WHAT WE LOSE IN PRUNING: A GENERALIZATION PERSPECTIVE ON MODEL PRUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Model pruning, which masks a fraction of model weights, primarily aims to preserve in-distribution accuracy; however, its hidden costs remain less understood. We ask: *Beyond accuracy, what do we lose when we prune a network?* We argue that pruning significantly reshapes the loss geometry, thereby damaging generalization and making the model more sensitive to data disturbances despite high clean accuracy. To theoretically investigate these effects, we analyze the consequences across two sequential stages: parameter masking and subsequent fine-tuning. First, immediately after masking, we characterize how the gradients of the remaining weights change through Hessian interactions. By grouping weights based on individual importance and interaction strength, we derive group-specific bounds on these gradient shifts. Second, after fine-tuning, we analyze sharpness via the maximum Hessian curvature with respect to the remaining weights. We decompose the change in sharpness relative to the original model into geometric shifts caused by masking and fine-tuning, showing how the final model becomes sharper. Through experiments on vision models and large language models across image classification, language modeling, commonsense reasoning, and code generation, we demonstrate that pruning incurs hidden costs in parameter sensitivity, sharpness, and robustness—costs that in-distribution performance fails to capture.

1 INTRODUCTION

The advent of large-scale models has pushed the boundaries of deep learning across various domains (Devlin et al., 2019; Brown et al., 2020). However, their immense size creates substantial bottlenecks in storage and inference. Model pruning mitigates these challenges by removing selected parameters from a trained dense network while seeking to preserve performance.

The dominant paradigm in post-training pruning is to maintain fidelity to the original dense model. Parameters are ranked by criteria such as weight magnitude (Han et al., 2015), second-order loss information (Theis et al., 2018), local reconstruction error (Frantar & Alistarh, 2023), or activation-aware statistics (Sun et al., 2024), and the resulting sparse model is often fine-tuned with its mask fixed. Despite these varying heuristics, the primary objective remains achieving high sparsity with minimal degradation in in-distribution performance.

Such loss- and accuracy-centric evaluations, however, overlook critical properties beyond zero-order performance, particularly the underlying loss geometry and solution robustness. While a pruned model may preserve its in-distribution accuracy, its altered geometry can render it far more sensitive to parameter perturbations and less robust to input shifts. This raises a fundamental question: *Beyond accuracy, what do we truly lose when we prune a network?*

To address this question from a generalization perspective, we examine how parameter removal and subsequent adaptation change the loss geometry. Prior work often relates pruning during training to improved generalization and flatter solutions, viewing parameter dropping as a form of implicit noise injection (Bartoldson et al., 2020), or employs sharpness-aware optimization to mitigate robustness drops (Bair et al., 2024). In contrast, we focus on the widely adopted post-training pruning regime, taking the fully trained dense model as a direct baseline. Through this lens, we theoretically quantify the separate geometric effects of parameter masking and subsequent fine-tuning. Disentangling these two stages is critical: while fine-tuning can partially reduce curvature relative to the immediately masked state, we show that it still leaves the final model sharper than the original dense network.

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.

Our analysis divides the post-training pruning process into two stages: immediately after masking and after fine-tuning. In the first stage, we characterize the immediate gradient shifts caused by masking, which abruptly sets selected weights to zero. Although the retained weights remain numerically unchanged, their gradients shift through interactions with the removed weights, critically altering early fine-tuning dynamics. To decouple individual importance from cross-parameter interactions, we categorize parameters using deletion saliency from Optimal Brain Damage (LeCun et al., 1989) and an off-diagonal Hessian interaction score. This taxonomy separates low-saliency parameters and further divides higher-saliency parameters into low- and high-interaction groups. Based on this categorization, we derive group-specific bounds that explicitly account for reinforcement and cancellation among the contributions of removed weights.

In the second stage, we characterize the local geometry after fine-tuning via the maximum Hessian curvature along the remaining parameter directions. We decompose the net change in maximum curvature from the dense model to the fine-tuned model into three distinct components: **i)** the geometric restriction of perturbations to the remaining parameter subspace at the dense solution, **ii)** the Hessian deformation induced by masking, and **iii)** the subsequent curvature shift from fine-tuning. Evaluated at the dense model, restricting perturbations to a subset of weights naturally upper-bounds the maximum curvature by that of the full parameter space. However, we show that the combined curvature increase driven by masking and fine-tuning can substantially exceed this subspace-induced reduction, ultimately leaving the fine-tuned model sharper than the original dense baseline.

Together, our two-stage theoretical framework and extensive empirical evaluations uncover the hidden geometric costs of pruning—revealing that high clean accuracy can paradoxically coexist with increased sharpness, heightened parameter sensitivity, and degraded robustness.

Our main contributions are summarized as follows:

- **A theory of pruning and loss geometry:** We categorize weights by individual deletion saliency and interaction strength, deriving bounds on the gradient shifts of retained weights caused by masking. For subsequent fine-tuning, we decompose the change in maximum retained-space curvature into the curvature reduction from restricting perturbations to the retained subspace and the Hessian shifts induced by masking and fine-tuning.
- **Empirical characterization of hidden pruning costs:** Experiments with ResNet-50 (He et al., 2016) and ViT-B/16 (Dosovitskiy et al., 2021) on CIFAR-100 (Krizhevsky, 2009) and Tiny-ImageNet (Deng et al., 2009) reveal the hidden costs of pruning: even with high clean accuracy restored after fine-tuning, models exhibit increased sensitivity to weight perturbations and degraded robustness to corruptions and adversarial attacks. Similar costs appear in language and code models without fine-tuning: evaluations of Qwen2 (Yang et al., 2024), Qwen2.5 (Yang et al., 2025), and Pythia (Biderman et al., 2023) on WikiText-2 (Merity et al., 2017), HellaSwag (Zellers et al., 2019), and PIQA (Bisk et al., 2020), and Qwen2.5-Coder (Hui et al., 2024) on WikiText-2 and HumanEval (Chen et al., 2021), reveal downstream performance losses despite comparatively preserved WikiText-2 perplexity.

2 Related Works

2.1 Model Pruning

Model pruning methods can be categorized by when parameters are removed relative to model training: before, during, or after training (Cheng et al., 2024).

Pruning before and during training. The Lottery Ticket Hypothesis and its strong variant ask whether effective sparse subnetworks exist at or near initialization (Frankle & Carbin, 2019; Ramanujan et al., 2020). During-training methods remove weights while optimization proceeds, allowing the network to adapt as sparsity increases (Zhu & Gupta, 2017; Bartoldson et al., 2020). These settings concern sparse trainability; our setting applies pruning to an already learned task solution.

Post-training pruning. Post-training pruning begins from a trained model and asks how parameters can be removed while preserving its performance. Early methods use weight magnitude (Han et al., 2015) or a local quadratic loss model, as in Optimal Brain Damage (OBD) (LeCun et al., 1989) and Optimal Brain Surgeon (Hassibi & Stork, 1993). Subsequent work develops saliency scores based on

movement during task adaptation (Sanh et al., 2020), Taylor approximations (Molchanov et al., 2019), and Fisher information (Theis et al., 2018; Kuznedelev et al., 2023). More recent methods scale these ideas through approximate second-order layerwise reconstruction, as in SparseGPT (Frantar & Alistarh, 2023), or combine weight and activation magnitudes, as in Wanda (Sun et al., 2024).

2.2 Generalization and Loss Landscape

The generalization of deep models is often studied through the geometry of the loss landscape. A long-standing hypothesis is that models converging to flat and wide minima generalize better than those converging to sharp and narrow ones (Hochreiter & Schmidhuber, 1997; Keskar et al., 2017). The underlying intuition is that a flat neighborhood is less sensitive to perturbations of the parameters.

This geometric perspective has also led to practical learning methods. Sharpness-Aware Minimization explicitly optimizes a finite-neighborhood objective and biases training toward flatter solutions (Foret et al., 2021). Related ideas have been applied to domain generalization (Cha et al., 2021), federated learning (Qu et al., 2022; Caldarola et al., 2022; Lee & Yoon, 2024), image generation (Lee et al., 2025b), and reinforcement learning (Lee & Yoon, 2025).

Several recent studies connect pruning directly to loss geometry. AdaSAP uses sharpness-aware perturbations to obtain robust sparse networks (Bair et al., 2024), SAFE searches for sparse and flat solutions (Lee et al., 2025a), and a complementary fundamental-limit analysis uses statistical dimension and Hessian spectra to characterize the maximum tolerable pruning ratio (Zhang et al., 2024). These studies either optimize the geometry of a sparse solution or characterize the sparsity that a model can tolerate. In contrast, we analyze how a given post-training mask changes an already learned solution, first at the moment of masking and then after fine-tuning. We use finite-radius parameter sensitivity as an operational description of local geometry and evaluate input robustness separately, without treating sensitivity as a universal or causal measure of generalization.

3 Theory for the Hidden Costs of Model Pruning

3.1 Definitions and Assumptions

Let $W_0 = (w_1, \dots, w_d) \in \mathbb{R}^d$ denote the trained dense model before pruning. Throughout our analysis, we distinguish three model *states*: the initial dense W_0, the immediately masked W_M, and the fine-tuned model W_{FT}. We denote the loss function by $\mathcal{L}$, with its gradient $g = \nabla\mathcal{L}$ and Hessian $H = \nabla^2\mathcal{L}$, writing $H_0 = H(W_0)$ for the Hessian evaluated at the dense model. Furthermore, $\|\cdot\|_2$ denotes the Euclidean norm for vectors or the spectral norm for matrices, $[x]$ is a set of positive integers less than or equal to x, $[x]_+ := \max\{x, 0\}$, and $e_j \in \mathbb{R}^d$ is the j-th standard basis vector.

Definition 3.1 (**States and retained-subspace curvature**). For a mask $M \in \{0,1\}^d$, $M_j = 1$ means that coordinate j is retained and $M_j = 0$ means that it is pruned. Let $P_M = \mathrm{diag}(M)$ and

$$W_M := P_M W_0, \quad q_M := (I - P_M)W_0 = W_0 - W_M, \quad W_{\mathrm{FT}} := W_M + \Delta W_{\mathrm{FT}}.$$

Let $\mathcal{Z}_M = \{j : M_j = 0\}$. Fixed-mask fine-tuning obeys $P_M \Delta W_{\mathrm{FT}} = \Delta W_{\mathrm{FT}}$. For a nonempty retained subspace, let us define

$$\Lambda_M(W) := \max_{\substack{\|v\|_2=1 \\ P_M v = v}} v^\top H(W) v, \qquad \Lambda_0 = \lambda_{\max}(H_0).$$

$\Lambda_M(W)$ is the largest Hessian curvature along unit perturbations that change only retained parameters. We call it the *retained-subspace curvature*.

Definition 3.2 (**Saliency-and-interaction taxonomy**). For coordinate j, define its individual deletion saliency $a_j \in \mathbb{R}$ and global interaction score $s_j \in \mathbb{R}$ by

$$a_j := \tfrac{1}{2} w_j^2 |(H_0)_{jj}|, \qquad s_j := |w_j| \sqrt{\sum_{i \neq j} (H_0)_{ij}^2}.$$

For thresholds $\tau_1, \tau_2 > 0$, partition the weight index set $[d]$ into

$$\mathcal{I}_{\mathrm{low}} = \{j : a_j \le \tau_1\}, \quad \mathcal{C}_{\mathrm{low}} = \{j : a_j > \tau_1, s_j < \tau_2\}, \quad \mathcal{C}_{\mathrm{high}} = \{j : a_j > \tau_1, s_j \ge \tau_2\}.$$

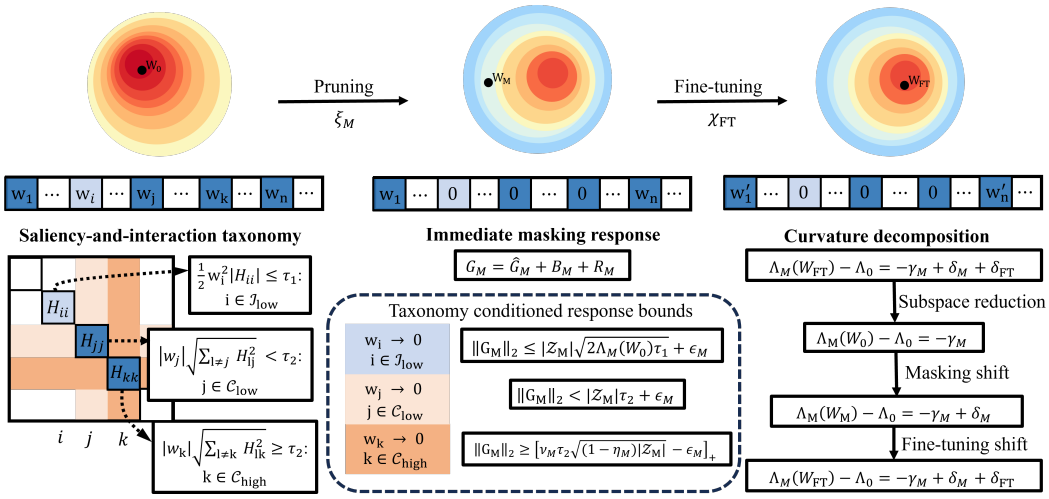


Figure 1: Overview of our stagewise theoretical analysis of post-training pruning.

