A loss curvature account of fine-tuning fragility

Ivaylo Dimitrov[* 1 2] **Leo Karoubi**[* 1] **Sunny Howard**[* 1 3] **Dmitrii Krasheninnikov**

Abstract

Fine-tuning on narrow distributions often produces fragile changes that are easily reversed by further training, with implications for the durability of safety fine-tuning. Mixing pre-training data into fine-tuning is a known mitigation, but why varying the proportion of fine-tuning data (which we term concentration) modulates forgetting is poorly understood. During a reversion phase (subsequent training on pre-training data after fine-tuning), we decompose the per-step change in fine-tune loss into its first- and second-order Taylor terms. We then track how each varies with concentration. In experiments on LLMs (Pythia-70M), we find that the second-order (curvature) term grows in importance with concentration. This curvature is much larger along the reversion update direction than along a random direction at every concentration, with the directional curvature itself increasing with concentration. Curvature therefore contributes to forgetting even when fine-tune and pre-train gradients are not in conflict, and its relative share grows with concentration, providing empirical support for recent theoretical accounts of curvature-driven forgetting.

1. Introduction

Safety fine-tuning is easily undone by subsequent training, even on seemingly unrelated tasks (Qi et al., 2024; Betley et al., 2026), and narrow fine-tuning can leave the underlying model capabilities largely intact while layering a small, removable "wrapper" on top (Panigrahi et al., 2023; Jain et al., 2024; Minder et al., 2026). If alignment can be stripped this easily, then understanding *why* fine-tuning is shallow is a prerequisite for making it robust (Wang et al., 2024). Mixing pre-training data into fine-tuning is a well-established mitigation (Ibrahim et al., 2024; Gu et al., 2024; Béthune et al., 2025; Baek et al., 2026; Kotha & Liang, 2026), and the same intuition motivates interleaving alignment and task data during safety training (Korbak et al., 2023). Yet prior work typically treats the concentration of new-task data as a hyperparameter – the mechanism by which concentration modulates forgetting remains underexplored.

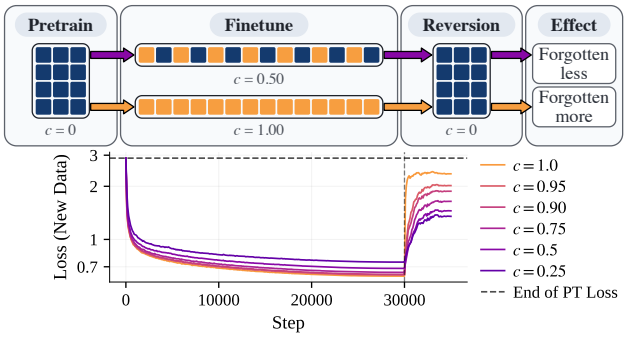


Figure 1. **Top:** the three-phase protocol. Every run shares the same pre-training, then fine-tunes on a mixture of c new-task data and $(1 - c)$ pre-training data, then reverts to pure pre-training. **Bottom:** fine-tuning and reversion loss on a chemistry dataset (SMILES/ChEMBL) under a step-matched budget. Higher concentration leads to faster forgetting.

To understand forgetting, we study a reversion phase in which a model fine-tuned on new data is further trained on pre-training data alone. Taylor-expanding the new-task loss under each reversion step decomposes the per-step change into a *first-order* term measuring the alignment between the new-task gradient and the reversion weight update, a *second-order* term measuring the curvature of the new-task loss landscape along the reversion update direction, and higher-order terms (§2.3). Existing mechanistic accounts of forgetting primarily focus on the first-order gradient conflict (Liu et al., 2021; Yang et al., 2026), although Springer et al. (2026) argue theoretically that curvature alone can drive rapid forgetting even when gradients are not in conflict, echoing observations of high-curvature "safety basins" (Peng et al., 2024) and findings that flatter fine-tuning minima correlate with reduced forgetting (Li et al., 2024). Concurrent work by Kalra et al. (2026) measures the curvature of the new task along the direction of the pre-training task and uses it to identify favourable concentration rates. A mechanistic account connecting the gradient and curvature

[*]Equal contribution [1]LASR Labs [2]SUMMIT Intelligent Systems Research Group, Sofia University "St. Kliment Ohridski" [3]Munich Center for Machine Learning. Correspondence to: Ivaylo Dimitrov <first name dot last name at protonmail dot com>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

findings to forgetting dynamics, remains underdeveloped.

We use concentration as an instrument to bring these threads together. Our experimental setup is shown in Fig. 1: we fine-tune on a concentration c of new data and $(1-c)$ of pre-training data, then revert to pure pre-training and measure how quickly the new task is forgotten (Jain et al., 2024), while analyzing the loss landscape.

Our contributions are as follows: **(1)** we systematically establish the correlation of forgetting rate with concentration; **(2)** we track the first- and second-order terms of the decomposition across the forgetting period, finding that both contribute, with the second-order term becoming increasingly significant at larger concentrations, and therefore at faster forgetting rates; and **(3)** we show that the curvature term is direction-dependent: the new task loss landscape is much sharper along the reversion update direction than along a random parameter direction at all concentrations, and directional curvature grows with concentration.

2. Methods

2.1. Three-phase training protocol

Every experiment follows the same three phases: (1) *Pre-train* a model on $\mathcal{D}_{\text{pre}}$ until convergence (or take an existing pre-trained model), (2) *Fine-tune* on a mixture of $\mathcal{D}_{\text{pre}}$ and a new distribution $\mathcal{D}_{\text{new}}$, (3) *Reversion* to training on $\mathcal{D}_{\text{pre}}$ alone. A concentration parameter $c \in [0, 1]$ controls the mixture during fine-tuning and describes the expected fraction of new examples per fine-tuning step. At c=1 all examples are new, and lower c interleaves more pre-training data across the same number of fine-tuning steps. We measure how quickly the model forgets the fine-tuned capability (more on forgetting speed and its definition in §B.5 and §D). We compare two budget conventions: *step-matched* fixes the number of fine-tuning steps across c, so lower c sees fewer new examples in total. *Volume-matched* scales steps as $1/c$ to hold the total number of new examples fixed, so the number of steps is higher for lower c.

2.2. Experimental settings

We run all experiments on Pythia (Biderman et al., 2023), which is an LLM pre-trained on the Pile, following the three-phase protocol and sweeping across concentrations c. We fine-tune on chemistry (SMILES/ChEMBL; molecular structure strings), music (ABC musical notation; Irish folk tunes), and biomedical text (PubMed; scientific paper abstracts), before performing the reversion phase on the Pile. Full experimental details are in §B, §E, and §F.

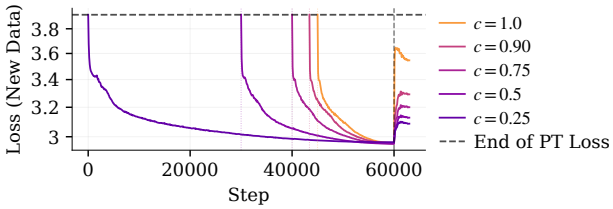


Figure 2. Validation loss during fine-tuning and reversion under a volume-matched budget. Vertical dashed line marks the fine-tuning→reversion boundary. Lower concentration produces monotonically smaller loss rebounds. Biomedical (PubMed) data.

2.3. Mechanistic metrics

During reversion, we perform per-step pre-training updates using the AdamW optimizer: at step k, $\Delta\theta_k = -\eta_k \cdot \text{AdamW}(g_{\text{pre},k})$, where η_k is the learning rate, θ are the model's parameters, $g_{\text{pre},k} = \nabla_\theta L_{\text{pre}}(\theta_k)$ is the pre-training gradient, and θ_0 is the fine-tuned checkpoint. The Taylor expansion below is evaluated per-step: $\Delta\theta$ denotes the single-step update, and T_1, T_2 are measured at each θ_k independently, not cumulatively. We are interested in the corresponding change to the new-task loss, which we decompose using a Taylor expansion. Writing $g_{\text{new}} = \nabla_\theta L_{\text{new}}(\theta)$ and $H_{\text{new}} = \nabla^2_\theta L_{\text{new}}(\theta)$:

$$\Delta L_{\text{new}} = \underbrace{g_{\text{new}}^\top \Delta\theta}_{T_1} + \underbrace{\frac{1}{2}\Delta\theta^\top H_{\text{new}}\,\Delta\theta}_{T_2} + \mathcal{O}(\|\Delta\theta\|^3)\,. \quad (1)$$

Our goal is to understand how each term of Eq. 1 varies with concentration. Noting both terms' dependence on $\Delta\theta$, we measure its alignment with the local structure of L_{new}. For T_1 we measure the cosine similarity, $\cos(g_{\text{new}}, \Delta\theta)$, the alignment with the new-task gradient. For T_2 we measure the curvature of L_{new} along the direction of the weight change, $\widehat{\Delta\theta} := \Delta\theta/\|\Delta\theta\|$ (the unit-norm update direction), reporting the directional curvature $\lambda_{\text{dir}} = \widehat{\Delta\theta}^\top H_{\text{new}}\,\widehat{\Delta\theta}$ (via a finite-difference Hessian-vector product, §C.3); we also find the curvature along the average direction, $\lambda_{\text{avg}} = \text{Tr}(H_{\text{new}})/d$, where d is the number of trainable parameters, as a reference scale. λ_{avg} averages over all (predominantly low-curvature) directions, so it is a conservative reference scale (see §5). These alignments determine whether a reversion step decreases or increases the new task loss, and we measure all quantities by sampling separate batches from $\mathcal{D}_{\text{new}}$ and $\mathcal{D}_{\text{pre}}$. We also consider the magnitude of the weight update, which determines the relative sizes of the two terms as T_1 is linear and T_2 is quadratic in $\|\Delta\theta\|$.

