 0.8% of the variance. A change is harmful when the base model was right and the compressed model is wrong; on the cached base models these are 12% of changes.

G Additional datasets and zero-shot evaluation

Table 4 reports directional error and anticipation for eight model–dataset pairs on CIFAR-100, DTD, ImageNet, and ImageNet-Sketch. Quantization configurations satisfy the task-retention tolerance. The source population for the introductory vision plot excludes base models with top-1 accuracy at

Table 6: **Machine-checked statements**: theorem names and what each establishes, over $\mathbb{R}$ with Mathlib (first block) and over integer scores in core Lean (second).

Where	Theorem	Statement
Eq. (1)	`logit_change_identity`	the logit change equals $W\Delta h + \Delta W h + \Delta W \Delta h + \Delta b$ exactly
Eq. (2)	`overtake_iff`, `prediction_preserved_iff`	a competitor overtakes iff its gap change exceeds the margin; the winner survives iff none does
Sec. 3	`fixed_head_contrast`, `gap_unchanged_of_orthogonal`	with a fixed head the gap change is $(w_j - w_y)^\top \Delta h$; an error orthogonal to the contrast leaves the margin unchanged
Eq. (4)	`spectral_decision`, `decision_bound`	the gap change splits over an orthonormal basis into sensitivity times directional error, and is bounded by their absolute products
Eq. (3)	`noise_dominated_iff`, `tail_share_le_count_share`	$\mathrm{SNR}_k < 1$ iff the error exceeds the variance; a set of lowest-variance directions holds at most its count share of the variance
Eq. (10)	`transform_invariance`, `sem_fold_costs_only_H`	the compensated layer is unchanged for any invertible T; for $T = \mathrm{diag}(s)H$ both diagonals fold into adjacent parameters and only H remains
App. D	`variance_of_linear_form`, `crossing_prob_le_chebyshev`	with pairwise-independent entries $\mathrm{Var}\langle G, \Delta W\rangle = \sum_i G_i^2 \sigma_i^2$; a zero-mean change crosses a margin m with probability at most V/m^2
Sec. 5	`forecast_error_is_cross_term`, `forecast_exact_fixed_head`	for one linear readout with a supplied feature perturbation, the residual is $(\Delta W_j - \Delta W_y)^\top e$; fixing that readout removes this term
Sec. 5	`magnitude_bound_blind`	once a bound exceeds the margin, a closing and a widening perturbation of that size are indistinguishable to it
Sec. 6	`feature_error_certificate`	if $\|w_j - w_y\|\,\|\Delta h\| \le m_{yj}$ for every competitor, the base model's winner survives: the distillation objective bounds decisions, a margin certificate of the kind used for certified perturbation invariance (Tsuzuku et al., 2018) applied to the feature error
Sec. 4	`white_error_drowns_tail`, `white_error_share_le_count_share`	an error flat along a decreasing spectrum leaves a tail of directions noise-dominated, and that tail holds at most its count share of the variance
App. D, Sec. 4	`variance_two_operands`, `snr_scale_invariant`	independent weight and activation terms add in variance; feature SNR is invariant to the readout scale
Secs. 4, 4	`retention_lower_bound`, `snr_gain_per_bit`, `weight_error_blind_to_activations`, `margin_ratio_scale_invariant`	pooled retention is at least half the surviving variance share; a bit multiplies every ratio by four; activation-only configurations share one weight error; the margin ratio is scale-free
Eq. (4)	`spectral_expansion`, `cauchy_schwarz_dot`	any inner product expands over an orthonormal basis as $\sum_k (u^\top v_k)(v_k^\top x)$; the Cauchy–Schwarz bound used for the contrast term
Eq. (10)	`sem_inverse_transpose`	the compensating transform of $\mathrm{diag}(s)H$ is $\mathrm{diag}(1/s)H$, so the scale folds into the weight rows
Sec. 6	`feature_error_controls_all_margins`	the squared gap changes against all competitors sum to at most the head's contrast energy times $\|\Delta h\|^2$: the feature objective bounds every margin at once
App. F	`forecast_is_differential`, `two_layer_residual`	the anticipated margin change is the differential of any differentiable margin map, with remainder $o(\|\Delta\theta\|)$; for two composed linear layers the remainder is exactly $\Delta_2 \Delta_1 h$
Sec. 5	`flip_iff`, `runnerup_sound`, `subset_sound`	a flip occurs iff some gap goes negative; the runner-up event implies disagreement, so its rate is a lower bound
Sec. 5	`subset_incomplete`, `subset_exact`	a top-k detector can miss a flip, and catches every flip once the excluded margins exceed the largest gap change
Sec. 6	`no_flip_of_bounded`, `preserved_iff_within_margins`	a margin above the worst gap change cannot be overturned; the prediction is preserved iff every gap change stays within its margin
Sec. 5	`flip_count_le_atrisk`, `flip_depends_only_on_gap_change`	over an evaluation set flips are at most the at-risk inputs; two compressions with equal gap changes flip identically
Sec. 5	`flip_of_overtake`, `no_overtake_of_argmax`	a strict overtake by any class is a flip, and no class overtakes a preserved winner: the exact characterisation behind the all-class statistic
Sec. 5	`runnerup_misses`, `flip_happened`, `runnerup_depends_only_on_gap_change`	a worked instance where the runner-up detector is silent while a third class overtakes, so the runner-up rate is a lower bound; that rate is a functional of the gap-change law alone
Sec. 5	`subset_complete`, `countP_le_of_imp`	with excluded margins at least M and gap changes at most $D < M$, a top-k detector misses no flip; the counting lemma behind the at-risk bound
Sec. 6	`no_flip_of_margins`, `margin_small_of_flip`, `perturbation_must_be_below_margin`	non-negative gaps preserve the prediction; a flip shows some margin was at or below the perturbation; a preserved prediction means the perturbation stayed below the threatened margin
Secs. 3, 4	`margin_scale`, `gap_change_scale`, `argmax_scale_invariant`	margins and gap changes scale by the same factor, so their ratio is scale-free and a positive rescaling leaves the argmax unchanged