$\mathcal{I}_{\text{low}}$ contains coordinates with low individual deletion saliency. Among the remaining coordinates, $\mathcal{C}_{\text{low}}$ and $\mathcal{C}_{\text{high}}$ contain those with low and high interaction scores, respectively. The score s_j aggregates the first-order gradient disturbance that deleting w_j can transmit to all other coordinates; it does not threshold a single Hessian entry; "low" and "high" refer to the magnitude of this score.

When $H_0 \succeq 0$, a_j recovers the classical OBD deletion saliency (LeCun et al., 1989). The score a_j estimates a weight's individual deletion cost, whereas s_j measures how strongly its removal can affect other gradients through Hessian interactions.

Assumption 3.3 (Local paths and bounded fine-tuning). To compare the changes induced by masking and fine-tuning, we consider the straight-line masking path $\xi_M(t) = W_0 - tq_M$ and the straight-line fine-tuning path $\chi_{\text{FT}}(t) = W_M + t\Delta W_{\text{FT}}$ for $t \in [0, 1]$. For either path $\zeta \in \{\xi_M, \chi_{\text{FT}}\}$, we assume that $g(\zeta(t))$ and $H(\zeta(t))$ are discontinuous at only finitely many points; the left and right limits exist at each such point; and both are absolutely continuous between consecutive discontinuities. On each such interval, the chain rule $\frac{d}{dt} g(\zeta(t)) = H(\zeta(t))\zeta'(t)$ holds almost everywhere, and $\left\| \frac{d}{dt} H(\zeta(t)) \right\|_2 \le L_H \|\zeta'(t)\|_2$. Fine-tuning keeps the mask fixed, with $\|\Delta W_{\text{FT}}\|_2 \le r_{\text{FT}}(M)$.

The masking and fine-tuning paths are used to compare the three model states and separate stage-specific changes, rather than to model the actual optimization trajectory. The bound $r_{\text{FT}}(M)$ controls the distance between the masked and fine-tuned models. These conditions accommodate discontinuities while separating smooth Hessian variation from discontinuous changes along the paths.

Figure 1 summarizes our stagewise analysis: the dense Hessian defines the saliency–interaction taxonomy, masking yields the retained-gradient response and taxonomy-conditioned bounds, and fixed-mask fine-tuning yields the curvature decomposition at the fine-tuned endpoint.

3.2 Immediate Parameter-Space Response to Pruning

We first study the effect of masking before fine-tuning. Although retained weights are unchanged, Hessian interactions with removed weights can alter their gradients. Since $W_M = W_0 - q_M$, define

$$G_M := P_M[g(W_M) - g(W_0)], \qquad \widehat{G}_M := -P_M H_0 q_M,$$

where $\widehat{G}_M$ predicts the masking response using the dense-model Hessian H_0. However, $\widehat{G}_M$ need not capture the full response, since masking can induce additional gradient changes at nonsmooth boundaries and alter the loss curvature as the parameters move from W_0 to W_M. We denote these effects by B_M and R_M, respectively.

We first make the boundary correction explicit. Along the continuous masking path ξ_M, the gradient $g(\xi_M(t))$ may have finitely many discontinuities. Let $\mathcal{T}_{\xi_M} \subset (0, 1)$ denote the points where its one-sided limits differ, with $g(\xi_M(t^{\pm})) := \lim_{s \to t^{\pm}} g(\xi_M(s))$. Their contribution to the retained-gradient

change is

$$B_M := \sum_{t \in \mathcal{T}_{\xi_M}} P_M \big[g(\xi_M(t^+)) - g(\xi_M(t^-)) \big].$$

For the second effect, the Hessian may change as the parameters move along the masking path. We define the resulting correction as

$$R_M := \int_0^1 P_M [H(\xi_M(t)) - H_0] \xi_M'(t)\, dt.$$

Theorem 3.4 (Decomposition of the retained-gradient response). *Under Assumption 3.3,*

$$G_M = \widehat{G}_M + B_M + R_M.$$

Moreover,

$$\|B_M\|_2 \le A_g(\xi_M) := \sum_{t \in \mathcal{T}_{\xi_M}} \big\| P_M [g(\xi_M(t^+)) - g(\xi_M(t^-))] \big\|_2,$$

$$\|R_M\|_2 \le \rho_M := \|q_M\|_2 \int_0^1 \|P_M [H(\xi_M(t)) - H_0](I - P_M)\|_2 \, dt.$$

Hence,

$$\|G_M - \widehat{G}_M\|_2 \le \varepsilon_M := A_g(\xi_M) + \rho_M.$$

Here, $A_g(\xi_M)$ measures the accumulated contribution of gradient discontinuities, while ρ_M controls the effect of Hessian variation along the masking path. When both are small, the actual response G_M remains close to the dense-Hessian prediction $\widehat{G}_M$.

Proposition 3.5 (Response bounds conditioned on OBD saliency and interaction). *For a mask M with $1 \le |\mathcal{Z}_M| < d$:*

(i) Low-individual-saliency: *If $\mathcal{Z}_M \subseteq \mathcal{I}_{\text{low}}$ and $H_0 \succeq 0$, then $\|G_M\|_2 \le |\mathcal{Z}_M| \sqrt{2\Lambda_M(W_0)\tau_1} + \varepsilon_M$.*

(ii) Low-interaction: *If $\mathcal{Z}_M \subseteq \mathcal{C}_{\text{low}}$, then $\|G_M\|_2 < |\mathcal{Z}_M|\tau_2 + \varepsilon_M$.*

(iii) High-interaction: *If $\mathcal{Z}_M \subseteq \mathcal{C}_{\text{high}}$, let $z_j := -w_j P_M H_0 e_j$ and $\nu_M := \min_{j \in \mathcal{Z}_M} \|z_j\|_2 / s_j$. To avoid excessive cancellation, suppose some $\eta_M \in [0, 1)$ satisfies*

$$\sum_{\substack{j,k \in \mathcal{Z}_M \\ j \ne k}} \langle z_j, z_k \rangle \ge -\eta_M \sum_{j \in \mathcal{Z}_M} \|z_j\|_2^2 \quad \Longrightarrow \quad \|G_M\|_2 \ge \Big[\nu_M \tau_2 \sqrt{(1 - \eta_M)|\mathcal{Z}_M|} - \varepsilon_M \Big]_+.$$

Cases (i)–(ii) show that low-saliency and low-interaction masks admit controlled immediate responses, up to the correction ε_M. In contrast, Case (iii) gives a conditional lower bound when high-interaction contributions survive in the retained subspace and do not cancel excessively; the bound is informative when the interaction term dominates ε_M. Mixed-mask cases are discussed in Appendix A.5.

3.3 Retained-Subspace Curvature after Fine-Tuning

We next track how retained-subspace curvature changes through masking and fixed-mask fine-tuning. For either path $\zeta \in \{\xi_M, \chi_{\text{FT}}\}$, let $\mathcal{T}_\zeta \subset (0, 1)$ denote the points where the one-sided Hessian limits differ, and define the accumulated Hessian-discontinuity magnitude

$$A_H(\zeta) := \sum_{t \in \mathcal{T}_\zeta} \big\| H(\zeta(t^+)) - H(\zeta(t^-)) \big\|_2.$$

Theorem 3.6 (Decomposition of the post-fine-tuning curvature change). *Under Assumption 3.3,*

$$\Delta\Lambda_M := \Lambda_M(W_{\text{FT}}) - \Lambda_0 = -\gamma_M + \delta_M + \delta_{\text{FT}},$$

where

$$\gamma_M := \Lambda_0 - \Lambda_M(W_0) \ge 0, \quad \delta_M := \Lambda_M(W_M) - \Lambda_M(W_0), \quad \delta_{\text{FT}} := \Lambda_M(W_{\text{FT}}) - \Lambda_M(W_M).$$

Moreover,

$$|\delta_M| \le L_H \|q_M\|_2 + A_H(\xi_M), \qquad |\delta_{\text{FT}}| \le L_H r_{\text{FT}}(M) + A_H(\chi_{\text{FT}}).$$

Hence,

$$\Delta\Lambda_M > 0 \quad \Longleftrightarrow \quad \delta_M + \delta_{\text{FT}} > \gamma_M.$$

In parallel with Theorem 3.4, $\Delta\Lambda_M$ decomposes into the retained-subspace reduction $-\gamma_M$ and the stage-specific curvature shifts from masking and fine-tuning, $\delta_M + \delta_{\mathrm{FT}}$. The final retained-space curvature exceeds the dense curvature exactly when their combined curvature shift exceeds γ_M. Section 4.4 measures all three contributions separately.

4 Experiments

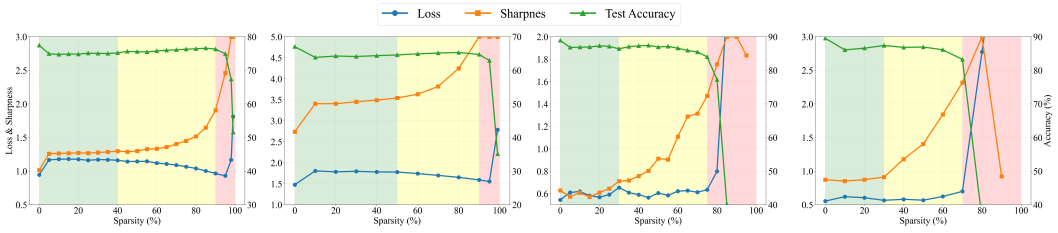
We empirically examine the analysis in Section 3 and pruning costs beyond clean accuracy. Section 4.2 evaluates clean accuracy, parameter sensitivity, and input robustness after fine-tuning. Section 4.3 compares taxonomy-conditioned masking responses and the dense-Hessian prediction, Section 4.4 quantifies the curvature decomposition, and Section 4.5 examines language-modeling and downstream performance in language and code models without fine-tuning.

4.1 Experiment Setup

In computer vision tasks, we evaluate the most widely used architectures, including ResNet-50 (He et al., 2016) and ViT-B/16 (Dosovitskiy et al., 2021), on CIFAR-100 (Krizhevsky, 2009) and Tiny-ImageNet (Deng et al., 2009). For the practical pruning algorithms, we select representative methodologies, such as unstructured magnitude (Han et al., 2015) and Fisher pruning (Theis et al., 2018), followed by five-epoch fixed-mask fine-tuning. All results are based on five statistical seeds unless noted. In language and code-generation tasks, we apply Wanda pruning (Sun et al., 2024) at 20–40% sparsity without fine-tuning to the popular large language models, such as Qwen2-0.5B, Qwen2.5-1.5B, Pythia-1.4B, and Qwen2.5-Coder at 0.5B–3B (Yang et al., 2024; 2025; Biderman et al., 2023; Hui et al., 2024). Details are in Appendix B.

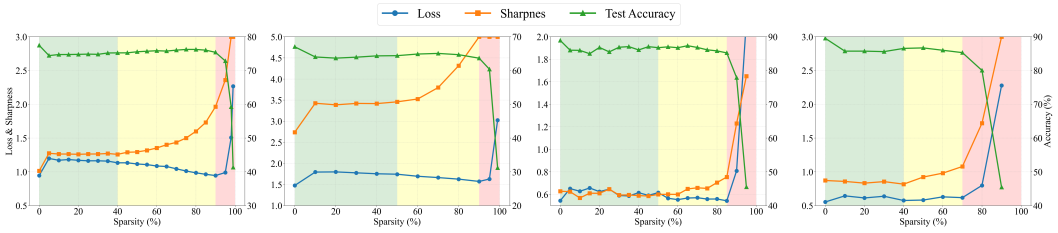
4.2 Hidden Costs Beyond Clean Accuracy

We measure retained-space *sharpness*: $S_\rho = \max_{\|\epsilon\|_2 \le \rho}\{\mathcal{L}(W+\epsilon) - \mathcal{L}(W)\}$ with $\rho = 0.05$, following Foret et al. (2021), and report it as parameter sensitivity.