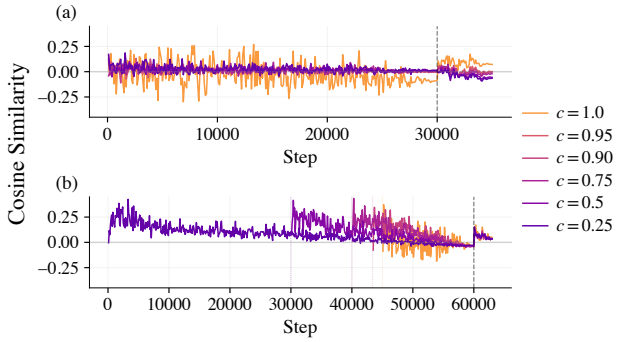


Figure 3. Cosine similarity between the weight update and the gradient of the new task, $\cos(g_{\text{new}}, \Delta\theta)$, during fine-tuning and reversion. (a) is on the SMILES/ChEMBL dataset, matching Fig. 1. (b) is on PubMed data, matching Fig. 2.

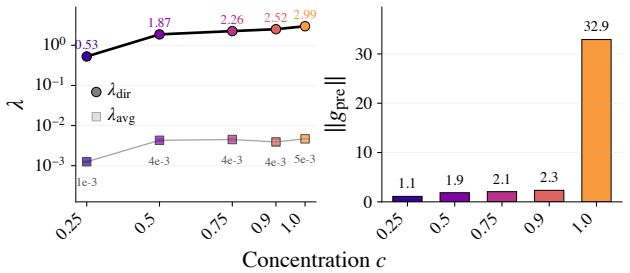


Figure 4. (a) The directional and average curvature, λ_{dir} and λ_{avg}, at the start of the reversion phase (PubMed data). We find that across all concentrations, $\lambda_{\text{dir}} \gg \lambda_{\text{avg}}$, showing that the new-task loss landscape is more sharply curved along the $\Delta\theta$ direction than along a random direction. This plot is also shown in linear scale with error bars in Fig. 14 (§D). (b) The pre-train gradient norm at the end of fine-tuning.

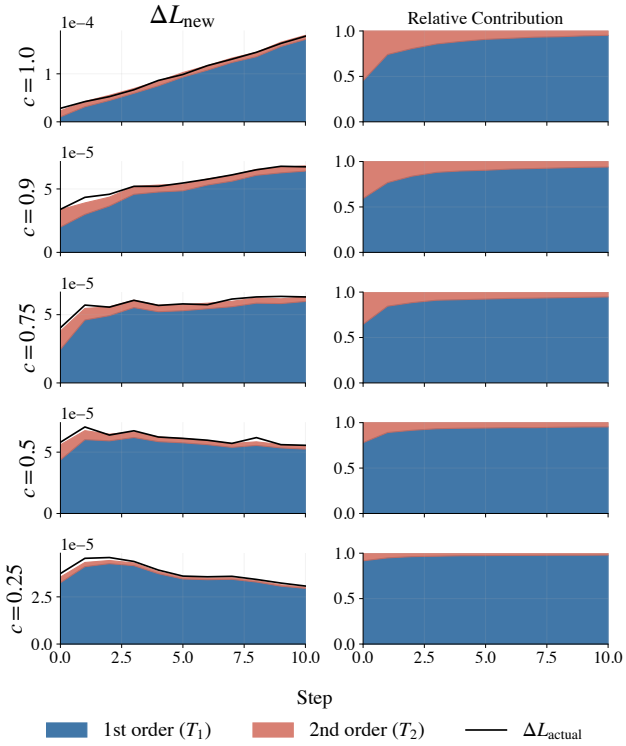


Figure 5. Per-step first-order (T_1, gradient-alignment) and second-order (T_2, curvature) Taylor terms during the first $N{=}10$ reversion steps, for each concentration c (one row per c, PubMed data, mean of six seeds). Each panel's y-axis is in units of loss-change per step (*not* shared across rows). The black line shows the actual ΔL_{new}; its deviation from $T_1{+}T_2$ reflects higher-order terms. The learning rate was fixed to its average over these steps rather than following the warmup schedule (see §C.3). Higher concentrations lead to higher absolute magnitudes and a larger relative share of the curvature term.

3. Results

3.1. Concentration drives forgetting speed

We find that forgetting speed increases monotonically with concentration across every setup and domain we test: step-matched (Pythia-70M on chemistry Fig. 1, music Fig. 9, PubMed Table 4) and volume-matched (Pythia-1B on PubMed Fig. 2, chemistry Fig. 8). Further, mixing even a small fraction of pre-training data during fine-tuning incurs negligible new-task loss, while significantly reducing the pre-training distribution degradation (Fig. 10).

3.2. Mechanistic analysis

We examine the effects of the first and second orders of the Taylor expansion of the fine-tune loss, T_1 and T_2 respectively (Eq. 1), across concentrations.

First-order alignment. Consistent with Yang et al. (2026), $\cos(g_{\text{new}}, \Delta\theta)$ becomes slightly negative during fine-tuning on PubMed (Fig. 3b). Under reversion, where the update is pure pre-training, this opposition drives T_1 to increase the new-task loss. However, on other datasets, such as SMILES, the cosine remains near zero (Fig. 3a) for most concentrations, while performance still degrades. Aggregate gradient conflict is therefore not a sufficient explanation for forgetting in this setting, motivating our analysis of the second-order term.

Second-order alignment. Analyzing the curvature at the end of fine-tuning, we find that at every concentration, $\lambda_{\text{dir}} \gg \lambda_{\text{avg}}$ (Fig. 4a): the new-task loss landscape is sharp along the direction of $\Delta\theta$ relative to a random direction. Both λ_{dir} and λ_{avg} increase with c. This directional curvature persists even when first-order alignment vanishes, indicating that the Hessian can amplify updates along $\Delta\theta$ even when g_{new} and $\Delta\theta$ are near-orthogonal, providing empirical support for Springer et al. (2026).

Magnitude. The relative contributions of T_1 and T_2 also depend on the magnitude of the update, since T_1 is linear and T_2 quadratic in $\|\Delta\theta\|$. The pre-train gradient norm

$\|g_{\text{pre}}\|$ at the end of fine-tuning, a measure of pre-training drift, grows monotonically with c and is especially large at c=1 (Fig. 4b). This confirms that higher-concentration fine-tuning leaves the model further from a pre-training optimum, consistent with mixing even a small amount of pre-training data reducing forgetting - roughly $4\times$ smaller relative degradation at c=0.25 vs. c=1 on chemistry (Table 4). $\|g_{\text{pre}}\|$ is step-like in c, markedly larger only at c=1, whereas the T_2 share (Fig. 5) varies smoothly. Under AdamW the early-step $\|\Delta\theta\|$ is largely decoupled from $\|g_{\text{pre}}\|$ (§C.4, Fig. 6), so the smooth growth of the T_2 share is driven by the smoothly increasing directional curvature λ_{dir} (Fig. 4a) rather than by $\|g_{\text{pre}}\|$, which we therefore treat as a qualitative drift indicator. The update norms still inherit the monotonic c-ordering (Fig. 6), so larger c produces both sharper directional curvature and larger steps along it. The SGD/Adam coupling and a check that clipping was never triggered during reversion are in §C.4.

4. Potential mechanism

Both T_1 (the slope term) and T_2 (the curvature term) contribute to forgetting during reversion. As c grows, T_2's share of the per-step change grows and T_1's share shrinks. Both terms grow in absolute terms, but T_2 grows faster, increasing T_2's relative share and its indirect amplification of T_1 through gradient–Hessian coupling. In our results, T_1 remains the larger *direct* contributor in absolute terms (Fig. 5), while T_2 is not the dominant term.

T_2 grows because the reversion step moves in a direction where the fine-tune landscape is unusually sharp ($\lambda_{\text{dir}} \gg \lambda_{\text{avg}}$; Fig. 4a), and this directional curvature increases with c. Either (a) high-c fine-tuning lands the model in intrinsically sharper regions (the landscape is uniformly more curved in every direction there), or (b) the reversion direction $\widehat{\Delta\theta}$ progressively aligns with the sharpest directions of H_{new} (the landscape bends precisely where we step), or both. A possible explanation for (a) is that at lower c the optimiser must simultaneously reduce L_{new} and L_{pre}, and such joint solutions tend to sit in flatter minima, where the loss barely rises as parameters move away from the minimum. Flatter minima correlate with reduced forgetting (Li et al., 2024).

T_1's relative share shrinks because at c=1 the optimiser sees only L_{new} and converges close to a fine-tune minimum θ^*, where $g_{\text{new}} \approx 0$ makes the ratio $|T_1|/\sqrt{T_2}$ small by construction. At lower c the competing pre-training objective leaves fine-tuning at a displaced point θ^d where g_{new} is non-zero and T_1 takes a larger share. Linearising g_{new} around θ^* and applying Cauchy–Schwarz (§C.2) bounds T_1 by the displacement,

$$|T_1| \leq \sqrt{2\,T_2}\;\|\theta^d - \theta^*\|_{H^*}, \tag{2}$$

so at fixed T_2, T_1 is large only if fine-tuning ended far from θ^* in this Hessian-weighted sense: high c keeps the displacement (and $|T_1|/\sqrt{T_2}$) small, while low c loosens the bound. We use $\|g_{\text{new}}\|$ as a proxy for the displacement; at sufficiently large batch size it decreases with c (Fig. 15), consistent with this account.

