SIB: Reparameterization of LLMs for Better Learning-Forgetting under SFT

Albert Catalan-Tatjer [1] **Jonas Geiping** [1]

Abstract

Supervised finetuning (SFT) of pretrained language models trades off the acquisition of new domain capabilities against retention of prior knowledge. Recently, post-training quantization (PTQ) and catastrophic forgetting from finetuning are increasingly seen as a loss geometry problem, where flatness leads to lower degradation. In this work, we adopt a unified view of post-training perturbations. In particular, inspired by PTQ we propose **Scale Invariant Balancing (SIB)** a functionally equivalent reparameterization within the weight-space symmetries that flattens the loss landscape. We extensively characterize the learning-forgetting trade-off for SFT and SIB. Across models and methods, two regimes universally develop. Either baseline SFT performance appears as a gradual trade-off between learning and forgetting, in which case SIB can be applied to approximate Pareto optimality, or, baseline SFT is already not forgetting, in which case SIB does not substantially intervene.

1. Introduction

Modern large language models are the result of a sequence of stages: pre-training, mid-training, supervised finetuning (SFT), and increasingly some form of reinforcement learning from human or verifiable feedback (Yang et al., 2025; Olmo et al., 2026; Apertus et al., 2025). Each stage targets a different capability, and each new release of an open-weight model effectively adds another post-training phase on top. This sequential nature of post-training naturally induces a continual learning setting, wherein each successive stage has to balance learning a new domain while retaining previously learned behaviors. The central problem that arises when continually learning in neural networks however, is catastrophic forgetting (McCloskey & Cohen, 1989), which remains a core challenge in modern post-training.

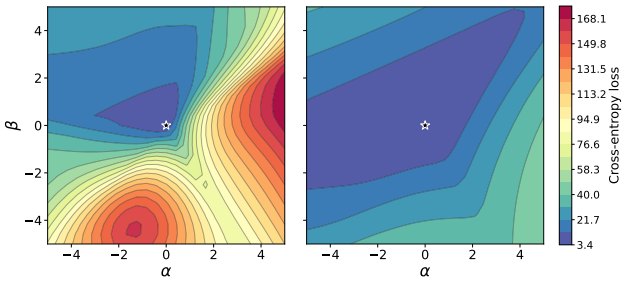

Figure 1. **Effect of Scale Invariant Balancing (SIB) on the Landscape of the loss.** We reparametrize the original model (left) by minimizing the gradient of the loss on a retain dataset within a scale-symmetry group. Which results in stretching or shrinking directions in the landscape according to how sensitive they are to weight perturbations, relative to the scale-symmetry pair.

Recent work points to the local geometry of the previous task loss landscape as a key factor to post-training forgetting. Catalan-Tatjer et al. (2026) to post-training quantization (PTQ) degradation and Watts et al. (2026) link flatness to forgetting under finetuning. From this perspective, a model is the joint product of its data, architecture, and training recipe, where the resulting weight-space geometry shapes how gracefully it can absorb a new optimization stage. Interestingly, this recent connection between PTQ and forgetting under finetuning indicates that some of the techniques and tricks that PTQ researchers have found may generalize to other other post-training perturbations.

In this paper, inspired by Liu et al. (2025) we leverage the weight-space symmetries to minimize the norm of the gradient of the loss on a *retain dataset*. In doing so, we find a functionally equivalent reparameterization of the model that is flatter w.r.t. a retain set. We analyze how a given method reshapes the trade-off between domain adaptation and performance retention, by sweeping the learning rates and studying the Pareto-frontier.

Our contributions are as follows:

1. We propose **SIB**, a functionally equivalent reparameterization of a given model with flatter landscape loss on a chosen retention set.

[1]ELLIS Tübingen, Max Planck for Intelligent Systems, Tübingen AI Center, Germany.. Correspondence to: Catalan-Tatjer <albert.catalan-tatjer@tue.ellis.eu>.

Workshop on Weight-Space Symmetries, held in conjunction with the 43rd International Conference on Machine Learning, Seoul, South Korea. 2026. Copyright 2026 by the author(s).