(a) ResNet-50 on CIFAR-100 (b) ResNet-50 on Tiny-ImageNet (c) ViT-B/16 on CIFAR-100 (d) ViT-B/16 on Tiny-ImageNet

Figure 2: Sparsity (i.e., a portion of pruned parameters) vs. test loss, test accuracy, and sharpness S_ρ for ResNet-50/ViT trained on the CIFAR-100/Tiny-ImageNet datasets under Magnitude pruning



(a) ResNet-50 on CIFAR-100 (b) ResNet-50 on Tiny-ImageNet (c) ViT-B/16 on CIFAR-100 (d) ViT-B/16 on Tiny-ImageNet

Figure 3: Sparsity (i.e., a portion of pruned parameters) vs. test loss, test accuracy, and sharpness S_ρ for ResNet-50/ViT trained on the CIFAR-100/Tiny-ImageNet datasets under Fisher pruning

Figures 2 and 3 show three qualitative regimes: a green-colored region with stable performance and sensitivity, a yellow-colored region where sensitivity rises while clean accuracy remains comparatively

preserved, and a red-colored region with severe performance degradation. We call the yellow regime the *tipping phase*, using the term descriptively rather than as a formal phase boundary. As shown consistently across architectures and benchmarks, test accuracy remains stable through the end of the tipping phase, but sharpness gradually increases during the phase; loss geometry begins to degrade.

Table 1: Generalization and robustness in the tipping phase. Accuracy entries report relative changes from the dense model. Sparsity 90% means that we mask 90% of parameters.

Dataset	Model / Pruning	Sparsity	Clean Acc.	Corruption Acc.	PGD Acc.	ΔS_ρ
CIFAR-100	ResNet-50 / Fisher	90%	−2.6%	−5.5%	−26%	+0.95
CIFAR-100	ResNet-50 / Magnitude	90%	−1.5%	−3.7%	−20%	+0.89
CIFAR-100	ViT-B/16 / Fisher	80%	−6.2%	−7.9%	−46%	+0.07
CIFAR-100	ViT-B/16 / Magnitude	70%	−3.9%	−7.2%	−58%	+0.68
Tiny-ImageNet	ResNet-50 / Fisher	80%	−3.5%	−6.7%	−18%	+1.58
Tiny-ImageNet	ResNet-50 / Magnitude	80%	−2.6%	−4.1%	−15%	+1.51
Tiny-ImageNet	ViT-B/16 / Fisher	70%	−4.7%	−7.3%	−44%	+0.20
Tiny-ImageNet	ViT-B/16 / Magnitude	70%	−7.0%	−12%	−82%	+1.45

Table 1 reports how this damage to loss geometry influences the model robustness. We evaluate accuracies against data corruption (Hendrycks & Dietterich, 2019) and PGD-10 attack at $\epsilon = 1/255$ (Madry et al., 2018). We partially report cases in the tipping phase to contrast severe deterioration of robustness even with relatively stable clean accuracy. Robustness degradation exceeds the corresponding clean-accuracy drop, showing that pruning and fine-tuning can preserve comparatively high clean accuracy while increasing parameter sensitivity and reducing robustness. Full results across sparsity and their confidence intervals are in Appendix C.

4.3 Immediate Effects of Pruning Different Parameter Groups

Here, we investigate the theory-simulation alignment. First, we ask whether the dense-Hessian leading term $\widehat{G}_M := -P_M H_0 q_M$ in Theorem 3.4 aligns with the immediate retained-gradient change G_M. If so, it means that loss geometry is consistent and monotone across masking-induced changes from W_0 to W_M; otherwise, the masking significantly reshapes the curvature of the loss geometry. To evaluate vector-level agreement between G_M and $\widehat{G}_M$, we compute the following measurements:

$$c_{\text{pred}}(M) := \cos(G_M, \widehat{G}_M), \qquad r_{\text{pred}}(M) := \|G_M - \widehat{G}_M\|_2 / \|G_M\|_2.$$

Here c_{pred} measures directional alignment, while r_{pred} is the normalized prediction residual. With substantial geometry changes, c_{pred} decreases and r_{pred} increases.

The key part of the theory is the parameter taxonomy: $\mathcal{I}_{\text{low}}$, $\mathcal{C}_{\text{low}}$, $\mathcal{C}_{\text{high}}$. As formalized by Proposition 3.5, we argue that loss geometry changes per group would be different; pruning $\mathcal{I}_{\text{low}}$ minimally tunes geometry, $\mathcal{C}_{\text{low}}$ slightly reshapes it, and $\mathcal{C}_{\text{high}}$ causes considerable damage. To scrutinize per-group differences, we select a proper set of thresholds, i.e., τ_1, τ_2, thus splitting parameters into 40:40:20 ratios of each taxonomy in order. We then apply masking of 5% sparsity to keep the local proximity sufficient to evaluate the dense-Hessian leading term against higher-order drift. For each layer, the three masks remove the same number of weights. We select the removed weights so that the sum of their squared magnitudes within each layer is as similar as possible across the three groups.

We report $\|q_M\|_2$ to show any remaining differences in the total parameter displacement caused by masking. For each mask, we compute $\widehat{G}_M$ using exact Hessian–vector products of the loss at W_0. Further details are in Appendix B.3.

In Table 2, for ResNet-50, pruning $\mathcal{C}_{\text{high}}$ rather than $\mathcal{I}_{\text{low}}$ or $\mathcal{C}_{\text{low}}$ makes $\widehat{G}_M$ deviate more from G_M: c_{pred} decreases from $0.818/0.791$ to 0.737 and r_{pred} increases from $0.651/0.653$ to 0.803, indicating more severe geometric changes. For ViT-B/16, pruning the interaction sets $\mathcal{C}_{\text{low}}$ and $\mathcal{C}_{\text{high}}$ shows a more disruptive impact: c_{pred} decreases from 0.917 to $0.430/0.020$ and r_{pred} increases from 0.446 to $1.023/0.998$, implying that masking interaction groups significantly manipulates geometry. Overall, the empirical results align with the theory-based per-taxonomy diagnostics of disruptive masking behavior, showing that misalignment between G_M and $\widehat{G}_M$ leads to larger residual terms that capture the fine-grained curvature changes via path-based integration.

Table 2: Immediate masking responses at 5% sparsity across the three groups of the 40/40/20 operational taxonomy on CIFAR-100.

Architecture	Group	$\|q_M\|_2$	$\|\widehat{G}_M\|_2$	$\|G_M\|_2$	c_{pred}	r_{pred}	Clean acc. (%)
ResNet-50	$\mathcal{I}_{\text{low}}$	4.110 ± 0.578	0.051 ± 0.018	0.097 ± 0.041	0.818 ± 0.040	0.651 ± 0.019	77.33 ± 0.43
	$\mathcal{C}_{\text{low}}$	4.302 ± 0.763	0.058 ± 0.033	0.069 ± 0.020	0.791 ± 0.104	0.653 ± 0.093	77.47 ± 0.46
	$\mathcal{C}_{\text{high}}$	4.558 ± 1.043	0.164 ± 0.100	0.259 ± 0.216	0.737 ± 0.139	0.803 ± 0.232	76.73 ± 1.15
ViT-B/16	$\mathcal{I}_{\text{low}}$	33.063 ± 1.100	0.017 ± 0.011	0.018 ± 0.008	0.917 ± 0.050	0.446 ± 0.177	89.39 ± 0.11
	$\mathcal{C}_{\text{low}}$	33.431 ± 1.353	0.135 ± 0.167	0.171 ± 0.104	0.430 ± 0.102	1.023 ± 0.183	88.86 ± 0.10
	$\mathcal{C}_{\text{high}}$	33.899 ± 1.735	0.194 ± 0.122	15.618 ± 7.073	0.020 ± 0.095	0.998 ± 0.005	80.99 ± 2.69

4.4 Curvature at the Fine-Tuned Endpoint

Here, we investigate whether the curvature increase induced by masking persists after fixed-mask fine-tuning. Following Theorem 3.6, we decompose

$$\Delta\Lambda_M = -\gamma_M + \delta_M + \delta_{\text{FT}},$$

where $-\gamma_M$ is the retained-subspace reduction, while δ_M and δ_{FT} are the curvature shifts induced by masking and fine-tuning, respectively. A positive δ_M indicates that masking increases retained-space curvature, whereas a negative δ_{FT} indicates that fine-tuning counteracts this increase. We report 80% sparsity for ResNet-50 and 70% for ViT-B/16 because ViT magnitude pruning shows severe accuracy degradation at 80% sparsity.

Table 3: Retained-subspace curvature decomposition on CIFAR-100 at 80% sparsity for ResNet-50 and 70% for ViT-B/16. All contributions are normalized by the dense-model curvature Λ_0. Values are mean $\pm$ SD over five seeds. Dense accuracies are 77.47% for ResNet-50 and 88.91% for ViT-B/16.

Architecture	Method	$-\gamma_M/\Lambda_0$	δ_M/Λ_0	$\delta_{\text{FT}}/\Lambda_0$	$\Delta\Lambda_M/\Lambda_0$	Clean acc. (%)
ResNet-50	Fisher	-0.31 ± 0.05	127.41 ± 93.35	-124.78 ± 91.22	2.32 ± 2.76	76.25 ± 0.57
	Magnitude	-0.44 ± 0.04	17.90 ± 13.66	-16.78 ± 12.84	0.68 ± 0.89	76.41 ± 0.53
ViT-B/16	Fisher	-0.20 ± 0.02	29.93 ± 30.99	-27.99 ± 31.06	1.74 ± 2.53	85.83 ± 1.40
	Magnitude	-0.83 ± 0.02	3.60 ± 2.63	-1.88 ± 1.51	0.89 ± 1.26	85.47 ± 0.28

For ResNet-50, masking produces a strong positive curvature shift, especially under Fisher pruning: $\delta_M/\Lambda_0 = 127.41$, followed by an opposing fine-tuning shift of $\delta_{\text{FT}}/\Lambda_0 = -124.78$. Magnitude pruning exhibits the same pattern at a smaller scale, with corresponding shifts of 17.90 and -16.78. Thus, fine-tuning reverses most of the masking-induced curvature increase, but the mean final curvature change remains positive for both methods.

ViT-B/16 exhibits the same masking–fine-tuning opposition. Fisher pruning produces normalized shifts of 29.93 and -27.99 from masking and fine-tuning, while magnitude pruning yields 3.60 and -1.88. Again, fine-tuning offsets much of the masking-induced increase without fully restoring the dense-model curvature on average.

Overall, masking substantially increases retained-subspace curvature, while fine-tuning reverses much of this change on average. Nevertheless, the combined masking and fine-tuning shifts exceed the retained-subspace reduction γ_M, leaving the final curvature above the dense baseline on average. Importantly, clean accuracy remains comparatively high after fine-tuning despite this residual curvature increase. Thus, even after fine-tuning, masking-induced geometric effects persist, revealing a hidden curvature cost of pruning. Additional results appear in Appendix E.3.

4.5 Extension to Large Language Models

We next examine whether the hidden costs observed in vision extend to large language and code models. Unlike the vision setting, we apply pruning without fine-tuning, so the masked model is directly used for task evaluation. This allows us to isolate task-level degradation induced by masking.

We distinguish base language-modeling quality from downstream capability. For language models, we consider Qwen2-0.5B, Qwen2.5-1.5B, and Pythia-1.4B. We use WikiText-2 perplexity as a proxy

for language-modeling quality and zero-shot HellaSwag and PIQA as downstream evaluations (Merity et al., 2017; Zellers et al., 2019; Bisk et al., 2020). For code models, we evaluate Qwen2.5-Coder at 0.5B, 1.5B, and 3B on HumanEval for downstream code generation (Chen et al., 2021).

Table 4: Task performance after Wanda pruning without fine-tuning. Each cell reports 20%, 30%, and 40% sparsity from left to right. WikiText-2 PPL ratios measure changes in base language-modeling quality relative to the dense model, while ΔHS, ΔPIQA, and ΔHumanEval denote downstream changes in percentage points.

(a) Language models

Model	PPL ratio	ΔHS (pp)	ΔPIQA (pp)
Qwen2-0.5B	1.02 / 1.07 / 1.19	−0.56 / −2.00 / −4.96	−0.27 / +0.11 / −0.87
Qwen2.5-1.5B	1.02 / 1.06 / 1.18	−0.53 / −2.26 / −5.94	+0.27 / −0.54 / −2.77
Pythia-1.4B	1.02 / 1.08 / 1.21	−0.52 / −1.08 / −3.31	−0.38 / −0.82 / −1.90

(b) Code models

Model	PPL ratio	ΔHumanEval (pp)
Coder-0.5B	1.03 / 1.11 / 1.35	−4.27 / −9.76 / −12.80
Coder-1.5B	1.02 / 1.07 / 1.23	−1.83 / −3.66 / −10.98
Coder-3B	1.01 / 1.06 / 1.20	−3.66 / −1.22 / −5.49

At low and moderate sparsity, base language-modeling quality remains comparatively preserved: WikiText-2 PPL ratios are only 1.01–1.11 at 20–30% sparsity across all models. Nevertheless, substantial downstream degradation already emerges. HumanEval drops by up to 9.76 percentage points, while HellaSwag consistently degrades across all three language models. Thus, relatively stable language-modeling performance can conceal considerable losses in downstream capability.