5. Limitations and future work

Our mechanistic results are limited to Pythia-70M (forgetting-speed results extend to Pythia-1B; Figs. 2 and 8), and we cannot rule out that the mechanistic picture attenuates or changes character at larger scale. The empirical picture is also not uniformly clean across hyperparameter regimes: the speed of forgetting depends on learning rate (Kim et al., 2025; Ibrahim et al., 2024) and optimiser (Panigrahi et al., 2023), and absolute numbers should not be compared across regimes (§D.1). The proposed mechanism (§4) is validated only over the N=10-step horizon where the quadratic approximation is reliable; once the model drifts far from θ^* (at later reversion steps, or at low c), the Taylor approximation degrades and T_1, T_2 lose their interpretation as distinct gradient- and curvature-driven contributions. The mechanistic measurements (Figs. 4 and 5) use the step-matched 70M sweep, in which lower c also sees less total new-task data, so we cannot fully separate concentration from total new-task exposure as the driver of the curvature terms.

Our directional-curvature claim compares λ_{dir} against λ_{avg}, the curvature averaged over all parameter directions. Because the Hessian spectrum is dominated by near-zero eigenvalues, λ_{avg} is a baseline almost any structured direction would exceed: it shows $\Delta\theta$ is sharp relative to a random direction, but not relative to other task-relevant directions (e.g. the new-task gradient or a different fine-tuning task).

Future work includes testing whether the concentration effect holds for DPO or RLHF, which would connect these results to how frontier safety training is done, and examining the role of curvature in the localised wrapper-like changes (Panigrahi et al., 2023; Jain et al., 2024). Our experiments use language-modelling and synthetic tasks rather than safety or preference fine-tuning, so the safety implications above are motivational extrapolations pending these tests.

6. Conclusion

We used the concentration of fine-tuning data as an instrument to probe the mechanisms of fine-tuning forgetting. We decompose the per-step change in fine-tune loss during reversion into its first- and second-order Taylor terms. Both contribute, with the second-order term becoming more important at higher concentrations, providing empirical support for theoretical accounts of curvature-driven forgetting

(Springer et al., 2026).

Our results suggest that the temporal structure of training, and not just its quantity, shapes how durable the learned behaviour is. Thus, for alignment, interleaving safety data throughout training rather than concentrating it at the end is a natural direction to explore.

ACKNOWLEDGMENTS

We would like to thank Clément Dumas, Stefan Heimersheim, Irakli Shalibashvili, Sid Baines, Niclas Küpper, and Einar Urdshals for their useful conversations and feedback. Arcadia Impact's LASR Labs and their funders like Coefficient Giving provided compute, funding, and organizational support.

This work was partially supported by the European Union's NextGenerationEU Program, through the National Recovery and Resilience Plan of the Republic of Bulgaria [Grant Project No. BG-RRP-2.004-0008].

References

Baek, C., Monti, R. P., Schwab, D., Abbas, A., Adiga, R., Blakeney, C., Böther, M., Burstein, P., Carranza, A. G., Deng, A., et al. The finetuner's fallacy: When to pretrain with your finetuning data. *arXiv preprint arXiv:2603.16177*, 2026.

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 41–48, 2009. doi:10.1145/1553374.1553380.

Berglund, L., Stickland, A. C., Balesni, M., Kaufmann, M., Tong, M., Korbak, T., Kokotajlo, D., and Evans, O. Taken out of context: On measuring situational awareness in LLMs. *arXiv preprint arXiv:2309.00667*, 2023.

Béthune, L., Grangier, D., Busbridge, D., Gualdoni, E., Cuturi, M., and Ablin, P. Scaling laws for forgetting during finetuning with pretraining data injection. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 4020–4042. PMLR, 2025. URL https://proceedings.mlr.press/v267/bethune25a.html.

Betley, J., Warncke, N., Sztyber-Betley, A., Tan, D., Bao, X., Soto, M., Srivastava, M., Labenz, N., and Evans, O. Training large language models on narrow tasks can lead to broad misalignment. *Nature*, 649(8097):584–589, 2026. doi:10.1038/s41586-025-09937-5.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., Skowron, A., Sutawika, L., and van der Wal, O. 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. URL https://proceedings.mlr.press/v202/biderman23a.html.

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N. Machine unlearning. *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159, 2021. doi:10.1109/SP40001.2021.00019.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., and Rohrer, D. Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132 (3):354–380, 2006. doi:10.1037/0033-2909.132.3.354.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9268–9277, 2019. doi:10.1109/CVPR.2019.00949.

French, R. M. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999. doi:10.1016/S1364-6613(99)01294-2.

Gu, J., Yang, Z., Ding, C., Zhao, R., and Tan, F. CMR scaling law: Predicting critical mixture ratios for continual pre-training of language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16143–16162. Association for Computational Linguistics, 2024.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Ibrahim, A., Thérien, B., Gupta, K., Richter, M. L., Anthony, Q., Lesort, T., Belilovsky, E., and Rish, I. Simple and scalable strategies to continually pre-train large language models. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=DimPeeCxKO.

Jain, S., Kirk, R., Lubana, E. S., Dick, R. P., Tanaka, H., Grefenstette, E., Rocktäschel, T., and Krueger, D. S. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=A0HKeKl4Nl.

Kalra, D. S., Gagnon-Audet, J.-C., Gromov, A., Mediratta, I., Niu, K., Miller, A. H., and Shvartsman, M. A scalable measure of loss landscape curvature for analyzing the training dynamics of llms. *arXiv preprint arXiv:2601.16979*, 2026.

Kim, M., Kwak, J. M., Alssum, L., Ghanem, B., Torr, P., Krueger, D., Barez, F., and Bibi, A. Rethinking safety in LLM fine-tuning: An optimization perspective. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=ZnOoEA2nDn.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi:10.1073/pnas.1611835114.

Korbak, T., Shi, K., Chen, A., Bhalerao, R. V., Buckley, C., Phang, J., Bowman, S. R., and Perez, E. Pretraining language models with human preferences. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17506–17533. PMLR, 2023. URL https://proceedings.mlr.press/v202/korbak23a.html.

Kotha, S. and Liang, P. Replaying pre-training data improves fine-tuning. *arXiv preprint arXiv:2603.04964*, 2026.

Krasheninnikov, D., Turner, R. E., and Krueger, D. Fresh in memory: Training-order recency is linearly encoded in language model activations, 2025. URL https://arxiv.org/abs/2509.14223.

Li, H., Ding, L., Fang, M., and Tao, D. Revisiting catastrophic forgetting in large language model tuning. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 4297–4308, 2024.

Liu, B., Liu, X., Jin, X., Stone, P., and Liu, Q. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18878–18890, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/9d27fdf2477ffbff837d73ef7ae23db9-Abstract.html.

Lopez-Paz, D. and Ranzato, M. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/f87522788a2be2d171666752f97ddebb-Abstract.html.

Lubana, E. S., Bigelow, E. J., Dick, R. P., Krueger, D., and Tanaka, H. Mechanistic mode connectivity. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22965–23004. PMLR, 2023. URL https://proceedings.mlr.press/v202/lubana23a.html.

McCloskey, M. and Cohen, N. J. Catastrophic interference in connectionist networks: The sequential learning problem. In Bower, G. H. (ed.), *Psychology of Learning and Motivation*, volume 24, pp. 109–165. Academic Press, 1989. doi:10.1016/S0079-7421(08)60536-8.

Minder, J., Dumas, C., Slocum, S., Casademunt, H., Holmes, C., West, R., and Nanda, N. Narrow finetuning leaves clearly readable traces in activation differences. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=pxaCoKZWVf.

Panigrahi, A., Saunshi, N., Zhao, H., and Arora, S. Task-specific skill localization in fine-tuned language models. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27011–27033. PMLR, 2023. URL https://proceedings.mlr.press/v202/panigrahi23a.html.

Peng, S., Chen, P.-Y., Hull, M., and Chau, D. H. Navigating the safety landscape: Measuring risks in finetuning large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 95692–95715, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ada93fa6643735f294be51dc31eebbd4-Abstract-Conference.html.

Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., and Henderson, P. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hTEGyKf0dZ.

Qi, X., Panda, A., Lyu, K., Ma, X., Roy, S., Beirami, A., Mittal, P., and Henderson, P. Safety alignment should be made more than just a few tokens deep. In *International Conference on Learning Representations*, pp. 11401–11431, 2025. URL https://openreview.net/forum?id=6Mxhg9PtDE.

Ramesh, R., Lubana, E. S., Khona, M., Dick, R. P., and Tanaka, H. Compositional capabilities of autoregressive transformers: A study on synthetic, interpretable tasks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42074–42103. PMLR, 2024. URL

https://proceedings.mlr.press/v235/ramesh24a.html.

Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., and Hadsell, R. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Springer, M., Lee, C. P., Metevier, B., Castleman, J., Turbal, B., Jung, H., Shen, Z., and Korolova, A. The geometry of alignment collapse: When fine-tuning breaks safety. *arXiv preprint arXiv:2602.15799*, 2026.

Wang, Z., Yang, E., Shen, L., and Huang, H. A comprehensive survey of forgetting in deep learning beyond continual learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47:1464–1483, 2024. doi:10.1109/TPAMI.2024.3498346.

Xie, S. M., Pham, H., Dong, X., Du, N., Liu, H., Lu, Y., Liang, P., Le, Q. V., Ma, T., and Yu, A. W. DoReMi: Optimizing data mixtures speeds up language model pretraining. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2305.10429.

Yang, M., Zhan, Z., Chen, Y., Li, H., Wang, K., Zheng, K., Wang, Y., Wang, Q., Gao, J., and Wu, J. Learning the mechanism of catastrophic forgetting: A perspective from gradient similarity. *arXiv preprint arXiv:2601.21577*, 2026.

Zhang, C., Kjellström, H., and Mandt, S. Determinantal point processes for mini-batch diversification. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017. URL https://arxiv.org/abs/1705.00607.

Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., Zhang, S., Ghosh, G., Lewis, M., Zettlemoyer, L., and Levy, O. LIMA: Less is more for alignment. In *Advances in Neural Information Processing Systems*, volume 36, pp. 55006–55021, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ac662d74829e4407ce1d126477f4a03a-Abstract-Conference.html.

Appendices

A Notation 8

B Experimental details 9

C Taylor analysis of forgetting 10

D Extended LLM results 12

E Compositional capabilities results 14

F Wrappers and concentration 15

G Related Work 17

A. Notation

This appendix collects every symbol used in the main text and appendices. Symbols are grouped by role.

A.1. Schedule parameters

$c \in [0, 1]$ Concentration. Expected fraction of new-task examples per fine-tuning step (§2.1). c=1 is pure fine-tuning on the new distribution; c=0 is pure pre-training.

$\rho \in [0, 1]$ Correlation between the pre-training and new-task distributions used in the CC and PCFG settings (§F). ρ=1 gives fully correlated tasks; ρ=0 gives independent tasks.

η, η_k Learning rate; η_k is the learning rate at reversion step k.

A.2. Distributions and losses

$\mathcal{D}_{\textbf{pre}}$, $\mathcal{D}_{\textbf{new}}$ Pre-training and new-task data distributions.

$L_{\textbf{pre}}(\theta)$, $L_{\textbf{new}}(\theta)$ Expected losses on $\mathcal{D}_{\text{pre}}$ and $\mathcal{D}_{\text{new}}$ at parameters θ.

$\Delta L_{\textbf{new}}$ Per-step change in the new-task loss during reversion (Eq. 1).

A.3. Parameters and optima

d Number of trainable model parameters (used to normalise λ_{avg} in §2.3).

θ, θ_t Model parameters; θ_t denotes parameters at training step t.

k, θ_k Reversion step index and parameters at that step; Taylor terms are evaluated independently at each θ_k (§2.3).

θ_0 Fine-tuned checkpoint at the start of reversion ($\theta_0 = \theta_k$ for k=0).

$\Delta\theta$, $\Delta\theta_k$ Single-step parameter change during reversion on pre-train data; $\Delta\theta_k = -\eta_k \cdot \text{AdamW}(g_{\text{pre},k})$.

θ^* A fine-tune minimum, i.e. a local minimiser of L_{new} (used in the Taylor analysis of §4 and Appendix C).

θ^d Displaced fine-tune endpoint reached when mixing with pre-training data keeps the optimiser away from θ^* (§4).

$H^* = \nabla^2_\theta L_{\textbf{new}}(\theta^*)$ New-task Hessian at a fine-tune minimum; defines the metric for Hessian-weighted distances.

$\|\theta - \theta^*\|_{H^*}$ Hessian-weighted distance from θ to θ^*.

A.4. Taylor decomposition

$T_1 = g^\top_{\textbf{new}} \Delta\theta$ First-order (gradient-alignment) Taylor term in the per-step change ΔL_{new} (Eq. 1, §2.3).

$T_2 = \frac{1}{2}\Delta\theta^\top H_{\textbf{new}}\, \Delta\theta$ Second-order (curvature) Taylor term (Eq. 1).

$\mathcal{O}(\|\Delta\theta\|^3)$ Higher-order remainder in the Taylor expansion; deviations of ΔL_{new} from T_1+T_2 reflect this term (Fig. 5).

N Number of reversion steps in the Taylor-analysis horizon (N=10 in §5).

A.5. Gradient and curvature quantities

$g_{\textbf{new}} = \nabla_\theta L_{\textbf{new}}(\theta)$ New-task gradient.

$g_{\textbf{pre}} = \nabla_\theta L_{\textbf{pre}}(\theta)$ Pre-training gradient.

$g_{\textbf{pre},k} = \nabla_\theta L_{\textbf{pre}}(\theta_k)$ Pre-training gradient at reversion step k.

$\|g_{\textbf{pre}}\|$, $\|g_{\textbf{new}}\|$, $\|\Delta\theta\|$ Euclidean norms of the pre-training gradient, new-task gradient, and reversion update; $\|g_{\text{pre}}\|$ and $\|g_{\text{new}}\|$ measure drift from respective optima, and $\|\Delta\theta\|$ scales the relative magnitudes of T_1 and T_2 (§3.2, §4).

$\cos(g_{\textbf{new}}, \Delta\theta)$ First-order alignment: cosine similarity between new-task gradient and the reversion weight update.

$H_{\textbf{new}} = \nabla^2_\theta L_{\textbf{new}}(\theta)$ Hessian of the new-task loss at θ.

$\widehat{\Delta\theta} = \Delta\theta/\|\Delta\theta\|$ Unit-norm reversion update direction (§2.3).

$\lambda_{\textbf{dir}} = \widehat{\Delta\theta}^\top H_{\textbf{new}}\, \widehat{\Delta\theta}$ Directional curvature of L_{new} along $\widehat{\Delta\theta}$ (the second-order alignment signal). Computed with a finite-difference Hessian-vector product (Appendix C.3).

$\lambda_{\textbf{avg}} = \text{Tr}(H_{\textbf{new}})/d$ Average curvature across all d parameter directions; a reference scale for λ_{dir}.

$\text{Tr}(H_{\textbf{new}})$ Trace of the new-task Hessian, estimated with the Hutchinson stochastic trace estimator (Appendix C).

A.6. Forgetting and retention metrics

$a_{\textbf{peak}}$ Peak new-task accuracy at the end of fine-tuning.

Reversion AUC Trapezoidal area under the new-task accuracy curve during reversion (Appendix B.5). Higher AUC indicates more durable learning.

Life threshold α For $\alpha \in \{0.95, 0.90, \ldots, 0.70\}$, the first reversion step at which new-task accuracy drops below $\alpha \cdot a_{\text{peak}}$ (Appendix B.5).

Normalised perplexity, $\text{NormPPL}(t)$ Rescales each condition to a common $[0, 1]$ range (1.0 = full domain specialisation retained, 0.0 = fully reverted to pre-training baseline).

B. Experimental details

B.1. Pythia architecture and training schedule

Both Pythia models are decoder-only transformers from the Pythia suite (Biderman et al., 2023), pre-trained on the deduplicated Pile ($\sim$207B tokens, $\sim$1.5 epochs) with rotary positional embeddings, parallelised attention and feedforward layers, Flash Attention, and untied input/output embeddings. Table 1 summarises the architecture; Table 2 gives the training schedule across all three phases.

All LLM runs share: Adam ($\beta = (0.9, 0.95)$, $\epsilon = 10^{-8}$), weight decay 0.01, gradient clipping 1.0, and fine-tuning sequence length 512.

	Pythia-70M	Pythia-1B
Total params	70.4M	1011.8M
Non-embedding	18.9M	805.7M
Layers	6	16
Model dimension	512	2048
Attention heads	8	8
Peak LR	10^{-3}	3×10^{-4}
Sequence length	2048	2048

Table 1. Architecture and pre-training hyperparameters for the Pythia models used in our experiments (Biderman et al., 2023). Both use the deduplicated variant.

Volume-matched control (Pythia-1B). Because low-concentration schedules in the fixed-steps regime see fewer domain examples total, we additionally run a volume-matched experiment where ft_steps $= 3000/c$ (so all schedules see the same total domain examples): 100% gets 3,000 steps, 50% gets 6,000, 25% gets 12,000.

B.2. LLM datasets and budget modes

Models. `EleutherAI/pythia-70m-deduped` and `EleutherAI/pythia-1b-deduped`.

Datasets (HuggingFace IDs). Chemistry: `antoinebcx/smiles-molecules-chembl`. Music: `sander-wood/irishman`. Biomedical: `ccdv/pubmed-summarization`. General: `monology/pile-uncopyrighted`. All text tokenized with the Pythia tokenizer, chunked into non-overlapping 512-token blocks. Fine-tune mixing is applied *per-sample within each batch*.

	Pre-train	Fine-tune	Reversion
Pythia-1B — SMILES			
Steps	143k (original)	10,000	3,000
Peak LR	3×10^{-4}	5×10^{-5}	5×10^{-5}
LR schedule	cosine $\rightarrow$ 10%	constant	constant
Warmup	1% (1,430 steps)	200 steps	—
Eff. batch	1024	16	16
Pythia-70M — SMILES / PubMed			
Steps	143k (original)	12k / 8k	3,000
Peak LR	10^{-3}	5×10^{-5}	5×10^{-5}
LR schedule	cosine $\rightarrow$ 10%	constant	constant
Warmup	1% (1,430 steps)	400 / 200 steps	—
Eff. batch	1024	20 / 40	16 / 40

Table 2. Training schedule for each Pythia experiment. "Pre-train" refers to the original Pythia pre-training (Biderman et al., 2023); we load the final checkpoint and apply our fine-tune and reversion phases. Where two values are shown (X / Y), the first is SMILES and the second PubMed. Effective batch size accounts for gradient accumulation.

Budget modes. In *steps* mode every fine-tuning run uses the same number of steps, so low-c runs see proportionally less new-task data. In *volume* mode the step count scales as $\lceil \text{ft_steps}/b \rceil$ so every fine-tuning run sees the same number of new-task tokens. The 70M sweeps used *steps*; the 1B sweeps used *volume*.

B.3. CC training schedule

Table 3 gives the CC training schedule. All phases use AdamW ($\beta = (0.9, 0.9)$, weight decay 10^{-3}), gradient clipping at 1.0, mixed precision (bfloat16), and batch size 128. The LR follows a piecewise cosine schedule that decays continuously across all three phases. Schedules below c=0.40 are excluded from main results because they do not reliably converge to 100% new-task accuracy within the allocated fine-tune phase.