2. We evaluate extensively our proposed method, and provide detailed analysis of the geometry changes of the loss.

2. Related work

Sequential finetuning of a pre-trained model forces a choice between absorbing a new task and preserving what was learned before, a tension that has been studied in many forms since McCloskey & Cohen (1989) first articulated catastrophic forgetting. We focus on two threads of prior work that bear directly on this trade-off in modern LLMs: the structure of contemporary post-training pipelines, and recent evidence that the severity of forgetting is governed by the local loss geometry around the pre-trained weights.

2.1. Post-training of language models

Modern LLMs are the result of a sequence of stages: pre-training, mid-training, instruction tuning, preference alignment and, increasingly, reasoning-oriented post-training, each targeting a different capability (Yang et al., 2025; Olmo et al., 2026; Apertus et al., 2025). SFT is the principal tool for injecting domain expertise, instruction following, and safety behaviors, and is typically followed by RL stages that further refine the model. As a consequence, post-training is inherently a sequential learning problem in which earlier capabilities should not be inadvertently overwritten. RL stages typically address forgetting through KL penalties between the updated and reference policies and very small learning rates (OLMo Team, 2025; Olmo et al., 2026; Apertus et al., 2025), treating the magnitude of post-training updates as a central lever for retention.

2.2. Forgetting in LLMs

Recent work is concerned with the relationship between pre-training and post-training, with growing evidence that the severity of forgetting in LLMs is tied to the local geometry of the loss of the pre-trained weights. Watts et al. (2026) establish a direct connection between the flatness of the loss landscape and catastrophic forgetting during finetuning, showing that sharpness-aware minimization (Foret et al., 2021) can mitigate it. Closely related, Catalan-Tatjer et al. (2026) show that post-training quantization robustness is governed by the same training dynamics suggesting that **robustness to weight perturbations** is a unifying lens for both quantization and post-training forgetting. Post-training quantization has been successfully exploiting weight-space geometry: Liu et al. (2025) show that, although the network is invariant to certain rotations of its weights, different parameterizations within these symmetries differ substantially in quantization error, an observation we build on to shape sensitivity to SFT updates. Empirically, forgetting has been observed under naive finetuning (Kalajdzievski, 2024), where the most consistent practical lever in current recipes remains conservative learning rates, which have been shown to limit feature drift (Rofin et al., 2026; Lin et al., 2026).

3. Evaluating the Pareto frontier

3.1. Supervised finetuning

Backbones. Generally, open-weights models provide a base model and an instruction-tuned version. We choose instruction-tuned models for these are trained the result of all the post-training stages, where mathematical benchmarks are highly optimized for (Gulati et al., 2025). Thus, a relevant setting for studying forgetting under sequential learning. We work with Qwen2.5-Instruct at 0.5B, 1.5B, 3B, 7B (Qwen Team, 2024), Qwen3 at 0.6B and 4B (Yang et al., 2025). Llama-3.2-Instruct at 1B and 3B, Llama-3.1-Instruct at 8B (Grattafiori & Llama Team, 2024), OLMo-2-Instruct at 1B and 7B (OLMo Team, 2025).

Optimization process. Following common practice (Taori et al., 2023), we implement a standard instruction tuning training pipeline, including prompt loss masking, where the loss is computed only on the response tokens and not on the prompt and standard instruction message formats for each model. All finetuning runs use AdamW (Loshchilov & Hutter, 2019) with weight decay 0.01. We use a cosine learning-rate schedule (Loshchilov & Hutter, 2017) that warms up for the first 20 steps and then decays the peak learning rate to $0.1\times$ its maximum value, with an additional 50 cool-down steps at the end of training. To map out the retention-learning Pareto frontier per model, we exhaustively sweep the peak learning rate and the number of epochs for all models and methods. Unless indicated otherwise, we default to 3 epochs of SFT.

Learning. We conduct our experiments on **ChemL3**, the level-3 chemistry split of Feng et al. (2024), a multiple-choice question answer subset suitable for studying sequetial domain-specific fine-tuning because it is mostly unkown by popular open-weight language models.