or above 93%; the spectral experiment uses a separate evaluation set. All rows in that table use that experiment's measured rates, rather than rates from the compression-selection set.

For zero-shot CLIP, the image encoder is quantized while the text embeddings defining class scores remain fixed. For quantization, feature distillation adapts the encoder on ImageNet-A proxy images, with no evaluation labels or evaluation-domain training images (sparsification uses CC3M image–

Table 7: **Anticipation residuals** by margin quartile, and the offsetting count errors behind the rate; medians over vision configurations.

		residual / change, by margin quartile						
Family	**Configurations**	**Q1 (small m)**	**Q2**	**Q3**	**Q4 (large m)**	**False pos. (%)**	**False neg. (%)**	**Net rate error (%)**
Rounding	256	0.95	0.96	0.94	0.80	39	48	-6.5
Pruning	95	0.47	0.52	0.50	0.45	19	20	-1.5
Low rank	69	0.28	0.28	0.28	0.29	12	12	$+1.1$
All	420	0.85	0.87	0.85	0.72	33	40	-3.4

Table 8: **Anticipation error on the full candidate population**, before and after the task-retention screen.

Population	**Configurations**	**Median relative error (%)**	**Median absolute error (pp)**	**90th pct. absolute (pp)**
All candidates	602	6.6	0.61	3.36
pass the screen	420	5.6	0.38	2.28
fail the task-quality tolerance	178	8.2	2.08	4.20
fewer than 50 reversals	4	3.7	0.01	0.05
Quantization, all candidates	387	8.1	0.80	3.35
Pruning, all candidates	129	5.9	0.44	4.24
Low rank, all candidates	86	3.0	0.25	1.42

caption pairs). Table 9 gives the eight paired evaluations underlying the zero-shot quantization rows of Figure 7 and Table 11.

Table 9: **Zero-shot decision preservation.** Per-dataset top-1 change and disagreement of quantized CLIP image encoders after PTQ, feature distillation on proxy images, and SEM with feature distillation; the last column is relative to PTQ.

Model	**Dataset**	**Zero-shot top-1 (%)**	**Bits**	**PTQ Δtop-1**	**PTQ changed (%)**	**Distill. Δtop-1**	**Distill. changed (%)**	**SEM Δtop-1**	**SEM changed (%)**	**Fewer changes (%)**
ViT-B/16, OpenAI	ImageNet val	65.2	W6A8	-0.65	9.75	$+0.05$	5.95	-0.15	4.25	56
ViT-B/16, OpenAI	ImageNet-R	73.4	W8A8	-1.00	5.20	-0.90	3.85	-0.20	2.40	54
ViT-B/16, OpenAI	ImageNet-Sketch	44.4	W8A8	-0.50	9.45	$+0.40$	7.80	-0.10	4.75	50
ViT-B/16, OpenAI	ImageNet-V2	57.1	W6A8	-1.00	11.70	-0.35	7.45	$+0.30$	3.85	67
ViT-B/32, OpenAI	ImageNet val	61.6	W8A8	-0.70	5.65	$+0.05$	5.20	-0.30	2.95	48
ViT-B/32, OpenAI	ImageNet-R	65.4	W6A8	-0.60	16.15	$+0.05$	9.40	$+0.10$	7.60	53
ViT-B/32, OpenAI	ImageNet-Sketch	38.7	W6A8	-0.95	25.90	-1.50	16.55	-0.40	12.25	53
ViT-B/32, OpenAI	ImageNet-V2	52.1	W7A8	-0.05	10.35	-0.60	7.20	-0.20	3.60	65