	Pre-train	Fine-tune	Reversion
Steps	600	50	500
Peak LR	10^{-3}	(continuing)	(continuing)
LR at phase end	30% of peak	15% of peak	5% of peak
LR schedule	piecewise cosine (continuous across phases)		
Warmup	50 steps	—	—
Batch size	128	128	128
Seeds	1 (shared)	10 per c	10 per c

Table 3. CC training schedule across all three phases. The LR follows a single piecewise cosine that decays continuously through pre-train, fine-tune, and reversion.

B.4. Concentration and correlation sweep grid values

LLM concentration levels. Pythia-1B (SMILES): $c \in \{1.0, 0.95, 0.90, 0.75, 0.50, 0.25\}$. Pythia-70M (SMILES): same levels. Pythia-70M (PubMed): $c \in \{1.0, 0.98, 0.95, 0.90, 0.75, 0.50, 0.25, 0.10\}$.

CC concentration levels. $c \in \{1.00, 0.98, 0.95, 0.90, 0.80, 0.70, 0.60, 0.50, 0.40\}$.

CC correlation $\times$ concentration grid. Correlation $\rho \in \{0/n_a, 1/n_a, \ldots, n_a/n_a\}$ crossed with $c \in \{100\%, 95\%, 90\%, 70\%, 50\%, 30\%\}$ at $n_a \in \{3, 4, 5, 6\}$ (3 – 5 seeds per cell).

PCFG correlation $\times$ concentration grid.

$$\rho \in \{0.00, 0.25, 0.50, 0.66, 0.75, 0.85, 0.92, 0.95, 1.00\}$$

crossed with

$$c \in \{0.10, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.85, 0.90, 1.00\}.$$

B.5. Reversion AUC

We measure how quickly the fine-tuned capability is forgotten using the area under the new-task performance curve during reversion. For the Compositional Capabilities task we integrate the new-task accuracy curve (the *Reversion AUC*); higher values indicate more durable learning. For the LLM language-modelling runs, where we track loss rather than accuracy, we instead integrate the new-task loss curve over reversion steps; here higher values indicate faster forgetting. In both cases, forgetting refers to how quickly the model loses the fine-tuned capability during reversion. Note that this is a different notion from the per-example forgetting studied in the training-dynamics literature, where an example is said to be "forgotten" if the model misclassifies it after having classified it correctly at an earlier step.

C. Taylor analysis of forgetting

C.1. Forgetting step impact on fine-tuning loss

During the continued pre-training stage we perform updates, $\Delta\theta$, to the model using pre-training data. The resulting change in fine-tuning loss can be obtained by Taylor expanding L_{new} around the current weights θ:

$$L_{\text{new}}(\theta + \Delta\theta) = L_{\text{new}}(\theta) + \nabla_\theta L_{\text{new}}(\theta)^\top \Delta\theta + \frac{1}{2}\Delta\theta^\top H_{\text{new}}\, \Delta\theta + \mathcal{O}(\|\Delta\theta\|^3), \tag{3}$$

where $H_{\text{new}} = \nabla^2_\theta L_{\text{new}}(\theta)$ is the Hessian of the fine-tuning loss and $g_{\text{new}} = \nabla_\theta L_{\text{new}}(\theta)$. We rearrange, and identify the first and second order terms,

$$\Delta L_{\text{new}} = \underbrace{g_{\text{new}}^\top \Delta\theta}_{T_1} + \underbrace{\frac{1}{2}\Delta\theta^\top H_{\text{new}}\, \Delta\theta}_{T_2} + \mathcal{O}(\|\Delta\theta\|^3). \tag{4}$$

The first term, T_1, is the dot product between the new-task gradient and the weight update: it is negative when the weight update opposes the new-task gradient, and positive when they align. The second term, T_2, captures how curved the fine-tuning loss landscape is in the direction the pre-training step takes the model. At a minimum, this term is always non-negative, meaning curvature always contributes to forgetting regardless of whether the gradients are aligned.

C.2. Relative magnitudes of T_1 and T_2 at a fine-tuning minimum

We now show that, near a fine-tuning minimum, T_1 is bounded by the distance from that minimum — implying that models fine-tuned at higher concentration, which sit closer to the minimum, will have forgetting dominated by the curvature term T_2.

To consider the relative magnitudes of the first and second order effects, suppose $c = 1$, and assume we will reach a minimum in the new-task loss, L^*, with parameters, θ^*. As one moves away from a pure concentration of fine-tuning data, $c < 1$, the model is also optimizing for the pre-training data, and thus at the end of fine-tuning, it may not be so deep in the fine-tuning minimum, and there will be a non-zero displacement term between its weights, θ, and the fine-tuning minimum, i.e. $(\theta - \theta^*) \neq 0$. Finding the Taylor expansion of the fine-tuning loss around this fine-tuned minimum yields a primary term that is quadratic,

$$L_{\text{new}}(\theta) \approx L^* + \tfrac{1}{2}(\theta - \theta^*)^\top H^*(\theta - \theta^*) + \mathcal{O}(\|(\theta - \theta^*)\|^3). \tag{5}$$

Thus, ignoring higher order terms, the gradient of the loss in this region is approximately linear in the displacement from the minimum,

$$\frac{\partial L_{\text{new}}(\theta)}{\partial \theta} = g_{\text{new}}(\theta) \approx H^*(\theta - \theta^*). \tag{6}$$

Now considering a forgetting step, $\Delta\theta$ at the current parameters, θ, the first and second order contributions to ΔL_{new} are,

$$T_1 = g_{\text{new}}^\top \Delta\theta \approx (\theta - \theta^*)^\top H^* \Delta\theta, \quad T_2 = \tfrac{1}{2}\Delta\theta^\top H^* \Delta\theta\,. \tag{7}$$

Notice the asymmetry: T_1 depends on the displacement $(\theta - \theta^*)$ but T_2 does not.

Applying the Cauchy-Schwarz Inequality [1], one obtains a relation between the magnitudes of these terms,

$$|T_1| \leq \sqrt{2\,T_2}\ \cdot\ \|\theta - \theta^*\|_{H^*}\,. \tag{8}$$

[1]The Cauchy–Schwarz inequality $|\langle a, b\rangle| \leq \|a\|\,\|b\|$ holds for

This bounds the first-order term by the product of the curvature term and the distance from the minimum. We note that this is general for any optimizer used (Adam, SGD, etc.). The displacement $\|\theta-\theta^*\|_{H^*}$ is not directly measurable, but $\|g_{\text{new}}\|$ serves as an empirical proxy: by the same linearisation, $g_{\text{new}}(\theta) \approx H^*(\theta-\theta^*)$, so it scales with displacement in the Hessian-weighted sense. Fig. 15 shows this proxy decreasing with c once the new-task data is sufficiently represented in the batch, consistent with higher-concentration fine-tuning ending closer to θ^*. The implication is:

At *high concentration* ($c \to 1$), the model ends fine-tuning close to θ^*, so $\|\theta-\theta^*\|_{H^*}$ is small. The bound is tight: T_1 is forced to be small relative to T_2, and forgetting is driven primarily by curvature. At *low concentration*, the model is pulled away from θ^* by the pre-training objective, so $\|\theta-\theta^*\|_{H^*}$ is large. The bound loosens: T_1 can contribute substantially, and forgetting reflects both gradient conflict and curvature.

C.3. Calculations: Taylor-term and curvature estimation

For each end-of-fine-tuning checkpoint $\theta_{\text{ft}}^{(c)}$, we simulate N_{steps} steps of the reversion phase on the pre-train loss starting from $\theta_{\text{ft}}^{(c)}$. At each step $k \in \{0,\dots,N_{\text{steps}}\}$ we record the two terms of the Taylor expansion of L_{new} around the current iterate θ_k:

$$T_1(\theta_k) = \langle g_{\text{new}}(\theta_k),\, \Delta\theta_k\rangle, \tag{9}$$

$$T_2(\theta_k) = \tfrac{1}{2}\,\|\Delta\theta_k\|^2\,\lambda_{\text{dir}}(\theta_k), \tag{10}$$

where g_{new} is the gradient of the fine-tune loss, and $\lambda_{\text{dir}}(\theta_k) = \widehat{\Delta\theta}^\top H_{\text{new}}(\theta_k)\,\widehat{\Delta\theta}$ is the Rayleigh quotient of the fine-tune Hessian along the unit pre-train gradient direction, $\widehat{\Delta\theta} = \Delta\theta/\|\Delta\theta\|$. After recording, we take one optimizer step (using Adam) along the pre-train gradient,

$$\theta_{k+1} = \theta_k + \eta \cdot \text{Adam}\left(L_{\text{pre}}(\theta_k)\right),$$

and repeat.

Gradient estimation. In what follows, due to the learning rate scheduler, the initial learning rate is 0, and so to calculate $\Delta\theta$, we instead use a value that is the average of the first 10 steps of the scheduler.

Both g_{new} and $\Delta\theta$ are estimated as the mean of the per-batch gradients over N_{bat} mini-batches of size B, drawn from the held-out validation splits of the domain (fine-tune) and pile (pre-train) datasets respectively.

Directional curvature λ_{dir}. Rather than instantiating H_{new}, we compute Hessian-vector products with a central finite-difference approximation:

$$H_{\text{new}}\,v \;\approx\; \frac{g_{\text{new}}(\theta+\varepsilon v) - g_{\text{new}}(\theta-\varepsilon v)}{2\varepsilon}, \qquad \|v\| = 1,$$

with ε = `FD_EPS`. The parameter vector is restored to θ after each probe. Setting $v = \widehat{\Delta\theta}$ gives $\lambda_{\text{dir}} = \widehat{\Delta\theta}^\top(H_{\text{new}}\widehat{\Delta\theta})$.