The hidden cost becomes pronounced at 40% sparsity. PPL ratios reach 1.18–1.21 for language models and 1.20–1.35 for code models, whereas HellaSwag and HumanEval drop by 3.31–5.94 and 5.49–12.80 points, respectively. PIQA is less sensitive and occasionally improves at low sparsity, showing task-dependent hidden costs. Overall, these results extend the phenomenon in Section 4.2: preserving the primary metric does not guarantee preservation of broader model behavior; pruning can retain base language-modeling quality while substantially degrading downstream capability.

5 Limitations and Potential

The thresholds τ_1 and τ_2 are selected as hyperparameters rather than by a principled rule. Appendix G.2 tests whether the response differences between low- and high-interaction groups persist across quantile thresholds; principled threshold selection remains future work.

Another limitation is that our theoretical bounds are local and conditional rather than universal monotonicity guarantees. In particular, the high-interaction lower bound is informative only when its interaction-based term dominates residual contributions from Hessian variation and discontinuous gradient changes at discontinuity points. Thus, our theory characterizes immediate gradient responses and curvature changes after fine-tuning relative to the dense model, rather than claiming that every pruning operation necessarily increases sharpness.

Our analysis provides a basis for studying the role of individual parameters beyond their immediate contribution to task performance, particularly from a generalization perspective in increasingly large models. As compressed models become more prevalent in resource-constrained and embedded-AI settings, characterizing what is lost through pruning may support more reliable model compression. Our findings also highlight that preserving clean accuracy alone may conceal losses in robustness under distributional or adversarial shifts.

6 Conclusion

We show that post-training pruning can preserve clean accuracy while increasing parameter sensitivity and reducing robustness. Our stagewise analysis characterizes the immediate masking response through Hessian interactions and decomposes the curvature change after fine-tuning, showing how masking and fine-tuning can leave the final model sharper than the dense baseline. Experiments on language and code models without fine-tuning further show that comparatively preserved perplexity can coexist with degraded downstream performance. Together, these results motivate evaluating pruning beyond accuracy by considering parameter sensitivity, loss curvature, and input robustness.

AI USE STATEMENT

In this work, we used generative AI tools for assisting in the verification and refinement of mathematical proofs. We have not used generative AI tools for developing theoretical models, formulating core mathematical claims, designing experimental methodologies, implementing methods, or interpreting results, and generating synthetic data sets, translation, or qualitative data analysis are not applicable to this work. Additionally, we used generative AI tools to edit the manuscript to improve readability and assist with editing and consolidating existing experimental code. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work analyzes model pruning using publicly available datasets and models. It does not involve human-subject data collection, personally identifiable information, or the release of new datasets, and does not introduce additional application-specific ethical risks.

REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce our theoretical and empirical results throughout the main paper and appendix. The assumptions, definitions, and theoretical results are presented in Section 3, with complete proofs and further theoretical details provided in Appendix A. The overall experimental setup is described in Section 4.1, with training, pruning, and fine-tuning details provided in Appendix B. Further implementation details are provided in Appendix B.2 for parameter sensitivity and robustness, Appendix B.3 for immediate gradient-response diagnostics, Appendix B.4 for retained-subspace curvature estimation, and Appendix B.5 for the language model experiments. We also provide the code used for the experimental analyses to facilitate reproducibility.

REFERENCES

Anna Bair, Hongxu Yin, Maying Shen, Pavlo Molchanov, and Jose M. Alvarez. Adaptive sharpness-aware pruning for robust sparse networks. In *International Conference on Learning Representations*, 2024.

Brian R. Bartoldson, Ari S. Morcos, Adrian Barbu, and Gordon Erlebacher. The Generalization-Stability Tradeoff In Neural Network Pruning. *Advances in Neural Information Processing Systems*, 33:20852–20864, 2020.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Debora Caldarola, Barbara Caputo, and Marco Ciccone. Improving Generalization in Federated Learning by Seeking Flat Minima. In *European Conference on Computer Vision*, pp. 654–672. Springer, 2022.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. SWAD: Domain Generalization by Seeking Flat Minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

Hongrong Cheng, Miao Zhang, and Javen Qinfeng Shi. A Survey on Deep Neural Network Pruning: Taxonomy, Comparison, Analysis, and Recommendations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:10558–10578, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 248–255, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware Minimization for Efficiently Improving Generalization. In *International Conference on Learning Representations*, 2021.

Jonathan Frankle and Michael Carbin. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. In *International Conference on Learning Representations*, 2019.

Elias Frantar and Dan Alistarh. SparseGPT: Massive Language Models Can be Accurately Pruned in One-Shot. In *International conference on machine learning*, pp. 10323–10337. PMLR, 2023.

Song Han, Jeff Pool, John Tran, and William Dally. Learning both Weights and Connections for Efficient Neural Network. *Advances in neural information processing systems*, 28, 2015.

Babak Hassibi and David G Stork. Second order derivatives for network pruning: Optimal Brain Surgeon. *Advances in neural information processing systems*, 5, 1993.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations*, 2019.

Sepp Hochreiter and Jürgen Schmidhuber. Flat Minima. *Neural computation*, 9(1):1–42, 1997.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, Kai Dang, Yang Fan, Yichang Zhang, An Yang, Rui Men, Fei Huang, Bo Zheng, Yibo Miao, Shanghaoran Quan, Yunlong Feng, Xingzhang Ren, Xuancheng Ren, Jingren Zhou, and Junyang Lin. Qwen2.5-Coder technical report, 2024. URL https://arxiv.org/abs/2409.12186.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. In *International Conference on Learning Representations*, 2017.

Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto, 2009.

Denis Kuznedelev, Eldar Kurtić, Elias Frantar, and Dan Alistarh. CAP: Correlation-Aware Pruning for Highly-Accurate Sparse Vision Models. *Advances in Neural Information Processing Systems*, 36:28805–28831, 2023.

Yann LeCun, John S Denker, and Sara A Solla. Optimal Brain Damage. *Advances in neural information processing systems*, 2, 1989.

Dongyeop Lee, Kwanhee Lee, Jinseok Chung, and Namhoon Lee. SAFE: Finding sparse and flat minima to improve pruning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 33300–33321. PMLR, 13–19 Jul 2025a.

Hyun Kyu Lee and Sung Whan Yoon. Flat Reward in Policy Parameter Space Implies Robust Reinforcement Learning. In *International Conference on Learning Representations*, 2025.

Taehwan Lee and Sung Whan Yoon. Rethinking the Flat Minima Searching in Federated Learning. In *International Conference on Machine Learning*, pp. 27037–27071. PMLR, 2024.

Taehwan Lee, Kyeongkook Seo, Jaejun Yoo, and Sung Whan Yoon. Understanding Flatness in Generative Models: Its Role and Benefits. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4908–4917, 2025b.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*, 2018.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*, 2017.

Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance Estimation for Neural Network Pruning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11264–11272, 2019.

Zhe Qu, Xingyu Li, Rui Duan, Yao Liu, Bo Tang, and Zhuo Lu. Generalized Federated Learning via Sharpness Aware Minimization. In *International conference on machine learning*, pp. 18250–18280. PMLR, 2022.

Vivek Ramanujan, Mitchell Wortsman, Aniruddha Kembhavi, Ali Farhadi, and Mohammad Rastegari. What's hidden in a randomly weighted neural network? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11893–11902, 2020.

Victor Sanh, Thomas Wolf, and Alexander Rush. Movement pruning: Adaptive sparsity by fine-tuning. *Advances in neural information processing systems*, 33:20378–20389, 2020.

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

Lucas Theis, Iryna Korshunova, Alykhan Tejani, and Ferenc Huszár. Faster gaze prediction with dense networks and Fisher pruning. *arXiv preprint arXiv:1801.05787*, 2018.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024.

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

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4791–4800, 2019.

Qiaozhe Zhang, Ruijie Zhang, Jun Sun, and Yingzhuang Liu. How sparse can we prune a deep network: A fundamental limit perspective. In *Advances in Neural Information Processing Systems*, volume 37, pp. 91337–91372, 2024. doi: 10.52202/079017-2898.

Michael Zhu and Suyog Gupta. To prune, or not to prune: exploring the efficacy of pruning for model compression. *arXiv preprint arXiv:1710.01878*, 2017.

A Proofs and Further Details

A.1 Piecewise gradient transport

Lemma A.1 (Piecewise path identity). *Let $\zeta \in \{\xi_M, \chi_{\mathrm{FT}}\}$ and let $0 < t_1 < \cdots < t_K < 1$ denote the points where $g(\zeta(t))$ is discontinuous. Under Assumption 3.3,*

$$g(\zeta(1)) - g(\zeta(0)) = \int_{[0,1]\setminus\{t_1,\ldots,t_K\}} H(\zeta(t))\zeta'(t)\,dt + \sum_{k=1}^{K}[g(\zeta(t_k^+)) - g(\zeta(t_k^-))].$$

Proof. Set $t_0 = 0$ and $t_{K+1} = 1$. Absolute continuity between consecutive discontinuities and the chain rule in Assumption 3.3 give

$$g(\zeta(t_{k+1}^-)) - g(\zeta(t_k^+)) = \int_{t_k}^{t_{k+1}} H(\zeta(t))\zeta'(t)\,dt.$$

Summing over these intervals and inserting the one-sided boundary differences proves the result. □

A.2 Proof of Theorem 3.4

Proof. Apply Lemma A.1 to $\xi_M(t) = W_0 - tq_M$, for which $\xi_M'(t) = -q_M$, and project onto the retained coordinates. This gives

$$G_M = -\int_0^1 P_M H(\xi_M(t))q_M\,dt + B_M.$$

Separating the dense-Hessian term inside the integral gives

$$G_M = \widehat{G}_M + B_M + R_M.$$

The triangle inequality gives $\|B_M\|_2 \le A_g(\xi_M)$. Since $q_M = (I - P_M)q_M$, we also have $\|R_M\|_2 \le \rho_M$. Therefore,

$$\begin{aligned}\|G_M - \widehat{G}_M\|_2 &= \|B_M + R_M\|_2 \\ &\le \|B_M\|_2 + \|R_M\|_2 \\ &\le A_g(\xi_M) + \rho_M = \varepsilon_M,\end{aligned}$$

which proves the response bound. □

If $Z_M := \|P_M g(W_0)\|_2$, then

$$[\|G_M\|_2 - Z_M]_+ \le \|P_M g(W_M)\|_2 \le \|G_M\|_2 + Z_M.$$

Thus, near a stationary dense model, $\|G_M\|_2$ also controls the scale of the gradient initially presented to the fixed-mask fine-tuning optimizer.