Forgetting. To measure retention, we evaluate the models on GSM8K (Cobbe et al., 2021). As mathematical reasoning is learned through substantial effort in pretraining for the open weight models we consider, it represents a sufficiently complex proxy of task retention. Given that GSM8k is also evaluated through verification of model generations, it is further preferable over task retention measurements solely through multiple choice benchmarks, which do not measure whether generative capabilities have been retained. In all of our experiments and baseline models, we observe that GSM8K performance is uncorrelated to ChemL3, providing a clear learning-forgetting trade-off. Therefore, we choose the retain set $\mathcal{D}_R$ to be GSM8K's train split throughout.

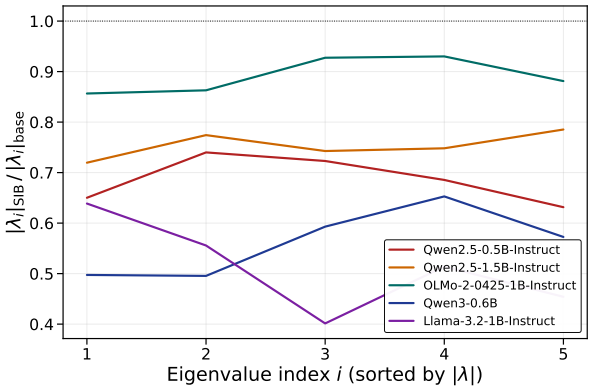

Figure 2. **Top 5 Eigenvalues of the Hessian decrease after SIB** We observe that SIB flattens the top 5 sharper directions of $\mathcal{L}(\mathcal{D}_R)$.

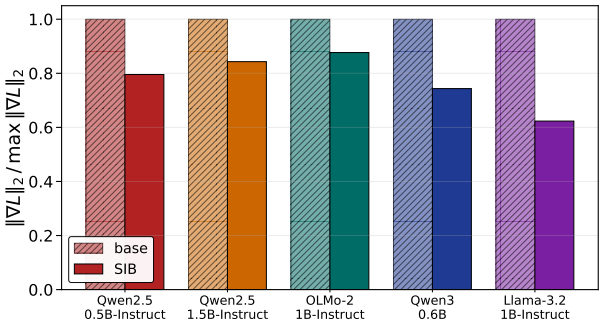

Figure 3. **Norm of the gradient of the loss decreases after SIB** We observe that SIB reparameterization considerably (10-35pp) reduces the norm of the gradient of $\mathcal{L}(\mathcal{D}_R)$.

4. Scale Invariant Balancing (SIB)

In this section we introduce our method, Scale Invariant Balancing (SIB). We start by describing some of the scale symmetries in the backbones that we consider, and then detail how we leverage those to flatten the loss landscape.

4.1. Common scale-symmetries in open-weights language models.

Weight-space symmetries have been used for improving model merging (Ainsworth et al., 2023), accelerated optimization (Zhao et al., 2022), and reduced quantization degradation (Liu et al., 2025). We focus on *scale symmetries*: transformations $W \mapsto W/\alpha$, $W' \mapsto W' \operatorname{diag}(\alpha)$ on a pair of adjacent weight tensors, with $\alpha \in \mathbb{R}^d_{>0}$, that leave the network's input–output map exactly unchanged. In modern transformers these arise at four sites:

- **V/O (per head):** $O \operatorname{diag}(\alpha) \operatorname{diag}(\alpha)^{-1} V = OV$, so we can rescale rows of V and columns of O within each head.
- **SwiGLU (Shazeer, 2020) (per neuron):** scaling rows of W_{up} by α and columns of W_{down} by $1/\alpha$ leaves $W_{\text{down}}\big(\sigma(W_{\text{gate}}x) \odot (W_{\text{up}}x)\big)$ unchanged, since the α factors out of the elementwise product.
- **RMSNorm / linear:** $\operatorname{RMSNorm}(x)_i = \gamma_i\, x_i / \|x\|_{\text{RMS}}$ is degree 1 in γ, so γ absorbs into the columns of any downstream linear sharing that norm: $\gamma \mapsto \gamma/\alpha$, $W_{\text{down}} \mapsto W_{\text{down}} \operatorname{diag}(\alpha)$.
- **Q/K with RoPE (per kv-head):** attention logits depend on $Q^\top K$, so $Q \mapsto \alpha Q$, $K \mapsto K/\alpha$ is invariant. RoPE's block-diagonal rotation forces α to be a single scalar per kv-head rather than per channel.