Average curvature λ_{avg}. For the post-fine-tune checkpoint of each concentration, we additionally estimate the average eigenvalue of H_{new} via Hutchinson's trace estimator. We draw N_{hutch} Rademacher vectors $v^{(i)} \in \{-1,+1\}^d$ (d = total parameter count), form unit vectors $\hat{v}^{(i)} = v^{(i)}/\|v^{(i)}\| = v^{(i)}/\sqrt{d}$, and compute the same finite-difference Rayleigh quotient $\kappa^{(i)} = (\hat{v}^{(i)})^\top H_{\text{new}}\,\hat{v}^{(i)}$ for each. Because $\mathbb{E}_{\hat{v}}[\hat{v}^\top H\hat{v}] = \text{Tr}(H)/d$ for isotropic unit directions, we estimate

$$\lambda_{\text{avg}} \;=\; \frac{\text{Tr}(H_{\text{new}})}{d} \;\approx\; \frac{1}{N_{\text{hutch}}}\sum_{i=1}^{N_{\text{hutch}}}\kappa^{(i)}.$$

The full trace estimate is recovered as $d \cdot \lambda_{\text{avg}}$.

Hyperparameters. We use the same reversion learning rate as the main experiments, $N_{\text{steps}} = 10$, $B = 16$, $N_{\text{bat}} = 10$, $\varepsilon = 10^{-3}$, and $N_{\text{hutch}} = 50$. Further, we use 6 seeds for the calculations of λ_{dir} and λ_{avg}, and the error bars on the plots represent ± 1 standard deviation. All computations are performed in single-precision float32.

C.4. Update magnitude under SGD vs. Adam

Under vanilla SGD, the reversion update is $\Delta\theta = -\eta\, g_{\text{pre}}$, so $\|\Delta\theta\| \propto \|g_{\text{pre}}\|$ exactly and any pre-training drift accumulated during fine-tuning translates directly into larger reversion steps. For adaptive optimisers such as Adam, this coupling is weaker. Adam's update is approximately $\Delta\theta \approx -\eta\,\hat{m}/(\sqrt{\hat{v}}+\epsilon)$, where $\hat{m}$ and $\hat{v}$ are bias-corrected first- and second-moment estimates of the gradient. At the first reversion step, the bias correction makes the update approximately a per-coordinate sign step, decoupling $\|\Delta\theta\|$ from $\|g_{\text{pre}}\|$. As reversion proceeds, $\hat{v}$ accumulates and begins to track the pre-training gradient statistics, restoring partial coupling between update magnitude and gradient norm.

This means $\|g_{\text{pre}}\|$ is a clean drift proxy under SGD but only a qualitative one under Adam, particularly in the first few reversion steps. Despite this, we observe empirically

any inner product. Taking $a = \theta-\theta^*$, $b = \Delta\theta$, and the inner product $\langle a,b\rangle_{H^*} = a^\top H^* b$:

$$|T_1| = |\langle\theta-\theta^*,\Delta\theta\rangle_{H^*}| \le \|\theta-\theta^*\|_{H^*}\;\|\Delta\theta\|_{H^*} = \|\theta-\theta^*\|_{H^*}\;\sqrt{2\,T_2},$$

where we used $\|\Delta\theta\|_{H^*}^2 = \Delta\theta^\top H^*\Delta\theta = 2\,T_2$.

(Fig. 6) that the reversion update norms remain monotonic in c throughout the early reversion window, indicating that the concentration ordering survives Adam's normalisation. We attribute this to the fact that higher-c fine-tuning increases $\|g_{\text{pre}}\|$ by a large enough margin (Fig. 4b, especially the $c = 1$ outlier) that even Adam's per-coordinate normalisation does not erase the ordering.

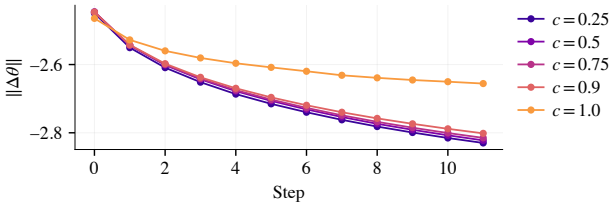


Figure 6. The magnitude of the weight update across the 11 steps of training from Fig. 5. One sees that it is higher for larger concentrations.

We also show additional results (Fig. 7) for the relative magnitudes of the T_1 and T_2 when SGD is used instead of AdamW, showing the same trend as Fig. 5.

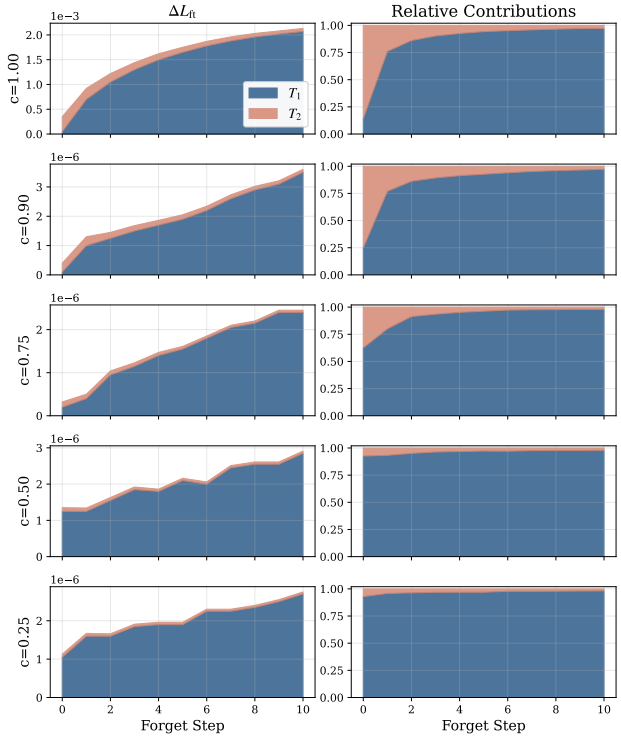


Figure 7. Analogous figure to Fig. 5, but for the case of Vanilla SGD instead of AdamW.

D. Extended LLM results

This appendix collects extended LLM results that complement the main findings in Section 3. Unless specified otherwise, all the subsequent results are for Pythia-70M fine-tuned on Chemistry (SMILES, ChEMBL), lr $= 5\times10^{-5}$, fine-tuning for 15k steps, and continued pre-training for 3k steps.

c	Chemistry	Music	Biomedical
	FT=15k, CPT=3k	FT=15k, CPT=3k	FT=15k, CPT=3k
1.00	0.20	2.43	1.45
0.90	0.12	1.59	0.83
0.75	0.09	1.25	0.64
0.50	0.07	0.87	0.47
0.25	**0.05**	**0.62**	**0.36**

Table 4. Relative degradation: $\ell_{\text{CPT}}/\ell_{\text{new}} - 1$, where ℓ_{new} is the domain validation loss at the end of fine-tuning and ℓ_{CPT} at the end of continued pre-training. Higher values indicate more degradation.

	Post-FT		Post-CPT	
c	Domain	Pile	Domain	Pile
Pretrained	3.92	3.06	—	—
1.00	2.96	5.92	3.55	3.02
0.90	2.97	3.10	3.31	3.01
0.75	2.98	3.04	3.25	3.00
0.50	3.02	3.01	3.23	2.99
0.25	**3.08**	**2.99**	**3.25**	**2.99**

Table 5. Validation cross-entropy loss after fine-tuning (Post-FT) and after continued pre-training on the Pile (Post-CPT). Lower concentration keeps lower loss post continued pre-training.

Volume-matched chemistry control (Pythia-1B). The main text shows the volume-matched PubMed control in Fig. 2. For completeness, we include the analogous Pythia-1B chemistry (SMILES) run here. Matching the total number of domain tokens across concentrations again preserves a monotonic relationship between concentration and the post-reversion rebound in domain loss.

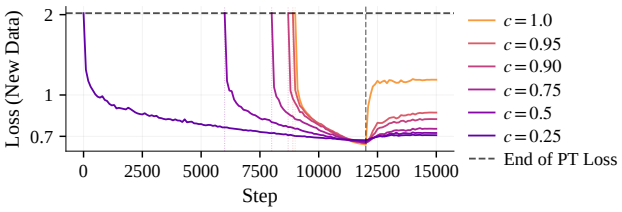


Figure 8. Pythia-1B domain validation loss on Chemistry (SMILES) during fine-tuning and continued pre-training on a volume-matched budget. As in the PubMed control shown in the main text, lower concentration yields smaller loss rebounds after the fine-tune→reversion boundary.

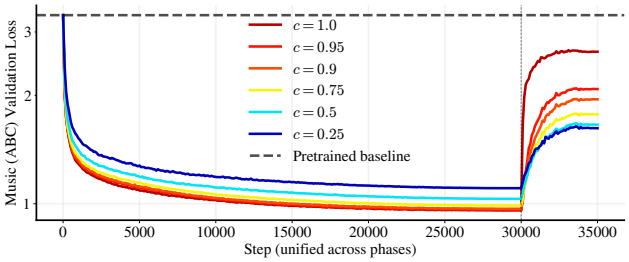


Figure 9. Domain validation loss during fine-tuning and reversion on Music (ABC notation), under a step-matched budget (Pythia-70M). Higher concentration leads to faster forgetting.

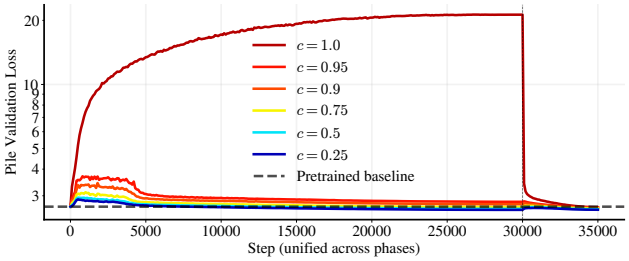


Figure 10. Pre-training data (Pile) Validation loss, fine-tuning. The pre-training loss is disproportionately higher in the 100% case.