A.3 Proof of Proposition 3.5

Proof. Since $q_M = \sum_{j\in\mathcal{Z}_M} w_j e_j$, define

$$z_j := -w_j P_M H_0 e_j, \qquad \widehat{G}_M = -P_M H_0 q_M = \sum_{j\in\mathcal{Z}_M} z_j.$$

For Case (i), suppose $\mathcal{Z}_M \subseteq \mathcal{I}_{\mathrm{low}}$ and $H_0 \succeq 0$. Write $H_0 = R^\top R$. Let $\mathcal{S}_M = [d] \setminus \mathcal{Z}_M$, and let R_S and R_Z contain the columns indexed by $\mathcal{S}_M$ and $\mathcal{Z}_M$, respectively. Writing $w_Z = (W_0)_{\mathcal{Z}_M}$ gives

$$\begin{aligned}\|\widehat{G}_M\|_2^2 &= \|R_S^\top R_Z w_Z\|_2^2 \\ &\le \Lambda_M(W_0)\|R_Z w_Z\|_2^2 \\ &\le \Lambda_M(W_0)|\mathcal{Z}_M| \sum_{j\in\mathcal{Z}_M} w_j^2 (H_0)_{jj} \\ &\le 2\Lambda_M(W_0)|\mathcal{Z}_M|^2\tau_1.\end{aligned}$$

The second inequality follows from Cauchy–Schwarz, and the last from $a_j = \frac{1}{2} w_j^2 (H_0)_{jj} \le \tau_1$ under $H_0 \succeq 0$. Hence

$$\|\widehat{G}_M\|_2 \le |\mathcal{Z}_M| \sqrt{2\Lambda_M(W_0)\tau_1}.$$

For Case (ii), because $P_M e_j = 0$ for every removed coordinate,

$$\|z_j\|_2 \le |w_j| \|(I - e_j e_j^\top) H_0 e_j\|_2 = s_j.$$

Thus, if $\mathcal{Z}_M \subseteq \mathcal{C}_{\text{low}}$,

$$\|\widehat{G}_M\|_2 \le \sum_{j \in \mathcal{Z}_M} \|z_j\|_2 < |\mathcal{Z}_M| \tau_2.$$

For Case (iii), the condition on η_M gives

$$\begin{aligned}
\|\widehat{G}_M\|_2^2 &= \sum_{j \in \mathcal{Z}_M} \|z_j\|_2^2 + \sum_{\substack{j,k \in \mathcal{Z}_M \\ j \ne k}} \langle z_j, z_k \rangle \\
&\ge (1 - \eta_M) \sum_{j \in \mathcal{Z}_M} \|z_j\|_2^2 \\
&\ge (1 - \eta_M) \nu_M^2 \sum_{j \in \mathcal{Z}_M} s_j^2 \\
&\ge (1 - \eta_M) \nu_M^2 |\mathcal{Z}_M| \tau_2^2.
\end{aligned}$$

Therefore,

$$\|\widehat{G}_M\|_2 \ge \nu_M \sqrt{1 - \eta_M}\, \tau_2 \sqrt{|\mathcal{Z}_M|}.$$

Finally, Theorem 3.4 gives

$$\|G_M\|_2 \le \|\widehat{G}_M\|_2 + \varepsilon_M, \qquad \|G_M\|_2 \ge [\|\widehat{G}_M\|_2 - \varepsilon_M]_+,$$

which yields all three stated bounds. □

A.4 Proof of Theorem 3.6

Proof. Adding and subtracting $\Lambda_M(W_M)$ and $\Lambda_M(W_0)$ proves the exact decomposition. The feasible unit vectors in the definition of $\Lambda_M(W_0)$ form a subset of those defining Λ_0; hence $\gamma_M \ge 0$. For symmetric operators A and B, Weyl's inequality gives

$$|\lambda_{\max}(A) - \lambda_{\max}(B)| \le \|A - B\|_2.$$

On each interval between consecutive Hessian discontinuities of a comparison path ζ, Assumption 3.3 bounds the accumulated Hessian variation by $L_H \int_0^1 \|\zeta'(t)\|_2 dt$. Telescoping these intervals and adding the Hessian-discontinuity contributions gives

$$\|H(\zeta(1)) - H(\zeta(0))\|_2 \le L_H \int_0^1 \|\zeta'(t)\|_2 dt + A_H(\zeta).$$

The lengths of ξ_M and χ_{FT} are $\|q_M\|_2$ and at most $r_{\text{FT}}(M)$, respectively. Let U_M have orthonormal columns spanning the retained subspace $\text{range}(P_M)$. Then

$$\Lambda_M(W) = \lambda_{\max}\left(U_M^\top H(W) U_M\right).$$

Hence Weyl's inequality and $\|U_M^\top A U_M\|_2 \le \|A\|_2$ imply

$$|\Lambda_M(W') - \Lambda_M(W)| \le \|H(W') - H(W)\|_2.$$

Applying this inequality to the masking and fine-tuning paths yields the two stagewise bounds.

Finally, the condition $\Delta\Lambda_M > 0$ if and only if $\delta_M + \delta_{\text{FT}} > \gamma_M$ follows directly from the exact decomposition. □

A.5 MASKS SPANNING MULTIPLE TAXONOMY GROUPS

Proposition 3.5 considers masks whose removed weights belong to a single taxonomy group. A general mask may remove weights from several groups simultaneously. For such a mask, partition the removed coordinates as

$$\mathcal{Z}_I := \mathcal{Z}_M \cap \mathcal{I}_{\text{low}},$$
$$\mathcal{Z}_L := \mathcal{Z}_M \cap \mathcal{C}_{\text{low}},$$
$$\mathcal{Z}_H := \mathcal{Z}_M \cap \mathcal{C}_{\text{high}}.$$

For $K \in \{I, L, H\}$, define the group contribution

$$\widehat{G}_{M,K} := -\sum_{j \in \mathcal{Z}_K} w_j P_M H_0 e_j.$$

All contributions use the same projector P_M associated with the full mask. They are components of the joint dense-Hessian prediction, rather than the actual gradient responses to applying three separate masks. In particular,

$$\widehat{G}_M = \widehat{G}_{M,I} + \widehat{G}_{M,L} + \widehat{G}_{M,H}.$$

Under Assumption 3.3, Theorem 3.4 and the triangle inequalities give

$$\|G_M\|_2 \le \sum_{K \in \{I,L,H\}} \|\widehat{G}_{M,K}\|_2 + \varepsilon_M,$$
$$\|G_M\|_2 \ge \Big[\|\widehat{G}_{M,H}\|_2 - \|\widehat{G}_{M,I}\|_2 - \|\widehat{G}_{M,L}\|_2 - \varepsilon_M\Big]_+.$$

Here, ε_M is evaluated along the masking path of the full mask and enters only once.

The low-interaction contribution satisfies

$$\|\widehat{G}_{M,L}\|_2 \le |\mathcal{Z}_L|\tau_2.$$

If $H_0 \succeq 0$, the low-saliency contribution also satisfies

$$\|\widehat{G}_{M,I}\|_2 \le |\mathcal{Z}_I|\sqrt{2\Lambda_M(W_0)\tau_1}.$$

These inequalities follow from the arguments for Cases (i)–(ii) of Proposition 3.5, applied to each subset while keeping the full-mask projector P_M fixed. Empty subsets contribute zero.

A lower bound on the high-interaction contribution additionally requires that its effects reach retained coordinates and do not cancel excessively within $\mathcal{Z}_H$, as in Case (iii). Even then, contributions from the other groups may oppose it. Thus, including high-interaction weights does not by itself guarantee a large joint gradient response: cancellation between groups and the residual ε_M must also be considered. The curvature decomposition in Theorem 3.6 already applies to mixed masks under its stated assumptions.

A.6 CONTINUOUS-GRADIENT SPECIALIZATION

If the gradient is continuous along the masking path and

$$\|H(\xi_M(t)) - H_0\|_2 \le tL_H\|q_M\|_2,$$

then $B_M = 0$ and

$$\|G_M - \widehat{G}_M\|_2 \le \frac{L_H}{2}\|q_M\|_2^2.$$

Thus, the familiar local Taylor expression is recovered as a special case of the general piecewise result.

A.7 OPERATIONAL EMPIRICAL-FISHER TAXONOMY

Let $g_n(W_0)$ denote the gradient of the n-th per-example loss and define the empirical Fisher at the dense model by

$$F_0 = \frac{1}{N}\sum_{n=1}^{N} g_n(W_0)g_n(W_0)^\top.$$

We define the operational individual saliency and interaction score as

$$a_j^F = \frac{1}{2} w_j^2 (F_0)_{jj}, \qquad s_j^F = |w_j| \left\| (I - e_j e_j^\top) F_0 e_j \right\|_2 .$$

Using thresholds τ_1^F, τ_2^F, we construct

$$\widehat{\mathcal{I}}_{\text{low}} = \{ j : a_j^F \le \tau_1^F \},$$

$$\widehat{\mathcal{C}}_{\text{low}} = \{ j : a_j^F > \tau_1^F,\ s_j^F < \tau_2^F \}, \qquad \widehat{\mathcal{C}}_{\text{high}} = \{ j : a_j^F > \tau_1^F,\ s_j^F \ge \tau_2^F \}.$$

In our experiments, these thresholds instantiate a fixed 40/40/20 quantile partition. Specifically, τ_1^F is the 40th percentile of $\{a_j^F\}$, and τ_2^F is the two-thirds quantile of $\{s_j^F : a_j^F > \tau_1^F\}$, yielding the 40/40/20 partition. Within the taxonomy analysis, the empirical Fisher is used only to construct the operational groups; the immediate-response predictor is $\widehat{G}_M = -P_M H_0 q_M$, computed using exact loss-Hessian–vector products.

A.8 Taxonomy transfer and diagnostics

The theoretical and operational scores obey

$$|s_j - s_j^F| \le |w_j| \|H_0 - F_0\|_2.$$

More sharply, one may replace the right-hand side by $|w_j| \|(I - e_j e_j^\top)(H_0 - F_0) e_j\|_2$. Hence the ordering of parameters j and k is certified whenever

$$s_j - s_k > |w_j| \|(H_0 - F_0) e_j\|_2 + |w_k| \|(H_0 - F_0) e_k\|_2.$$

For the individual OBD labels, define

$$a_j^H := \frac{1}{2} w_j^2 |(H_0)_{jj}|.$$

Then

$$|a_j^H - a_j^F| \le \frac{1}{2} w_j^2 |(H_0)_{jj} - (F_0)_{jj}|.$$

If $\|H_0 - F_0\|_{\max} \le \varepsilon_\infty$ and $\tau_1^F = \tau_1$, it follows that

$$a_j^F > \tau_1 + \frac{1}{2} w_j^2 \varepsilon_\infty \Longrightarrow a_j^H > \tau_1,$$

and

$$a_j^F < \tau_1 - \frac{1}{2} w_j^2 \varepsilon_\infty \Longrightarrow a_j^H < \tau_1.$$

Only coordinates in the resulting weight-dependent threshold band remain uncertified. When the Hessian and Fisher thresholds are independently quantile-calibrated, one may instead assess OBD-label overlap, score-rank agreement, and low/high group overlap between the two curvature constructions.

Diagnostic interpretation. Each parameter receives one individual OBD saliency and one global interaction score. The theoretical and operational taxonomies therefore use the same global grouping construction, but their memberships need not coincide: the former uses H_0, whereas the latter uses its empirical-Fisher proxy F_0. This transfer can be assessed through score-rank agreement and group overlap. Both scores include the weight because pruning changes coordinate j by $-w_j$ rather than by a unit perturbation.

The price is equally explicit. A large global score can be concentrated on coordinates that are pruned together and therefore removed by P_M; this effect is captured by ν_M. Even when the interaction survives the mask, different gradient contributions can cancel; the cancellation is captured by η_M. The global taxonomy is consequently most directly tied to the immediate pruning response. Its connection to fine-tuned curvature remains a separate, experimentally testable propagation question.

Operational diagnostics and group-wise interventions. The empirical-Fisher scores $\{a_j^F, s_j^F\}$ are used to construct the operational taxonomy, whereas the immediate masking response is evaluated using the dense-Hessian prediction

$$\widehat{G}_M = -P_M H_0 q_M.$$

We compare $\widehat{G}_M$ with the observed response G_M through $\|\widehat{G}_M\|_2$, $\|G_M\|_2$, directional agreement c_{pred}, and the normalized residual r_{pred}.

For the main experiment, we instantiate the operational taxonomy globally using the fixed 40/40/20 partition and construct one 5% pruning mask from each of $\mathcal{I}_{\mathrm{low}}$, $\mathcal{C}_{\mathrm{low}}$, and $\mathcal{C}_{\mathrm{high}}$. The three masks remove the same number of parameters and use identical layer quotas. Within each layer, removed-weight energy is matched as closely as the available group supports permit, and $\|q_M\|_2$ is reported to expose any remaining difference in masking displacement. Accordingly, this experiment is interpreted as a controlled group-wise comparison rather than an exactly covariate-matched causal intervention.

B Experimental Details

B.1 Vision Training and Pruning Setup

Datasets and dense models. We use CIFAR-100 and Tiny-ImageNet with their standard training and validation splits. ResNet-50 models are trained from scratch, whereas ViT-B/16 models are initialized from ImageNet-1K pretrained weights and adapted to the target dataset. The dense task models are fixed before pruning and serve as W_0 for every pruning method. The original dense-model recipe uses 100 epochs of SGD with momentum 0.9, weight decay 5×10^{-4}, initial learning rate 0.1, and cosine decay. We use five independently trained dense models for each architecture–dataset pair.

For ResNet-50, CIFAR-100 training uses random 32×32 crops with four-pixel padding and horizontal flips; Tiny-ImageNet uses random 64×64 crops with eight-pixel padding and horizontal flips. Evaluation uses unaugmented validation images. For ViT-B/16, training images use random resized crops to 224×224 with horizontal flips, while evaluation uses resize followed by a center crop. ViT inputs use the ImageNet-1K normalization of the pretrained weights.

Pruning scope and scores. Vision pruning is global and unstructured over convolutional and linear weight tensors. Biases and normalization parameters are not pruned. For ViT-B/16, the patch projection, MLP matrices, attention output projections, and fused query–key–value matrices are included; class and positional embeddings are excluded. Magnitude pruning ranks coordinate j by $|w_j|$. Fisher pruning ranks it by $w_j^2 \widehat{F}_{jj}$, where $\widehat{F}_{jj}$ is estimated from squared calibration-gradient entries. The factor $1/2$ in the corresponding OBD score is omitted because it does not change the ranking. A reported sparsity is the fraction removed within this prunable coordinate set.