H NOISE AND PERSISTENCE CONTROLS

We additionally compare prediction disagreement under matched initial feature error. Table 10 compares each compression with a Gaussian perturbation matched on feature-error energy. For quantization the ratio is close to one, indicating similar disagreement at matched error energy; for pruning and low-rank truncation the compression flips more inputs than its matched control, consistent with a contribution from error structure. Adapting ViT-B/16 at four bits under the constraint and after a matched perturbation for $4{,}000$ steps leaves output disagreement at 13.96% and 11.40% (best learning rate for each); the constrained model is higher at all six learning-rate and seed combinations tested, by a factor of 1.16 to 1.41.

I SUPPLEMENTARY TABLES

Tables 11 and 15 list the values of Figure 7. Tables 13, 14 and 15 support Section 6, every method at matched task performance. Table 12 is the language-model counterpart of Figure 2: free-running answers to trivia and arithmetic questions, before and after compression.

Table 10: **Matched-error controls.** Disagreement under each compression against a Gaussian perturbation of equal feature-error energy.

Perturbation	Models	Flips (%)	Matched control (%)	Control / compression	Flip-set overlap
Low-rank 0.99	8	9.25	6.15	0.6	0.18
Low-rank 0.9	8	25.84	25.10	1.0	0.32
Prune 30%	8	59.05	43.09	0.8	0.52
Prune 10%	8	14.04	11.15	0.8	0.25
W8	8	3.11	3.69	1.0	0.19
W6	8	14.43	15.68	1.1	0.29
W4	8	53.36	47.78	1.0	0.55

Table 11: **Preserving representation suppresses compression-induced decision changes.** Feature distillation restores decision fidelity after compression, while SEM further reduces disagreement by protecting high-energy inputs before adaptation. The effect extends across quantization, sparsification, and low rank at matched task performance.

Model	Runs	One-shot	Feature distillation	SEM + feature distillation	Δtop-1 (pp)	Reduction (%)
Quantization, W8A8; supervised on ImageNet-Sketch; runs are seeds						
ViT-B/16	3	12.61	4.24	1.86 ± 0.11	$+0.00$	85.2
DeiT-B/16	3	12.74	5.21	1.84 ± 0.18	-0.04	85.6
ViT-S/16	3	14.02	5.17	3.45 ± 0.80	$+0.04$	75.4
DeiT3-S/16	2	20.87	6.29	3.32 ± 0.29	-0.16	84.1
Quantization, zero-shot on ImageNet / V2 / R / Sketch with proxy images						
ViT-B/16, OpenAI	4	9.02 ± 2.74	6.26 ± 1.80	3.81 ± 1.01	-0.04	56.8 ± 7.4
ViT-B/32, OpenAI	4	14.51 ± 8.72	9.59 ± 4.95	6.60 ± 4.29	-0.20	54.7 ± 7.4
Sparsification, 30% of the weights (ViT-B/16: every layer but the classifier; CLIP: MLP layers)						
ViT-B/16	1	27.35	13.02	10.64	$+0.01$	61.1
CLIP ViT-B/16, zero-shot	4	24.24 ± 7.89	14.16 ± 4.99	12.10 ± 4.26	-0.79	50.3 ± 2.3
Low rank, half rank of each MLP weight						
ViT-B/16	1	57.34	21.81	20.22	-1.95	64.7

Values are means over runs, $\pm$ SD where applicable. One-shot denotes compression without adaptation.

Table 12: **Free-running question answering.** Exact match and changed-answer rate per configuration on TriviaQA and GSM8K.