We restrict our analysis to V/O and SwiGLU for two reasons. First, they are present in all architectures we study, while the RMSNorm symmetry is broken in Gemma 3 (Gemma Team, 2025) and OLMo 2, which parameterize the gain as $(1+\gamma)$; the additive constant destroys the homogeneity that the symmetry requires. Second, the two tensors in each V/O or SwiGLU pair have comparable parameter counts, whereas γ is a vector and its consumer is a matrix, which complicates questions of stability and normalization across the pair.

4.2. Minimizing the norm of the gradient

Now, let $(W_{\text{div}}, W_{\text{mul}})$ be a scale-symmetric pair of tensors, and $\mathcal{D}_R$ a retention dataset whose loss landscape we want to flatten, we define the follow objective to minimize the gradient of $\mathcal{L}(\mathcal{D}_R)$,

$$\|g_{\text{div}}/\alpha\|^2 + \|\alpha\, g_{\text{mul}}\|^2, \quad \text{where} \quad g_{\text{div}} = \nabla_{W_{\text{div}}} \mathcal{L}(\mathcal{D}_R), \quad g_{\text{mul}} = \nabla_{W_{\text{mul}}} \mathcal{L}(\mathcal{D}_R). \tag{1}$$

which admits the closed-form solution

$$\alpha^* = \left(\|g_{\text{div}}\|^2 / \|g_{\text{mul}}\|^2\right)^{1/4}, \qquad \alpha^* \in [1/R,\ R], \tag{2}$$

clamped to R (we use R=100 unless stated otherwise). The solution at the "balance" between $\|g_{\text{mul}}\|^2$ and $\|g_{\text{div}}\|^2$ gives name to this method.

Conceptually, this stretches the directions of higher weight sensitivity (as indicated by $\nabla_W \mathcal{L}(\mathcal{D}_R)$) and compresses the more robust, thereby protecting $\mathcal{L}(\mathcal{D}_R)$ from SFT weight updates. Figure 1 shows the loss landscape of a model (left) and its reparameterization following 1, it can be seen that the top eigenvalues of the hessian of $\mathcal{L}(\mathcal{D}_R)$ are much lower, indicating that the reparameterization has effectively flattened the landscape.

We perform SIB once the language model, and then proceed with vanilla SFT. Find the step-by-step algorithm of SFT pipeline with SIB described in Algorithm 1.

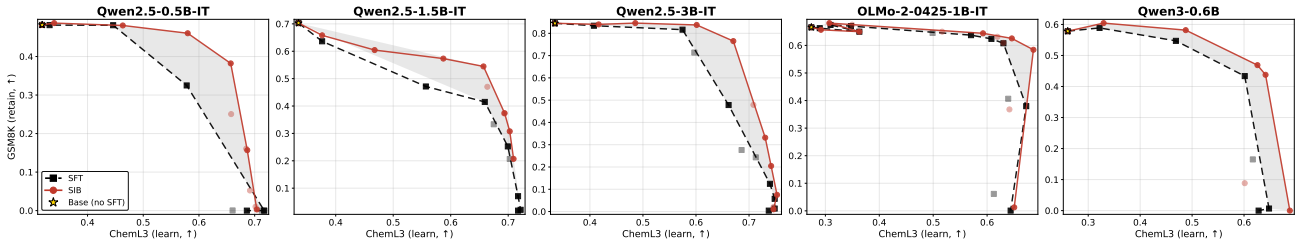

Figure 4. **Learning-forgetting Pareto comparison**. In the settings where there is gradual balance between learning and forgetting (e.g., Qwen2.5 and OLMo2 settings), SIB parameterizations dominate the Pareto frontier, but notably only move it upward, 'sharpening' the cliff, i.e. the sudden phase transition from complete retention to complete forgetting.

5. Results and discussion

As introduced in Section 3, sequential learning is inherently multi-objective optimization, where the quality of a solution is determined by how it balances different objectives, a balance that is regulated by the magnitude of SFT updates. In this section, we provide quantitative analysis of the impact of SIB reparameterization, and we discuss the Pareto frontier in different open-weights models.