Fixed-mask fine-tuning. After pruning, every removed coordinate is set to zero. During fine-tuning, its gradient is masked and the parameter mask is reapplied after every optimizer step, preventing regrowth. We fine-tune for five epochs with AdamW, zero weight decay, and batch size 128. The learning rate is 5×10^{-5} for ResNet-50 and 10^{-5} for ViT-B/16. All evaluations of W_{FT} load the corresponding dense model, pruning mask, and fine-tuned retained weights from the same model seed. We verify that all removed coordinates remain exactly zero in every saved checkpoint.

B.2 Sensitivity and Robustness Evaluation

Retained-parameter sensitivity. For a mini-batch $\mathcal{B}$, let $\mathcal{L}_{\mathcal{B}}$ denote the mean cross-entropy loss. At evaluation state W and mask M, we form the retained-gradient perturbation

$$\delta_{\mathcal{B}}^{\mathrm{act}}(W; M) := \rho \, \frac{P_M \nabla \mathcal{L}_{\mathcal{B}}(W)}{\|P_M \nabla \mathcal{L}_{\mathcal{B}}(W)\|_2}, \qquad \rho = 0.05,$$

and report

$$S_\rho(W; M) := \mathbb{E}_{\mathcal{B}} \left[\mathcal{L}_{\mathcal{B}}\big(W + \delta_{\mathcal{B}}^{\mathrm{act}}(W; M)\big) - \mathcal{L}_{\mathcal{B}}(W) \right].$$

The same mini-batch is used before and after perturbation. Both the gradient and perturbation are projected by P_M, and the pruning mask is reapplied before the perturbed forward pass. Thus, pruned

coordinates cannot be reactivated and the measured directions coincide with the retained subspace used by the theory. We average loss increases with weights proportional to mini-batch size over the full evaluation split and use the same radius and batching protocol for all sparsities.

Corruption and adversarial robustness. Corruption accuracy is averaged over the 15 standard corruption types and all five severity levels of CIFAR-100-C or Tiny-ImageNet-C. Adversarial accuracy uses a 10-step ℓ_∞ PGD attack with radius $1/255$, step size $0.25/255$, and a uniform random start in the perturbation ball. Updates and clipping are performed in pixel space, after which inputs are normalized for the model. Clean, corruption, and PGD accuracies are evaluated on the same fixed-mask endpoint W_{FT}; no additional adaptation is performed for a robustness test.

B.3 Hessian and Immediate-Response Diagnostics

Operational taxonomy. For the controlled experiment in Section 4.3, we estimate the empirical Fisher on a fixed stratified subset of $N = 1000$ training examples,

$$F_0 := \frac{1}{N}\sum_{n=1}^{N} g_n g_n^\top, \qquad g_n := \nabla \ell_n(W_0).$$

The individual deletion score is $a_j^F = \frac{1}{2} w_j^2 (F_0)_{jj}$. To obtain a scalable interaction score, we use $K = 32$ Rademacher probes r_k and compute

$$\widehat{s}_j^F := |w_j| \sqrt{\left[\frac{1}{K}\sum_{k=1}^{K} \big((F_0 r_k)_j\big)^2 - (\widehat{F}_{jj})^2\right]_+}.$$

This estimates the off-diagonal row energy associated with coordinate j; the diagonal contribution is subtracted so that individual deletion saliency and interaction are not counted twice. Scores are computed over the complete vision pruning scope rather than independently normalizing every layer.

Group-wise masking intervention. We instantiate the operational taxonomy globally over the complete prunable parameter space. The bottom 40% of coordinates by a_j^F form $\mathcal{I}_{\mathrm{low}}$. The remaining coordinates are ordered by $\widehat{s}_j^F$ and divided into $\mathcal{C}_{\mathrm{low}}$ and $\mathcal{C}_{\mathrm{high}}$, containing 40% and 20% of the full prunable space, respectively. For each group, we construct a mask that removes 5% of the full prunable parameter space, so the three interventions have identical removed-parameter counts.

The three masks use identical layer quotas. Within each layer, coordinates are selected so that removed-weight energy is matched as closely as the group supports permit. Exact energy matching is not always feasible at 5% sparsity, so we report $\|q_M\|_2$ for every intervention to make the remaining difference in masking displacement explicit.

Exact response measurement. The taxonomy is constructed with the empirical-Fisher surrogate, but the leading term reported in Section 4.3 uses the exact loss Hessian on the same fixed 1000-example subset. With

$$G_M := P_M[g(W_M) - g(W_0)], \qquad \widehat{G}_M := -P_M H(W_0) q_M,$$

we report $\|G_M\|_2$, $\|\widehat{G}_M\|_2$, and

$$c_{\mathrm{pred}}(M) := \frac{\langle G_M, \widehat{G}_M\rangle}{\|G_M\|_2 \|\widehat{G}_M\|_2}, \qquad r_{\mathrm{pred}}(M) := \frac{\|G_M - \widehat{G}_M\|_2}{\|G_M\|_2}.$$

The Hessian–vector product $H(W_0)q_M$ is obtained by automatic differentiation of the mean cross-entropy loss; it is not replaced by the Fisher in these reported diagnostics. All gradients and Hessian actions are projected onto exactly the same retained coordinates.

B.4 Curvature Estimation

We estimate the largest eigenvalue of the Hessian restricted to the retained parameter subspace,

$$\Lambda_M(W) = \lambda_{\max}\big(U_M^\top H(W) U_M\big),$$

where the columns of U_M form an orthonormal basis of $\mathrm{range}(P_M)$. We use 16-step Lanczos iteration with exact Hessian–vector products. In implementation, Lanczos vectors are initialized in $\mathrm{range}(P_M)$, and both the vectors and Hessian actions are projected by P_M after every operation. This implements the restricted operator without explicitly constructing U_M.

For every decomposition, W_0, W_M, and W_{FT} use the same fixed stratified subset of 1,000 training examples, the same mean cross-entropy objective, and the same mask. The operator is evaluated in FP32 with TF32 disabled, and trainable parameters outside the pruning scope remain active.

B.5 Language and Code Model Setup

Models and Wanda pruning. We evaluate the base checkpoints Qwen2-0.5B, Qwen2.5-1.5B, Pythia-1.4B, and Qwen2.5-Coder at 0.5B, 1.5B, and 3B parameters. No post-pruning fine-tuning is performed. We apply layerwise Wanda pruning at 20%, 30%, and 40% sparsity to every attention and MLP linear weight, using $|W_{ij}|\|X_j\|_2$ as the pruning score. Embeddings, normalization parameters, biases, and the language-model head are excluded. Wanda statistics use 32 calibration sequences of length 128. General-language models use WikiText-2 calibration text, whereas the Coder models use MBPP code text. Calibration examples are disjoint from the sequences used for the immediate-response measurement and from downstream evaluation tasks.

LLM Hessian diagnostics. We measure G_M on 16 held-out sequences of length 128 and compute the exact autograd action $P_M H(W_0) q_M$ over the full prunable parameter space. To fit the computation in memory, q_M is partitioned into groups of transformer blocks and we accumulate

$$P_M H(W_0) q_M = \sum_b P_M H(W_0) q_{M,b}.$$

Each summand is differentiated with respect to all retained output coordinates, so cross-block Hessian terms are preserved; the procedure is a memory decomposition of an exact Hessian action, not a block-diagonal Hessian approximation. Gradients are stored in bfloat16 and accumulated Hessian actions in FP32. We use $Z_M = \|P_M g(W_0)\|_2$ to normalize $\|G_M\|_2$ and report the same cosine and residual diagnostics as in the vision experiments.

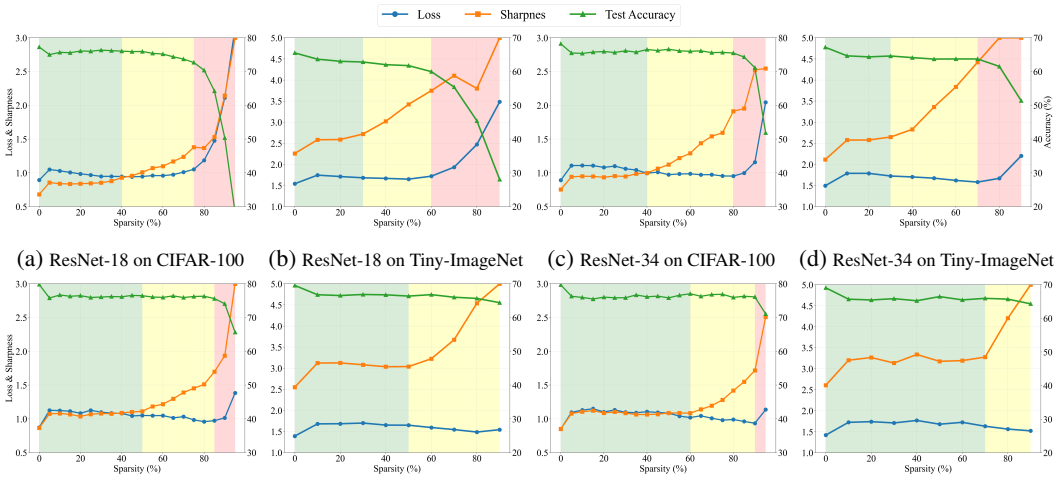
Downstream evaluation. We compute perplexity on the complete WikiText-2 test split by concatenating documents with double newlines, tokenizing without added special tokens, and evaluating non-overlapping 2,048-token blocks after dropping the final incomplete block. Zero-shot HellaSwag and PIQA use their complete validation sets with length-normalized multiple-choice accuracy. Prompt construction and answer scoring follow the official evaluation-harness implementation. For Coder models, HumanEval pass@1 uses all 164 problems, greedy decoding, a maximum of 1024 newly generated tokens, and one completion per problem. Generated programs are executed with restricted imports and a five-second timeout.

For the final reported tables, each task evaluation and its corresponding Hessian diagnostic use the same serialized pruning mask. We record the checkpoint revision, selected tensor names, number of removed parameters, and a mask hash for this audit.

C Supplement to Section 4.2

We conducted pruning experiments across a range of ResNet architectures, measuring test loss, accuracy, and sharpness as a function of sparsity under Fisher pruning. Consistent with the results in the main text, we were able to identify green, yellow, and red regions for all networks in Figure 4. Notably, as the model size increases, the span of the red region becomes smaller. Additionally, the sparsity level at which the tipping phase begins varies with model size. Specifically, on CIFAR-100, the tipping phase for ResNet-18 and ResNet-34 commences at 40% sparsity. For the larger ResNet-101 and ResNet-152 models, however, it begins later at 50% and 60%, respectively, and their severe performance degradation is also delayed to higher sparsity. A similar pattern is found on Tiny-ImageNet: both ResNet-18 and ResNet-34 enter the tipping phase early and collapse rapidly, whereas ResNet-101 and ResNet-152 enter the tipping phase and collapse later.