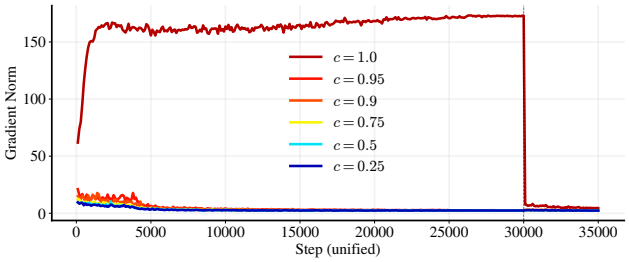


Figure 11. Gradient norm of the pre-training data, fine-tuning. The pre-training gradient norm is disproportionately higher in the 100% case.

Stepwise PubMed diagnostic. The mechanistic analysis in the main text uses Pythia-70M fine-tuned on PubMed to show that the pre-training task gradient norm grows with concentration. We include the full stepwise trace here for reference.

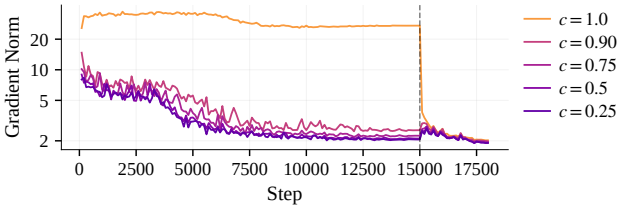


Figure 12. Biomedical dataset. Pre-training task gradient norm during fine-tuning.

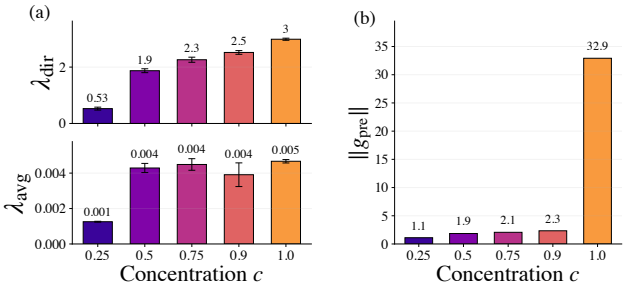


Figure 14. The same plot as in Fig. 4 in the main text, but (a) is shown in linear scale. The error bars are the $\pm 1\sigma$ from 6 seeds.

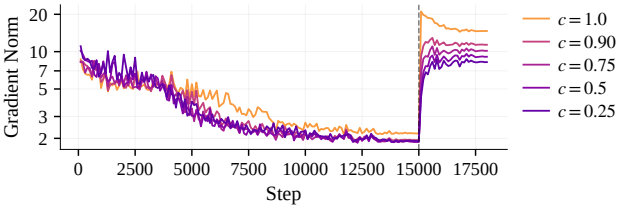


Figure 13. Biomedical dataset. Pre-training task gradient norm during PubMed fine-tuning, shown across training steps for each concentration. Higher concentration produces larger pre-training gradient norms throughout fine-tuning, matching the trend summarized in Fig. 12.

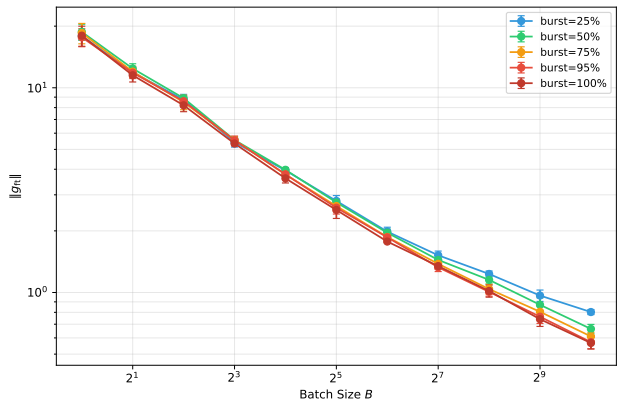


Figure 15. Gradient norm of the new-task data after fine-tuning. One sees the norm follows a $\sqrt{B}$ relation, and when the data is sufficiently represented, the norm of higher concentrations is lower than that of the lower concentrations.

D.1. Learning rate sensitivity

The magnitude of forgetting depends strongly on hyperparameter choices, especially learning rate (Kim et al., 2025; Ibrahim et al., 2024) and optimizer (Panigrahi et al., 2023). We also empirically observed that batch size had a significant effect. To quantify this interaction, we sweep the learning rate across the fine-tuning phase on the Compositional Capabilities setup while holding all other hyperparameters fixed, and measure forgetting at several concentration levels

(Fig. 16).

The learning rate has a dramatic effect on forgetting behaviour: with this setting, typically at low learning rates (relative to pre-training learning rate) fine-tuning is largely forgotten during reversion, while at high learning rates the fine-tuned capability is retained almost entirely for the allocated steps.

Our monotonic ordering replicates across the settings we tested, but absolute numbers should not be compared across hyperparameter regimes.

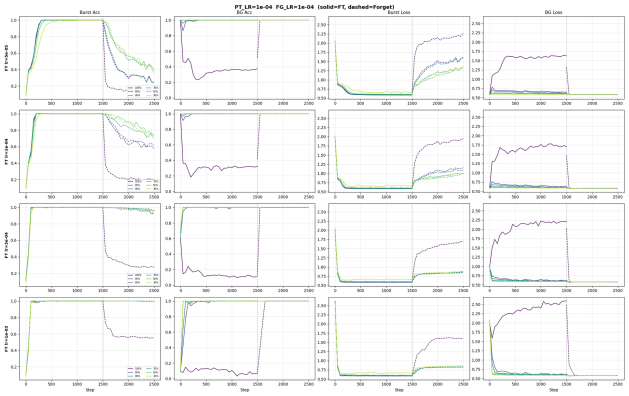


Figure 16. Learning Rate Sweep on Compositional Capabilities showing the effect of this parameter on forgetting.

E. Compositional capabilities results

This appendix presents the full Compositional Capabilities (CC) experimental setup and results, which complement the LLM findings in the main text. All figures use 10 seeds with concentration schedules $c \in \{40, 50, 60, 70, 80, 90, 95, 98, 100\}\%$ unless noted otherwise.

E.1. Setup

We use the Compositional Capabilities framework of Ramesh et al. (2024), in which 6-layer nanoGPT transformers are trained from scratch on pre-training bijection chains. During fine-tuning, a novel bijection is injected into the training mixture at concentration c, with the remaining $1-c$ fraction drawn from the original pre-training distribution. After fine-tuning, training reverts to the pre-training distribution and we measure how quickly the new task is forgotten. Training details (optimizer, learning rate schedule, batch size) are given in Table 3.

E.2. Concentration drives forgetting

All concentration schedules learn the new task to near-perfect accuracy (Table 6), but they differ sharply in how long that knowledge survives reversion. Fig. 17 shows both the Reversion AUC and the new-task accuracy trajectories during the early reversion window: higher concentration produces faster forgetting across all metrics. The effect is graded and monotonic — even small reductions in c (e.g. from 100% to 95%) yield measurably more durable learning. Pre-training task accuracy remains at 1.0 throughout for all schedules, confirming that lower concentrations do not impair the original task.

Schedule	Peak	Rev. AUC	90%-life	80%-life	70%-life
Concentration 100%	1.000	216 ± 5	25 ± 0	25 ± 0	25 ± 0
Concentration 98%	1.000	254 ± 8	25 ± 0	25 ± 0	32 ± 7
Concentration 95%	1.000	286 ± 4	25 ± 0	40 ± 8	60 ± 8
Concentration 90%	1.000	328 ± 8	40 ± 8	80 ± 14	150 ± 21
Concentration 80%	1.000	362 ± 7	72 ± 15	148 ± 16	238 ± 31
Concentration 70%	1.000	374 ± 7	78 ± 11	150 ± 14	272 ± 40
Concentration 60%	1.000	376 ± 4	72 ± 15	172 ± 16	280 ± 23
Concentration 50%	1.000	376 ± 5	80 ± 22	165 ± 10	275 ± 31
Concentration 40%	1.000	378 ± 4	85 ± 10	175 ± 21	268 ± 25

Table 6. CC task summary statistics (10 seeds, $\pm$ 95% CI). Peak accuracy is near-perfect for all schedules. Reversion AUC and $X\%$-life thresholds (steps to decay below $X\%$ of peak) both increase monotonically with decreasing concentration, quantifying the durability advantage of diluted training.

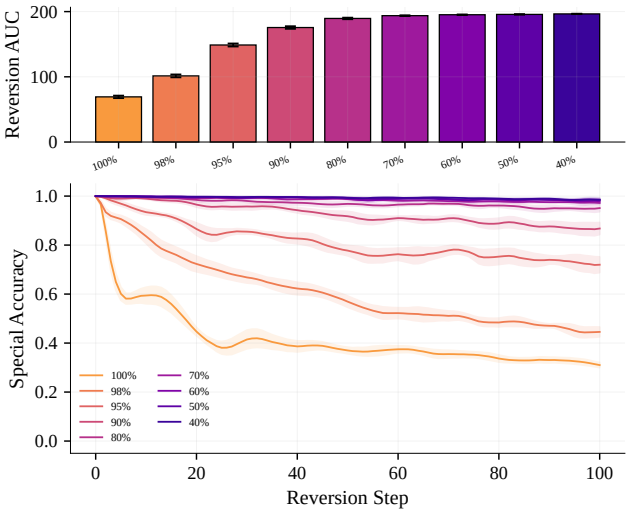


Figure 17. CC summary (10 seeds). **(A)** Reversion AUC across concentrations. **(B)** New-task accuracy during the early reversion window (shaded: 95% CI). Both panels show a monotonic, graded concentration effect: higher concentration yields faster forgetting, and the relationship is smooth rather than a sharp c=1 outlier.