5.1. Geometry

We begin by analyzing the norm of the gradient of $\mathcal{L}(\mathcal{D}_R)$ in Figure 3. We show $||\nabla\mathcal{L}(\mathcal{D}_R)||$ max-normalized, so each model has its maximum value at 1. Figure 3 demonstrates that the reparameterization objective is successful at minimizing $||\nabla\mathcal{L}(\mathcal{D}_R)||$, where SIB achieves between $\tilde{1}$0pp and 35pp lower values than the base model. Similarly, we study the 5 largest-by-magnitude eigenvalues of the hessian of $\mathcal{L}(\mathcal{D}_R)$ in Figure 2. We report the ratio between the i-th base eigenvalue and its SIB counterpart, which is always < 1. Although surprising, whether the beneficial geometry materializes into better learning-forgetting Pareto trajectories remains to be seen.

5.2. Pareto frontier

We continue by seeing in Figure 4 that SIB (red line) dominates SFT (dashed black line) the Pareto frontier in six cherry-picked settings, wherein at the same domain adaptation performance, SIB parameterizations retain considerably more (up to 35pp for Qwen2.5-3B) GSM8K accuracy. However, the study of more model families and sizes reveals a more nuanced picture. We show the entire grid of models under study in Figure 5, where we observe two interesting regimes that emerge during SFT. For the settings (Qwen2.5-0.5B, -1.5B, -3B, -7B and OLMo2-1B, Qwen3-0.6B) where learning and forgetting are gradually traded-off, SIB parameterization approach the optimal corner at which there is reduced forgetting and maximal learning. Surprisingly, there is another scenario (Llama3.2-1B, Llama3.2-3B, Llama3.1-8B) where SFT already approximates the optimal corner. In this case, SIB fails to offer better learning, indicating that these modulate forgetting rather than learning.

6. Conclusion

A unified perspective of post-training perturbations, opens up a stream of ideas from post-training quantization, where research has bloomed achieving extraordinary success, to catastrophic forgetting. In this work, we borrow some ideas from Liu et al. (2025), to leverage weight-space symmetries to find a *better* parameterization of the base model. Flatness of the loss landscape has been linked to PTQ degradation and SFT forgetting (Catalan-Tatjer et al., 2026; Watts et al., 2026), which motivates it as our optimization target. In fact, our method SIB, stretches and shrinks scale-symmetric pairs to considerably flatten the sharpest directions of the landscape and overall the norm of the gradient of the loss.

Finally, we uncover two regimes that emerge during SFT: one in which learning and forgetting are gradually traded-off, where SIB Pareto-dominates vanilla SFT, and a second regime where SFT already approaches the optimal corner, and SIB does not substantially intervene, indicating that improvements in learnability remain elusive.

References

Ainsworth, S. K., Hayase, J., and Srinivasa, S. Git rebasin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2209.04836.

Apertus, P., Hernández-Cano, A., Hägele, A., Huang, A. H., Romanou, A., Solergibert, A.-J., Pasztor, B., Messmer, B., Garbaya, D., Ďurech, E. F., and et al., I. H. Apertus: Democratizing open and compliant llms for global language environments, 2025. URL https://arxiv.org/abs/2509.14233.

Catalan-Tatjer, A., Ajroldi, N., and Geiping, J. Training dynamics impact post-training quantization robustness, 2026. URL https://arxiv.org/abs/2510.06213.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Feng, K., Ding, K., Wang, W., Zhuang, X., Wang, Z., Qin, M., Zhao, Y., Yao, J., Zhang, Q., and Chen, H. Sciknoweval: Evaluating multi-level scientific knowledge of large language models, 2024.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization, 2021. URL https://arxiv.org/abs/2010.01412.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Grattafiori, A. and Llama Team. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Gulati, A., Miranda, B., Chen, E., Xia, E., Fronsdal, K., de Moraes Dumont, B., and Koyejo, S. Putnam-AXIOM: A functional & static benchmark for measuring higher level mathematical reasoning in LLMs. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=kqj2Cn3Sxr.

Kalajdzievski, D. Scaling laws for forgetting when fine-tuning large language models, 2024. URL https://arxiv.org/abs/2401.05605.