Table 5 provides the full sparsity-wise robustness results used to construct the summary in the main text. For each dataset, architecture, and pruning method, we report clean accuracy, corruption



(a) ResNet-18 on CIFAR-100 (b) ResNet-18 on Tiny-ImageNet (c) ResNet-34 on CIFAR-100 (d) ResNet-34 on Tiny-ImageNet

(e) ResNet-101 on CIFAR-100 (f) ResNet-101 on Tiny-ImageNet (g) ResNet-152 on CIFAR-100 (h) ResNet-152 on Tiny-ImageNet

Figure 4: Sparsity (i.e., a portion of pruned parameters) vs. test loss, test accuracy, and sharpness S_ρ for ResNet architectures trained on the CIFAR-100/Tiny-ImageNet datasets under Fisher pruning

Table 5: A generalization test on CIFAR-100-C and Tiny-ImageNet-C

ResNet-50 for Fisher pruning

Sparsity(%)	0	20	40	60	70	80	90
CIFAR-100	77.47	74.80	75.24	75.87	76.05	76.25	75.40 (-2.6%)
CIFAR-100-C	50.01	46.76	47.92	48.02	48.50	48.19	47.25 (-5.5%)
PGD	21.31	17.94	18.72	19.09	19.30	18.05	15.68 (-26%)
S_ρ	1.01	1.26	1.26	1.35	1.44	1.60	1.96 (+0.95)

ResNet-50 for Magnitude pruning

Sparsity(%)	0	20	40	60	70	80	90
CIFAR-100	77.47	74.82	75.20	75.78	76.11	76.41	76.33 (-1.5%)
CIFAR-100-C	50.01	47.45	47.36	47.88	48.13	48.53	48.14 (-3.7%)
PGD	21.31	18.20	18.90	18.97	18.80	18.70	16.97 (-20%)
S_ρ	1.01	1.27	1.30	1.33	1.40	1.52	1.90 (+0.89)

ViT-B/16 for Fisher pruning

Sparsity(%)	0	20	40	60	70	80	90
CIFAR-100	88.91	86.85	86.11	86.84	85.83	83.37 (-6.2%)	77.91
CIFAR-100-C	71.05	67.78	66.39	68.16	68.12	65.45 (-7.9%)	56.03
PGD	24.52	20.79	18.56	16.75	14.80	13.26 (-46%)	6.09
S_ρ	0.63	0.61	0.59	0.60	0.65	0.70 (+0.07)	1.23

ViT-B/16 for Magnitude pruning

Sparsity(%)	0	20	40	60	70	80	90
CIFAR-100	88.91	87.30	87.29	86.59	85.47 (-3.9%)	77.26	31.56
CIFAR-100-C	71.05	69.35	70.02	68.64	65.95 (-7.2%)	58.22	21.45
PGD	24.52	18.28	19.24	14.28	10.38 (-58%)	3.13	5.99
S_ρ	0.63	0.61	0.76	1.11	1.31 (+0.68)	1.75	2.21

ResNet-50 for Fisher pruning

Sparsity(%)	0	20	40	60	70	80	90
Tiny-IN	67.04	63.68	64.35	64.89	65.08	64.70 (-3.5%)	63.66
Tiny-IN-C	25.19	23.21	23.56	23.65	23.51	23.51 (-6.7%)	22.94
PGD	22.04	17.65	18.32	18.77	18.83	18.18 (-18%)	17.17
S_ρ	2.74	3.39	3.42	3.53	3.80	4.32 (+1.58)	5.67

ResNet-50 for Magnitude pruning

Sparsity(%)	0	20	40	60	70	80	90
Tiny-IN	67.04	64.26	64.38	64.90	65.15	65.32 (-2.6%)	64.73
Tiny-IN-C	25.19	23.73	23.74	24.24	24.04	24.15 (-4.1%)	23.48
PGD	22.04	17.99	18.20	19.02	19.03	18.67 (-15%)	17.09
S_ρ	2.74	3.41	3.49	3.63	3.82	4.25 (+1.51)	5.59

ViT-B/16 for Fisher pruning

Sparsity(%)	0	20	40	60	70	80	90
Tiny-IN	89.55	85.75	86.56	86.01	85.31 (-4.7%)	80.04	45.43
Tiny-IN-C	58.31	54.29	56.17	54.94	54.04 (-7.3%)	49.52	17.87
PGD	30.52	12.42	14.10	15.94	17.18 (-44%)	7.8	5.91
S_ρ	0.87	0.83	0.82	0.98	1.07 (+0.20)	1.72	3.58

ViT-B/16 for Magnitude pruning

Sparsity(%)	0	20	40	60	70	80	90
Tiny-IN	89.55	86.60	86.83	86.12	83.26 (-7.0%)	36.05	12.37
Tiny-IN-C	58.31	54.28	54.84	54.10	51.34 (-12%)	15.68	6.30
PGD	30.52	11.48	12.17	9.43	5.63 (-82%)	1.12	1.94
S_ρ	0.87	0.87	1.18	1.84	2.32 (+1.45)	4.03	0.92

accuracy, PGD accuracy, and sharpness across sparsity levels. Table 6 additionally reports 95% Student-t confidence intervals for the ResNet-50 corruption and PGD results, quantifying the variation across the five seeds.

Consistent with the main results, degradation under corrupted and adversarial inputs is generally larger than the clean-accuracy drop near the end of the tipping phase, while sharpness increases. This shows that increased sharpness and robustness degradation can co-occur while clean accuracy remains comparatively preserved.

Table 7 reports a finer-grained sparsity sweep for ResNet-50 on CIFAR-100 and Tiny-ImageNet, including clean accuracy, test loss, and S_ρ with 95% confidence intervals. Across both datasets and pruning criteria, clean accuracy remains comparatively stable over a broad sparsity range, while S_ρ increases more noticeably at higher sparsities. This further illustrates that clean accuracy and test loss alone do not fully capture the change in retained-parameter sensitivity.

Table 6: Confidence intervals for corruption and PGD accuracy with ResNet-50. Entries are means ± 95% Student-t confidence interval half-width.

Sparsity	Corruption accuracy ↑				PGD-10 accuracy ↑			
	CIFAR-100-C		Tiny-ImageNet-C		CIFAR-100		Tiny-ImageNet	
	Fisher	Magnitude	Fisher	Magnitude	Fisher	Magnitude	Fisher	Magnitude
0%	50.01±1.46	50.01±1.46	25.19±0.48	25.19±0.48	21.31±4.19	21.31±4.17	22.04±1.52	22.04±1.52
20%	46.76±0.30	47.45±0.76	23.21±0.67	23.73±1.20	17.94±2.73	18.20±3.30	17.65±0.94	17.99±1.06
40%	47.92±1.46	47.36±1.17	23.56±0.73	23.74±0.86	18.72±3.59	18.90±3.65	18.32±0.87	18.20±1.09
60%	48.02±0.91	47.88±0.40	23.65±0.77	24.24±0.88	19.09±3.41	18.97±3.59	18.77±1.18	19.02±1.43
70%	48.50±0.80	48.13±1.04	23.51±1.06	24.04±0.57	19.30±3.38	18.80±3.73	18.83±1.03	19.03±0.98
80%	48.19±0.62	48.53±0.98	23.51±0.78	24.15±0.47	18.05±2.96	18.70±3.63	18.18±0.66	18.67±0.68
90%	47.25±0.97	48.14±0.72	22.94±0.33	23.48±0.22	15.68±3.12	16.97±3.14	17.17±0.56	17.09±0.74

Table 7: ResNet-50 results across pruning sparsity. Entries are means ± 95% Student-t confidence intervals over five seeds. The dense checkpoint at 0% sparsity is shared by both pruning methods.

Sparsity	CIFAR-100			Tiny-ImageNet		
	Acc. ↑	Loss ↓	S_ρ ↓	Acc. ↑	Loss ↓	S_ρ ↓
Dense						
0%	77.47±0.47	0.946±0.039	1.015±0.053	67.04±1.01	1.478±0.057	2.739±0.072
Fisher pruning						
10%	74.74±0.91	1.172±0.044	1.264±0.113	64.03±0.70	1.795±0.055	3.426±0.160
20%	74.80±0.58	1.171±0.035	1.260±0.119	63.68±1.18	1.799±0.067	3.388±0.093
30%	74.81±0.64	1.164±0.037	1.265±0.092	63.96±1.15	1.777±0.064	3.422±0.067
40%	75.24±0.80	1.134±0.045	1.259±0.090	64.35±0.83	1.755±0.039	3.417±0.112
50%	75.57±0.46	1.117±0.037	1.295±0.095	64.41±0.61	1.744±0.044	3.459±0.057
60%	75.87±0.72	1.087±0.024	1.353±0.115	64.89±0.87	1.696±0.077	3.526±0.078
70%	76.05±0.56	1.045±0.025	1.436±0.134	65.08±0.87	1.665±0.063	3.798±0.133
80%	76.25±0.57	0.986±0.029	1.600±0.164	64.70±0.51	1.628±0.047	4.315±0.141
90%	75.40±0.82	0.946±0.034	1.965±0.218	63.66±0.61	1.573±0.052	5.673±0.399
95%	72.84±0.48	0.990±0.028	2.359±0.204	60.40±1.40	1.631±0.055	7.242±0.396
Magnitude pruning						
10%	74.75±0.30	1.178±0.034	1.265±0.109	63.86±1.23	1.806±0.075	3.406±0.110
20%	74.82±0.57	1.176±0.029	1.272±0.118	64.26±0.72	1.784±0.051	3.406±0.097
30%	75.05±0.77	1.171±0.045	1.277±0.109	64.13±0.96	1.796±0.061	3.452±0.069
40%	75.20±0.60	1.162±0.036	1.299±0.121	64.38±0.73	1.782±0.060	3.490±0.102
50%	75.56±0.57	1.145±0.040	1.300±0.101	64.59±0.61	1.780±0.062	3.543±0.080
60%	75.78±0.64	1.120±0.021	1.334±0.125	64.91±1.02	1.743±0.067	3.631±0.115
70%	76.11±0.65	1.090±0.023	1.403±0.125	65.16±0.54	1.699±0.053	3.818±0.094
80%	76.41±0.53	1.041±0.030	1.518±0.115	65.32±1.29	1.653±0.067	4.246±0.209
90%	76.33±0.58	0.966±0.022	1.905±0.150	64.73±0.84	1.591±0.051	5.589±0.337
95%	74.93±0.97	0.930±0.032	2.459±0.181	62.93±0.73	1.555±0.048	7.342±0.442

D SUPPLEMENT TO SECTION 4.3

Table 8 describes how ordinary Fisher pruning overlaps with the operational taxonomy. It is a mask-composition diagnostic, not a sequential taxonomy-guided pruning rule. At low sparsity, Fisher pruning predominantly removes low-saliency coordinates; as sparsity increases, it also enters both interaction groups.

The main text evaluates equal-count 5% interventions across the three groups of the 40/40/20 operational taxonomy. Table 9 applies the same group-wise construction at a more aggressive setting, removing 10% of the full prunable parameter space from one group at a time. This stress test examines how the immediate response changes beyond the main intervention regime and whether subsequent fixed-mask fine-tuning can recover task accuracy. Because the larger displacement may amplify non-leading terms, it is not a verification of the small-remainder regime.

Table 8: Taxonomy composition of Fisher-pruned coordinates on CIFAR-100. Entries are the mean number of removed coordinates, in millions, over five dense-model seeds.

Architecture	Group	20%	40%	60%	80%
ResNet-50	$\mathcal{I}_{\text{low}}$	4.61	7.96	9.21	9.43
	$\mathcal{C}_{\text{low}}$	0.12	1.35	4.27	7.43
	$\mathcal{C}_{\text{high}}$	0.00	0.15	0.71	2.05
ViT-B/16	$\mathcal{I}_{\text{low}}$	17.10	30.70	33.90	34.21
	$\mathcal{C}_{\text{low}}$	0.02	3.51	16.75	29.15
	$\mathcal{C}_{\text{high}}$	0.00	0.04	0.72	5.12

Table 9: Stand-alone 10% taxonomy stress test on CIFAR-100. Entries are five-seed means; FT denotes five-epoch fixed-mask fine-tuning.

	ResNet-50				ViT-B/16			
	Dense	$\mathcal{I}_{\text{low}}$	$\mathcal{C}_{\text{low}}$	$\mathcal{C}_{\text{high}}$	Dense	$\mathcal{I}_{\text{low}}$	$\mathcal{C}_{\text{low}}$	$\mathcal{C}_{\text{high}}$
Accuracy at W_M (%)	77.47	77.11	77.06	66.57	89.33	89.28	85.29	3.51
$\|G_M\|_2$ at W_M	0.000	0.225	0.194	9.052	0.000	0.068	2.143	11.435
$S_{0.05}$ at W_M	1.050	1.002	0.997	1.426	0.672	0.653	0.749	0.710
Accuracy after FT (%)	77.47	75.52	75.40	75.62	89.33	89.04	88.67	87.36
$S_{0.05}$ after FT	1.050	1.282	1.275	1.243	0.672	0.674	0.700	0.876

At 10% sparsity, high-interaction pruning yields the largest $\|G_M\|_2$ on both architectures and causes severe immediate accuracy loss, particularly for ViT-B/16. Fine-tuning recovers much of the accuracy, but the endpoint sensitivity no longer follows the immediate $\|G_M\|_2$ ordering. This is consistent with the separation between the immediate-response theorem and the endpoint-curvature decomposition.

E Supplement to Section 4.4

E.1 Fine-tuning locality across sparsity

Table 10 reports parameter movement over the complete ResNet-50/CIFAR-100 sparsity sweep. We additionally define

$$R_{\text{move}} := \frac{\|W_{\text{FT}} - W_M\|_2}{\|q_M\|_2}$$

to compare the absolute movements induced by fine-tuning and masking. We also define the relative masking displacement, relative fine-tuning displacement, and parameter-vector cosine similarity as

$$d_{\text{mask}} := \frac{\|q_M\|_2}{\|W_0\|_2}, \qquad d_{\text{FT}} := \frac{\|W_{\text{FT}} - W_M\|_2}{\|W_M\|_2}, \qquad c_{\text{FT}} := \cos(W_M, W_{\text{FT}}).$$

The first two quantities are reported as percentages.

Table 10: Global parameter displacement after five-epoch fixed-mask fine-tuning. Values are mean $\pm$ sample standard deviation over five seeds.