Model	Compression	Task	PPL ratio	Exact match (%) base $\rightarrow$ compressed	Answers changed (%)
Qwen2.5-0.5B	W8A8	TriviaQA	1.0475	$7.7 \rightarrow 7.3$	74.0
	W8A8	GSM8K	1.0475	$1.7 \rightarrow 2.0$	47.7
	W7A8	TriviaQA	1.0405	$7.7 \rightarrow 5.3$	81.0
	W7A8	GSM8K	1.0405	$1.7 \rightarrow 2.3$	58.0
	W6A8	TriviaQA	1.0931	$7.7 \rightarrow 9.0$	79.3
	W6A8	GSM8K	1.0931	$1.7 \rightarrow 1.7$	75.7
	A8	TriviaQA	1.0300	$7.7 \rightarrow 8.0$	71.0
	A8	GSM8K	1.0300	$1.7 \rightarrow 1.7$	48.3
	Prune 10% A8	TriviaQA	1.0890	$7.7 \rightarrow 9.3$	77.7
	Prune 10% A8	GSM8K	1.0890	$1.7 \rightarrow 3.7$	67.0
Llama-2-7B	W8A8	TriviaQA	1.0814	$41.0 \rightarrow 44.5$	27.5
	W8A8	GSM8K	1.0814	$5.5 \rightarrow 5.0$	34.5
	W7A8	TriviaQA	1.0787	$41.0 \rightarrow 43.0$	32.0
	W7A8	GSM8K	1.0787	$5.5 \rightarrow 5.5$	36.0
	W6A8	TriviaQA	1.0867	$41.0 \rightarrow 42.0$	38.0
	W6A8	GSM8K	1.0867	$5.5 \rightarrow 6.5$	38.5
	A8	TriviaQA	1.0764	$41.0 \rightarrow 45.5$	29.0
	A8	GSM8K	1.0764	$5.5 \rightarrow 4.0$	37.0
	Prune 10% A8	TriviaQA	1.0911	$41.0 \rightarrow 40.0$	32.5
	Prune 10% A8	GSM8K	1.0911	$5.5 \rightarrow 6.5$	42.0

Table 13: **Feature-preservation ablation.** Disagreement, added arithmetic, and top-1 change of each component on ViT-B/16 at W8A8.

Method	Flips	Rate (%)	vs. PTQ (%)	Added arithmetic (%)	Δtop-1 (pp)
Post-training quantization	1261	12.61	0	0.00	−0.02
Feature distillation	424	4.24	66	0.00	–
+ precision allocation	308	3.08	76	4.85	+0.05
+ fitted diagonal	383	3.83	70	0.00	+0.22
+ dense flat rotation	278	2.78	78	158.40	–
+ SEM	186 ± 11	1.86	85	2.82	+0.00
+ SEM, + precision allocation	189 ± 10	1.89	85	4.89	+0.07
+ dense energy rotation	168	1.68	87	158.40	−0.01
+ dense energy rotation, + precision	164	1.64	87	163.30	–

Table 14: **Precision allocation, held out.** Widths chosen on one half of an input set and every rule scored on the other, over the budgets where every rule is defined, since uniform allocation needs an integer average width. Intervals bootstrap budgets.

Rule	Flips, mean ± sd (%)	Anticipation is lower	Reduction (%), median [95% CI]
Activation bits, at fixed W8: 17 budgets where every rule is defined			
Uniform	27.24 ± 15.91	17/17	31 [27, 44]
Rounding energy	21.91 ± 9.81	17/17	22 [20, 32]
Anticipation (ours)	16.52 ± 8.03	–	–
Weight bits, at fixed A8: 29 budgets where every rule is defined			
Uniform	14.40 ± 11.46	25/29	10 [2, 16]
Rounding energy	14.67 ± 9.78	23/29	11 [5, 18]
Hessian (HAWQ-v2)	13.17 ± 7.38	24/29	5 [1, 18]
Anticipation (ours)	11.85 ± 7.60	–	–

Table 15: **Language-model decision preservation.** Token disagreement and perplexity ratio for each transform and for complete GPTQ pipelines; QuaRot is our R1+R4 reimplementation. SEM is the transform alone, without feature distillation; the Qwen W8A8 and W4A4 SEM values are means over three calibration seeds (range in parentheses), the other rows single runs. Figure 7 shows the rows within 10% of base model perplexity.

				perplexity ratio		
Configuration	No transform or mixing (%)	QuaRot R1+R4 or learned (%)	SEM (%)	Reference	SEM	SEM lower by (%)
Untransformed / QuaRot R1+R4 / SEM						
Qwen W8A8	12.9	3.27	1.97 (1.81–2.20)	1.006	1.000	40
Qwen W6A6	45.0	14.26	7.91	–	–	45
Qwen W4A8	77.1	45.95	27.15	2.185	1.274	41
Qwen W4A4	97.9	64.21	37.91 (37.30–38.67)	5.840	1.690	41
Llama W8A8	11.2	1.90	0.88	1.000	1.001	54
Mixing only / learned per-channel scale / SEM						
Qwen W8A8	3.16	2.33	2.03	1.000	1.003	+13
Qwen W6A6	11.72	9.59	8.77	1.017	1.022	+8
Qwen W4A8	39.34	29.96	27.18	1.291	1.272	+9
Qwen W4A4	52.64	38.53	40.35	1.742	1.754	−5
Llama W8A8	1.12	0.88	1.00	1.000	1.001	−14
GPTQ weights: untransformed, QuaRot + GPTQ, and SEM + GPTQ						
Qwen W8A8	13.96	3.08	2.00	1.004	1.002	35
Qwen W4A8	40.72	17.87	14.89	1.095	1.058	17
Llama W8A8	9.70	1.89	0.72	0.999	0.999	62