Lin, J., Wang, Z., Qian, K., Wang, T., Srinivasan, A., Zeng, H., Jiao, R., Zhou, X., Gesi, J., Wang, D., Guo, Y., Zhong, K., Zhang, W., Sanghavi, S., Chen, C., Yun, H., and Li, L. Sft doesn't always hurt general capabilities: Revisiting domain-specific fine-tuning in llms, 2026. URL https://arxiv.org/abs/2509.20758.

Liu, Z., Zhao, C., Fedorov, I., Soran, B., Choudhary, D., Krishnamoorthi, R., Chandra, V., Tian, Y., and Blankevoort, T. Spinquant: Llm quantization with learned rotations. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2405.16406.

Loshchilov, I. and Hutter, F. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1608.03983.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1711.05101.

McCloskey, M. and Cohen, N. J. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of learning and motivation*, 24:109–165, 1989.

Olmo, T., :, Ettinger, A., Bertsch, A., Kuehl, B., Graham, D., Heineman, D., Groeneveld, D., Brahman, F., Timbers, F., and et al, H. I. Olmo 3, 2026. URL https://arxiv.org/abs/2512.13961.

OLMo Team. 2 OLMo 2 furious. *arXiv preprint arXiv:2501.00656*, 2025. URL https://arxiv.org/abs/2501.00656.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Rofin, M., Varre, A., and Flammarion, N. (how) learning rates regulate catastrophic overtraining, 2026. URL https://arxiv.org/abs/2604.13627.

Shazeer, N. Glu variants improve transformer, 2020. URL https://arxiv.org/abs/2002.05202.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Watts, I., Li, C., Goyal, S., Springer, J. M., and Raghunathan, A. Sharpness-aware pretraining mitigates catastrophic forgetting. In *ICLR 2026 Workshop on Geometry-grounded Representation Learning and Generative Modeling*, 2026. URL https://openreview.net/forum?id=B2qTJi5s0M.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., and et al., B. Z. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Zhao, B., Dehmamy, N., Walters, R., and Yu, R. Symmetry teleportation for accelerated optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2205.10637.

A. Experimental set-up

Backbones. Generally, open-weights models provide a base model and an instruction-tuned version. We choose instruction-tuned models for two principal reasons. Firstly, these models are trained for more post-training stages, with more domain-specific capabilities already acquired, making them relevant for studying forgetting under sequential learning. Secondly, open-weight models are highly optimized for performance on math benchmarks (Gulati et al., 2025), as a consequence, the forgetting signal will be less noisy. We work with Qwen2.5-Instruct at 0.5B, 1.5B, 3B, 7B and 14B (Qwen Team, 2024), Qwen3 at 0.6B and 4B (Yang et al., 2025), Llama-3.2-Instruct at 1B and 3B, Llama-3.1-Instruct at 8B (Grattafiori & Llama Team, 2024), OLMo-2-Instruct at 1B and 7B (OLMo Team, 2025).

B. Scale symmetries

For each of the following weight pairs, there exists a positive vector α such that replacing $W \mapsto W/\alpha$ on one side and $W \mapsto W\alpha$ on the other leaves the network's input–output map exactly unchanged:

- **V/O**: V rows $\leftrightarrow$ O columns (per head).
- **SwiGLU (Shazeer, 2020)**: $(W_{\text{up}}, W_{\text{gate}})$ rows $\leftrightarrow$ W_{down} columns (per neuron).
- **RMSNorm / linear**: γ entries $\leftrightarrow$ columns of downstream linear layers sharing that norm.
- **Q/K** (with RoPE): per-kv-head scale, constrained to be shared within each kv-head to preserve RoPE.

We limit the symmetries to V/O pair and the SwiGLU MLPs.