Method	Sparsity	d_{mask} (%)	d_{FT} (%)	c_{FT}	R_{move}
Fisher	20%	5.39 ± 0.22	17.48 ± 1.50	0.9860 ± 0.0022	3.25 ± 0.34
	40%	14.01 ± 0.56	15.73 ± 1.35	0.9886 ± 0.0018	1.11 ± 0.08
	60%	26.03 ± 0.98	13.96 ± 1.16	0.9910 ± 0.0014	0.52 ± 0.03
	80%	43.87 ± 1.40	12.74 ± 0.88	0.9925 ± 0.0009	0.26 ± 0.01
Magnitude	20%	2.91 ± 0.22	17.45 ± 1.53	0.9860 ± 0.0022	6.04 ± 0.88
	40%	9.71 ± 0.38	15.89 ± 1.37	0.9884 ± 0.0018	1.63 ± 0.15
	60%	20.51 ± 0.61	13.75 ± 1.30	0.9913 ± 0.0015	0.66 ± 0.06
	80%	37.72 ± 0.82	11.60 ± 0.83	0.9939 ± 0.0008	0.28 ± 0.02

Although c_{FT} remains close to one, locality depends on the stage being considered. Fine-tuning movement dominates at low sparsity, whereas masking movement dominates at high sparsity. The two stages must therefore be quantified separately, as in Theorem 3.6.

Table 11: Layerwise relative fine-tuning displacement on ResNet-50/CIFAR-100. Statistics are first computed across the six network stages and then averaged over five seeds.

Method	Sparsity	Median (%)	90th percentile (%)	Maximum (%)
Fisher	20%	11.53 ± 1.56	18.87 ± 1.59	19.44 ± 1.81
	40%	10.54 ± 1.24	16.88 ± 1.39	17.28 ± 1.52
	60%	9.89 ± 1.00	14.69 ± 1.13	14.85 ± 1.10
	80%	9.20 ± 0.54	12.83 ± 0.80	14.21 ± 1.06
Magnitude	20%	12.41 ± 1.76	18.75 ± 1.62	19.35 ± 1.90
	40%	11.50 ± 1.63	16.89 ± 1.32	17.27 ± 1.45
	60%	10.62 ± 1.43	14.37 ± 1.26	14.50 ± 1.21
	80%	9.04 ± 0.55	12.46 ± 1.23	12.91 ± 1.80

E.2 Full curvature decomposition on CIFAR-100

Table 12 extends the representative main-text endpoint to all evaluated ResNet-50 sparsities. The masking and fine-tuning terms often oppose one another, and their balance varies with sparsity.

Table 12: Retained-subspace curvature decomposition on ResNet-50/CIFAR-100. All terms are normalized by Λ_0 and reported as mean $\pm$ sample standard deviation over five seeds.

Method	Sparsity	$-\gamma_M/\Lambda_0$	δ_M/Λ_0	$\delta_{\mathrm{FT}}/\Lambda_0$	$\Delta\Lambda_M/\Lambda_0$
Fisher	20%	-0.03 ± 0.00	-0.01 ± 0.06	3.12 ± 2.73	3.08 ± 2.73
	40%	-0.08 ± 0.01	0.19 ± 0.41	1.76 ± 0.73	1.88 ± 1.09
	60%	-0.17 ± 0.03	11.42 ± 9.15	-9.48 ± 7.52	1.77 ± 1.72
	80%	-0.31 ± 0.05	127.41 ± 93.35	-124.78 ± 91.22	2.32 ± 2.76
Magnitude	20%	-0.07 ± 0.01	0.01 ± 0.05	2.76 ± 3.25	2.70 ± 3.23
	40%	-0.17 ± 0.02	0.15 ± 0.21	1.46 ± 1.07	1.44 ± 1.10
	60%	-0.28 ± 0.03	1.26 ± 1.50	0.84 ± 0.69	1.82 ± 2.18
	80%	-0.44 ± 0.04	17.90 ± 13.66	-16.78 ± 12.84	0.68 ± 0.89

E.3 Endpoint Curvature Decomposition on Tiny-ImageNet

We repeat the endpoint analysis on ResNet-50/Tiny-ImageNet at 80% sparsity. All curvature terms use exact loss-Hessian–vector products on the set of 1,000 examples and 16-step Lanczos iteration.

Table 13: Parameter movement and retained-subspace curvature decomposition at 80% sparsity on ResNet-50/Tiny-ImageNet. Curvature terms are normalized by Λ_0. Values are mean $\pm$ sample standard deviation over five seeds.

	Parameter movement			Curvature decomposition			
Method	d_{mask} (%)	d_{FT} (%)	c_{FT}	$-\gamma_M/\Lambda_0$	δ_M/Λ_0	$\delta_{\mathrm{FT}}/\Lambda_0$	$\Delta\Lambda_M/\Lambda_0$
Fisher	38.99 ± 2.30	15.56 ± 0.91	0.9895 ± 0.0011	-0.23 ± 0.08	$(1.84 \pm 1.83) \times 10^3$	$(-1.84 \pm 1.83) \times 10^3$	2.86 ± 2.12
Magnitude	32.63 ± 1.72	13.85 ± 0.75	0.9921 ± 0.0007	-0.36 ± 0.09	53.90 ± 35.63	-52.79 ± 35.10	0.75 ± 0.71

As on CIFAR-100, masking and fine-tuning induce large, opposing curvature drifts despite relatively high parameter-space cosine similarity. The mean endpoint change is positive for both methods: it is positive in all five Fisher seeds and four of five Magnitude seeds. The substantial seed variation, particularly for Fisher pruning, supports an empirical tendency.

E.4 Numerical checks

For the CIFAR-100 sweep, the maximum mask violation is zero, the maximum normalized closure error is 4.97×10^{-13}, and the maximum relative Lanczos Ritz residual is 3.82×10^{-4}. For the Tiny-ImageNet replication, the maximum mask violation is also zero and the maximum normalized closure error is 4.09×10^{-13}. These checks validate the implementation of the fixed-mask decomposition but do not determine the sign of the empirical endpoint change.

F Supplement to Section 4.5

The main text reports changes relative to each dense model. Table 14 provides the corresponding absolute values; HellaSwag and PIQA use normalized zero-shot accuracy.

Table 14: Absolute performance without fine-tuning at dense/20%/30%/40% sparsity. PPL is WikiText-2 perplexity; HS and PIQA are normalized accuracies in percent.

Model	PPL	HS (%)	PIQA (%)
Qwen2-0.5B	13.37 / 13.60 / 14.25 / 15.94	49.05 / 48.50 / 47.05 / 44.09	69.21 / 68.93 / 69.31 / 68.34
Qwen2.5-1.5B	9.27 / 9.41 / 9.80 / 10.90	67.76 / 67.23 / 65.49 / 61.82	75.73 / 76.01 / 75.19 / 72.96
Pythia-1.4B	11.92 / 12.16 / 12.85 / 14.39	52.03 / 51.51 / 50.96 / 48.73	71.16 / 70.78 / 70.35 / 69.26

Table 15: Absolute HumanEval pass@1 (%) at dense/20%/30%/40% sparsity.

Model	HumanEval pass@1 (%)
Qwen2.5-Coder-0.5B	27.44 / 23.17 / 17.68 / 14.63
Qwen2.5-Coder-1.5B	42.68 / 40.85 / 39.02 / 31.71
Qwen2.5-Coder-3B	46.34 / 42.68 / 45.12 / 40.85

For the Coder models, the HVP and HumanEval runs use SDPA attention and the same deterministic Wanda calibration protocol. Across all nine conditions, an independently reconstructed HumanEval mask matches the HVP mask hash, and the removed count and $\|q_M\|_2$ agree exactly.

G Additional Ablation Studies

G.1 Exact Hessian action versus empirical Fisher

Table 16 isolates the choice of prediction operator on Qwen2-0.5B while holding the Wanda masks, gradient samples, and normalization fixed. The exact loss-Hessian action provides substantially stronger directional agreement than the empirical-Fisher action.

For each operator $A \in \{H_0, F_0\}$, define

$$\widehat{G}_M^{(A)} := -P_M A q_M.$$

The cosine and residual diagnostics use the corresponding operator-specific prediction in each row.

Table 16: Prediction-operator comparison for Qwen2-0.5B.

Sparsity	Operator	$\|G_M\|_2/Z_M$	$\|\widehat{G}_M^{(A)}\|_2/Z_M$	$c_{\text{pred}} \uparrow$	$r_{\text{pred}} \downarrow$
20%	Empirical Fisher	0.289	0.013	0.041	0.999
	Exact HVP	0.289	0.290	0.779	0.666
30%	Empirical Fisher	0.498	0.040	0.008	1.003
	Exact HVP	0.498	0.585	0.655	0.917
40%	Empirical Fisher	0.702	0.050	0.165	0.991
	Exact HVP	0.702	0.865	0.574	1.051

The Fisher matrix remains useful for scalable mask construction, but this ablation supports using the loss Hessian itself when testing the immediate gradient-transport identity. At 40% sparsity, even the exact leading term has $r_{\text{pred}} > 1$, consistent with increasingly important nonlinear remainder terms.

G.2 Sensitivity to Taxonomy Thresholds τ_1 and τ_2

The theoretical taxonomy is defined by thresholds τ_1 and τ_2. In our operational implementation, these thresholds are calibrated through quantiles of the empirical-Fisher scores. We examine whether the group-conditioned response observed in Section 4.3 is sensitive to the choice of these thresholds.

$$\tau_1^F = Q_{q_1}\left(\{a_j^F\}_j\right), \qquad \tau_2^F = Q_{q_2}\left(\{s_j^F : a_j^F > \tau_1^F\}\right).$$

The main experiments use a 40/40/20 split for $\mathcal{I}_{\text{low}}$, $\mathcal{C}_{\text{low}}$, and $\mathcal{C}_{\text{high}}$. To test whether the local ordering is specific to this choice, we vary the first quantile over $\{0.30, 0.40, 0.50\}$ and the interaction quantile within the retained high-saliency pool over $\{0.50, 2/3, 0.75\}$. For each cell and seed, the $\mathcal{C}_{\text{low}}$ and $\mathcal{C}_{\text{high}}$ masks remove 0.1% of all prunable weights and are matched for removal count, layer allocation, and removed-weight energy. Taxonomy construction and response evaluation use disjoint 1,000-example stratified subsets.

Table 17: Threshold robustness on ResNet-50/CIFAR-100. Ratios compare $\mathcal{C}_{\text{high}}$ with $\mathcal{C}_{\text{low}}$ and are geometric means with paired 95% confidence intervals over five seeds. "Seeds > 1" reports the number of seeds for which the corresponding high-to-low ratio exceeds one. The bold row is the prespecified 40/40/20 split used in the main experiments.

q_1	q_2	$\mathcal{I}_{\text{low}}$ / $\mathcal{C}_{\text{low}}$ / $\mathcal{C}_{\text{high}}$	$\|\widehat{G}_M^{\text{high}}\|_2/\|\widehat{G}_M^{\text{low}}\|_2$	Seeds > 1	$\|G_M^{\text{high}}\|_2/\|G_M^{\text{low}}\|_2$	Seeds > 1
0.30	0.50	30/35/35	2.51 [1.16, 5.44]	5/5	1.22 [0.71, 2.08]	4/5
0.30	0.67	30/47/23	3.13 [1.44, 6.76]	5/5	1.45 [1.12, 1.87]	5/5
0.30	0.75	30/53/17	2.45 [1.44, 4.15]	5/5	1.45 [1.04, 2.02]	4/5
0.40	0.50	40/30/30	1.64 [0.94, 2.86]	5/5	1.45 [1.18, 1.77]	5/5
0.40	**0.67**	**40/40/20**	**1.66 [0.79, 3.53]**	**5/5**	**1.42 [1.20, 1.69]**	**5/5**
0.40	0.75	40/45/15	2.12 [0.75, 5.96]	4/5	1.55 [1.10, 2.16]	5/5
0.50	0.50	50/25/25	1.85 [0.93, 3.71]	4/5	1.33 [0.85, 2.07]	4/5
0.50	0.67	50/33/17	1.78 [1.19, 2.67]	5/5	1.33 [1.14, 1.56]	5/5
0.50	0.75	50/38/12	2.42 [1.22, 4.79]	5/5	1.61 [1.29, 2.02]	5/5

The observed response is larger for $\mathcal{C}_{\text{high}}$ than for $\mathcal{C}_{\text{low}}$ in all five seeds for six of the nine threshold settings. In particular, the prespecified 40/40/20 split yields an observed-response ratio of 1.42 with a paired 95% confidence interval of $[1.20, 1.69]$, with the high-to-low ratio exceeding one in all five seeds. The exact-Hessian leading term shows the same seed-wise ordering at this split, although its confidence interval includes one. Thus, the taxonomy signal is not confined to a single threshold choice, while the remaining settings support treating the thresholds as calibrated hyperparameters rather than a universal partition.