C. SIB details

C.1. SIB equivalent to per-channel learning rate scales

Proposition C.1. *Let $\hat{W} := \alpha W$ and consider Adam updates applied to $\hat{W}$ with learning rate η. Then, up to the additive stabilizer ϵ, the induced dynamics on W are equivalent to Adam updates with learning rate η/α applied directly to W.*

Proof. The Adam update on the scaled parameters is given by

$$\hat{W}^{t+1} = \hat{W}^t - \eta \frac{\hat{m}^t}{\sqrt{\hat{v}^t} + \epsilon}, \tag{3}$$

where $\hat{m}^t$ and $\hat{v}^t$ denote the bias-corrected first and second moment estimates of

$$\hat{g}^t := \nabla_{\hat{W}} \mathcal{L}(\hat{W}^t).$$

Since $\hat{W} = \alpha W$ and $\mathcal{L}(\hat{W}) = \mathcal{L}(\alpha W)$, the chain rule yields

$$\hat{g}^t = \frac{1}{\alpha} g^t, \qquad g^t := \nabla_W \mathcal{L}(W^t).$$

By linearity of the Adam moment updates, this scaling propagates to the first and second moments:

$$\hat{m}^t = \frac{1}{\alpha} m^t, \qquad \hat{v}^t = \frac{1}{\alpha^2} v^t.$$

Substituting into the normalized update term gives

$$\frac{\hat{m}^t}{\sqrt{\hat{v}^t} + \epsilon} = \frac{(1/\alpha) m^t}{(1/\alpha)\sqrt{v^t} + \epsilon} \tag{4}$$

$$= \frac{m^t}{\sqrt{v^t} + \alpha\epsilon}. \tag{5}$$

Algorithm 1 SFT with Scale-Invariant Balancing (SIB), V/O + MLP symmetry pairs

Require: Pre-trained weights $\theta^\star$; fine-tuning data $\mathcal{D}_{\mathrm{FT}}$; retain (PT) data $\mathcal{D}_{\mathrm{PT}}$; the set of scale-invariant pairs we optimize across all architectures, $\mathcal{P} = \mathcal{P}_{\mathrm{V/O}} \cup \mathcal{P}_{\mathrm{MLP}}$, with

- $\mathcal{P}_{\mathrm{V/O}} = \{(W_V^{(\ell)},\, W_O^{(\ell)})\}_{\ell=1}^{L}$ (per-layer value $\rightarrow$ output projection pair);
- $\mathcal{P}_{\mathrm{MLP}} = \{(W_{\mathrm{up}}^{(\ell)},\, W_{\mathrm{down}}^{(\ell)}),\, (W_{\mathrm{gate}}^{(\ell)},\, W_{\mathrm{down}}^{(\ell)})\}_{\ell=1}^{L}$ (gate/up rows paired with the down-projection's columns);

clamp ratio R; FT optimizer $\mathcal{O}$ (AdamW), base LR η, epochs E.

$\theta \leftarrow \theta^\star$

Phase 1 — Geometry teleportation (one-shot, gradient pass on PT data)

for $(W_{\mathrm{div}}, W_{\mathrm{mul}}) \in \mathcal{P}$ **do**

  Compute gradient norms $g_{\mathrm{div}} = \left\|\nabla_{W_{\mathrm{div}}} \mathcal{L}_{\mathrm{PT}}(\theta)\right\|$, $g_{\mathrm{mul}} = \left\|\nabla_{W_{\mathrm{mul}}} \mathcal{L}_{\mathrm{PT}}(\theta)\right\|$ from a single PT minibatch

  $\alpha \leftarrow \mathrm{clip}\Big(\sqrt{g_{\mathrm{mul}}/g_{\mathrm{div}}},\ 1/R,\ R\Big)$ {equalises gradient magnitudes across the pair}

  $W_{\mathrm{div}} \leftarrow W_{\mathrm{div}}/\alpha, \quad W_{\mathrm{mul}} \leftarrow W_{\mathrm{mul}} \cdot \alpha$ {output-preserving reparametrisation}

end for

Phase 2 — Standard SFT from the rebalanced weights

for epoch $= 1 \ldots E$ **do**

  for minibatch $b \sim \mathcal{D}_{\mathrm{FT}}$ **do**

    $\theta \leftarrow \mathcal{O}(\theta,\, \nabla_\theta \mathcal{L}_{\mathrm{FT}}(\theta; b),\, \eta)$

  end for

end for

return θ =0

Hence,

$$\hat{W}^{t+1} = \hat{W}^t - \eta \frac{m^t}{\sqrt{v^t} + \alpha\epsilon}. \tag{6}$$

Dividing by α and defining $W^t := \hat{W}^t/\alpha$ yields

$$W^{t+1} = W^t - \frac{\eta}{\alpha} \frac{m^t}{\sqrt{v^t} + \alpha\epsilon}. \tag{7}$$

In regimes where $\alpha\epsilon \ll \sqrt{v^t}$, the stabilizer term can be neglected, yielding an effective learning-rate rescaling by $1/\alpha$. □

C.2. Algorithm

D. More results

D.1. Pareto frontiers of all the models

In Figure 5 we show the comparison of the Pareto frontiers for all of the models. Here we can see that there are two different regimes in the multi task learning setting. Firstly, there are models where learning and forgetting are traded off gradually, in these cases, such as Qwen2.5-0.5-IT, the Pareto trajectory of SFT showcases a curve or trapezoidal figure, where the Pareto optimality (top-right corner) is not achieved. In these situations, Qwen2.5-{0.5B, 1.5B, 3B}, Qwen3-0.6B, OLMo2-1B, we see that SIB dominates the Pareto frontier. Moreover, we identify aother regime, where Pareto optimality is achieved with vanila SFT. In these cases, Llama3.2-{1B, 3B} where there is maximal learning with minimal forgetting, SIB does not provide better learning, resulting in two lines that mostly track each other.

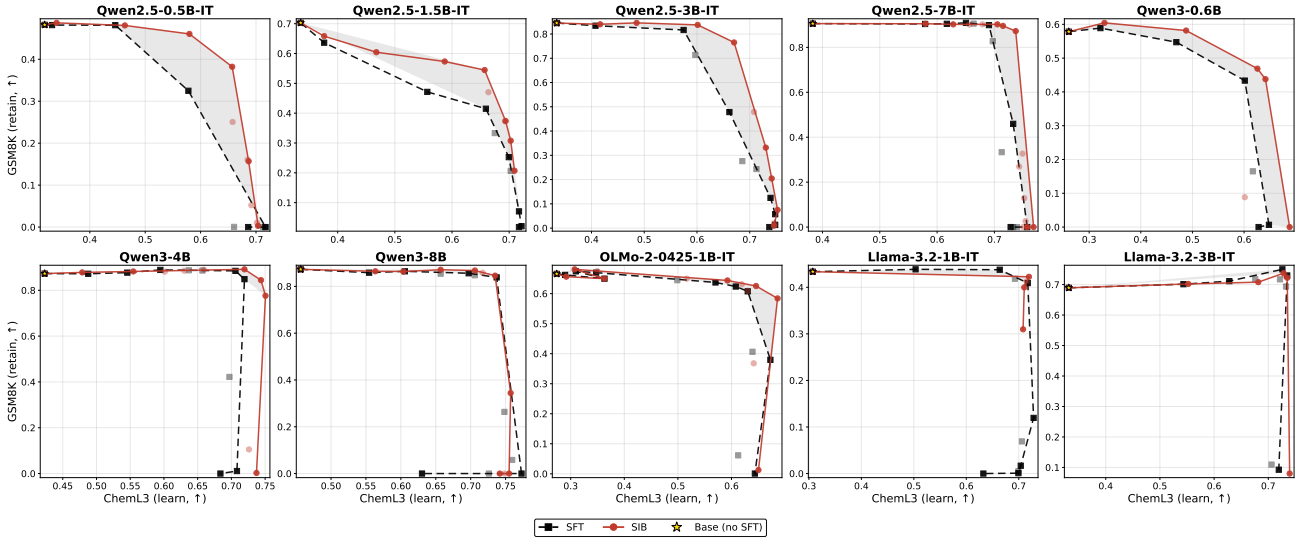

Figure 5. **Learning-forgetting Pareto comparison for the complete model zoo**. In the settings where there is gradual balance between learning and forgetting (e.g., Qwen2.5 and OLMo2 settings), SIB parameterizations dominate the Pareto frontier, but notably only move it upward, 'sharpening' the cliff, i.e. the sudden phase transition from complete retention to complete forgetting. When SFT already forms a cliff (e.g., Llama3.2 models), SIB barely perform better than SFT, indicating its role as performance retention rather than learning enhancement.