E.3. Mechanistic diagnostics

We report the same gradient diagnostics as in the main text, applied to the CC setting.

Gradient alignment. Consistent with Yang et al. (2026), the cosine similarity between new and pre-training gradients becomes increasingly negative during fine-tuning (Fig. 18), indicating active competition in parameter space. However, at $c = 1$ the cosine remains near zero while performance still

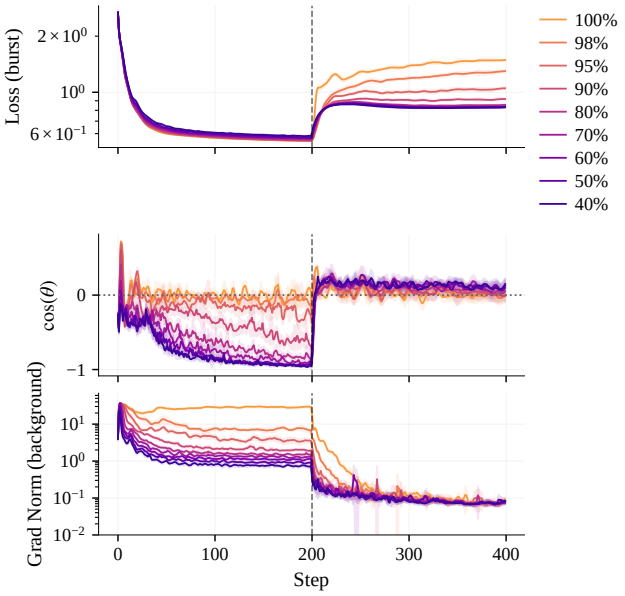


Figure 18. CC gradient diagnostics through fine-tuning and the first 200 reversion steps (log scale), averaged over 10 seeds. Panels show the new-task loss, $\cos(g_{\text{new}}, g_{\text{pre}})$, and the pre-training gradient norm $\|g_{\text{pre}}\|$. Higher concentration produces larger pre-training gradient norms and increasingly negative gradient cosine during fine-tuning.

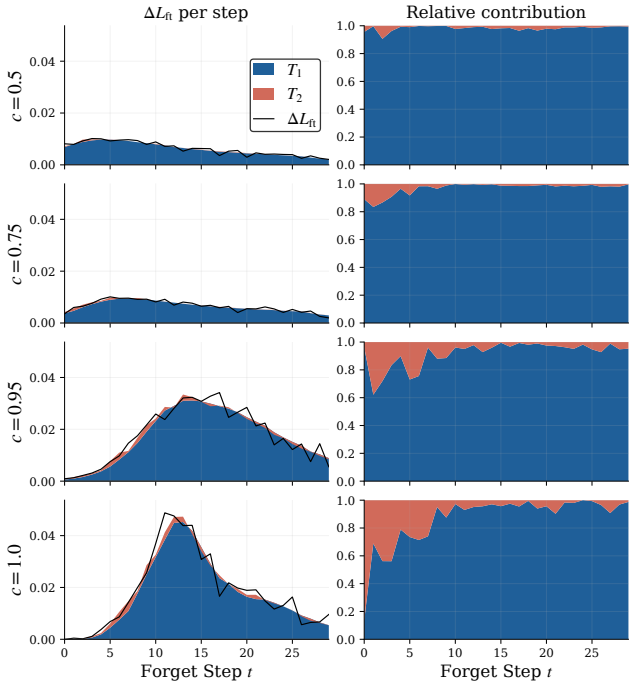


Figure 19. First- and second-order terms of the Taylor expansion during the first reversion steps on the CC task. Higher concentrations produce a larger relative contribution from the curvature term T_2.

degrades, demonstrating that first-order gradient conflict is not the only driver of forgetting.

Pre-training gradient norm. The pre-training gradient norm grows during fine-tuning and is monotonic with concentration (Fig. 18), consistent with the model moving further from the pre-training optimum at higher c. This matters for both Taylor terms of Eq. 1 in the main text, but unevenly: T_1 is linear in $\|g_{\text{pre}}\|$ while T_2 is quadratic, so larger g_{pre} at higher concentration amplifies the second-order contribution disproportionately.

Taylor decomposition. Fig. 19 decomposes the per-step change in fine-tuning loss during the first reversion steps into its first-order (T_1) and second-order (T_2) Taylor terms. As in the LLM setting, the curvature term's share of the total grows with concentration, providing converging evidence that the balance between first- and second-order forgetting shifts systematically with c. This analysis was done using vanilla SGD.

F. Wrappers and concentration

This section details additional results on the effect of varying fine-tuning data concentration on wrapper formation and forgetting under varying correlation. These experiments were mostly performed using the setup from (Jain et al., 2024) (as well as the CC setup), where a transformer is trained to solve tasks involving probabilistic context free grammars (PCFGs). In the original work, it was shown that wrappers form in the presence of spurious correlations between pre-training and fine-tuning tasks. In these experiments, we test how wrapper formation changes in the case where the fine-tuning data concentration is reduced, as well as probing how forgetting changes with simultaneous sweeps through concentration and correlation.

F.1. PCFG experiment details and training schedule

The setup from (Jain et al., 2024) involves generating PCFG strings consisting of the characters $\in [a, b, c]$, and then training a transformer to perform tasks such as counting occurrences of a, indexing the nth occurrence of a or finding what token is at position x. The first tokens of the input prompt to the transformer are dedicated to setting the task. For example, a full training string has the form of

$$\$,\ T,\ a,\ x,\ <,\ [\text{PCFG string}],\ >,\ = y,\ \$.$$

$\$$ is a start of sequence token, T is a task specific token, a is an operand token, x is a parameter token, and y is the answer token. For example, the string, 'Ca40', means to count the occurrences of 'a' in the last 40 tokens, 'Ib6' means to find the position of the 6th occurrence of 'b', and @_40 means to find the token at the 40th position from the end.

As in the original paper, the PCFG generation length was

255 and the tasks analyzed were:

- Ca40, Cb40, Cc40,
- Caa40, Cbb40, Ccc40
- Ia6,Ib6, Ic6
- Iaa6, Ibb6, Icc6
- @_40,

where 'aa' refers to an occurrence of consecutive characters. The fine-tune task is chosen to be Ca40, and the pre-train tasks are the remaining 12 tasks. As in the original paper, we implement the correlation, ρ, between the pre-training and fine-tuning data by enforcing that Ca40 = Cb40 + 1 for a fraction, ρ, of the data. This is done by generating many PCFG strings and extracting ones where the condition is true.

(Jain et al., 2024) demonstrated the formation of wrappers in the case of $\rho = 1$ by taking the model after fine-tuning, when it had 100% accuracy on Ca40 and 0% accuracy on Cb40, and selectively pruning neurons to improve the performance on Cb40. They demonstrated that pruning just $\approx$10 neurons was sufficient to return the model to its performance at the end of pre-training (0% accuracy on Ca40 and 100% accuracy on Cb40). In the results, we extend this analysis by testing wrapper formation in lower concentrations than $c = 1$.

Table 7 gives the PCFG training schedule. All phases use AdamW ($\beta = (0.9, 0.95)$, gradient clipping 1.0), batch size 96, and block size 512.

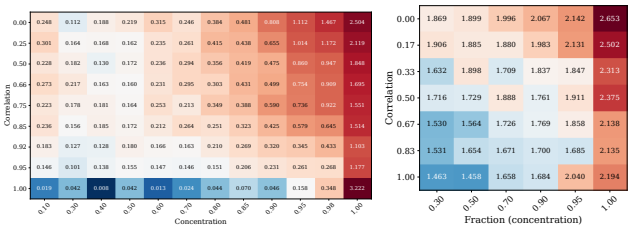
	Pretrain	**Fine-tune**	**Reverse**
Steps	200,000	10,000	8,000
Peak LR	10^{-3}	5×10^{-5}	5×10^{-5}
LR schedule	cosine decay	constant	constant
Batch size	96	96	96
Dropout	0.1	0.1	0.1

Table 7. PCFG training schedule across all three phases.

We also implement correlation in the case of the compositional capabilities setup (CC). Here, for the fine-tuning tasks we introduce six new possible functions. To create correlation, we simply make a fraction of these functions the same as the pre-training functions, meaning that we obtain 7 possible correlation values.

F.2. Results

Correlation. Firstly, we performed multiple training runs, simultaneously scanning through correlation and concentration on both the PCFG and CC setups; the accuracy on the fine-tune task for the PCFG task is shown in Fig. 20, and the forgetting result for both setups is seen in Fig. 21. We observe what has been discussed in the literature, that typically tasks with higher similarity are forgotten less fast. Further, the results suggest that when training on a task that is more correlated with the pre-training data, one can use a higher value of c than with a task that is less correlated.



(a) PCFG *(b)* Compositional Capabilities

Figure 21. Correlation vs concentration scan. **Left:** PCFG Count-A final reversion loss. **Right:** CC accuracy drop (end vs peak). Both tasks show the same interaction: fragility grows with concentration and shrinks with correlation.

Wrappers. Second, we perform an experiment to investigate wrapper formation at lower concentrations. Firstly, at parameters $[\rho = 1, c = 1]$, we replicate the wrapper-formation result from (Jain et al., 2024), that pruning 10 neurons from the network causes the model to regain its performance on Cb40, and dramatically to lose its performance on the fine-tune task, Ca40. A similar result is found (to a lesser extent) when reducing ρ to 0.95 — both results are seen in Fig. 22 (for the case of $c = 1$). We then perform the same analysis on models trained with lower concentration $\in [0.9, 0.95]$, and find that this is sufficient to prevent wrapper formation. By the end of fine-tuning, the accuracy on the Cb40 task has not degraded relative to the end of pre-training, and pruning neurons does not remove the performance on Ca40.

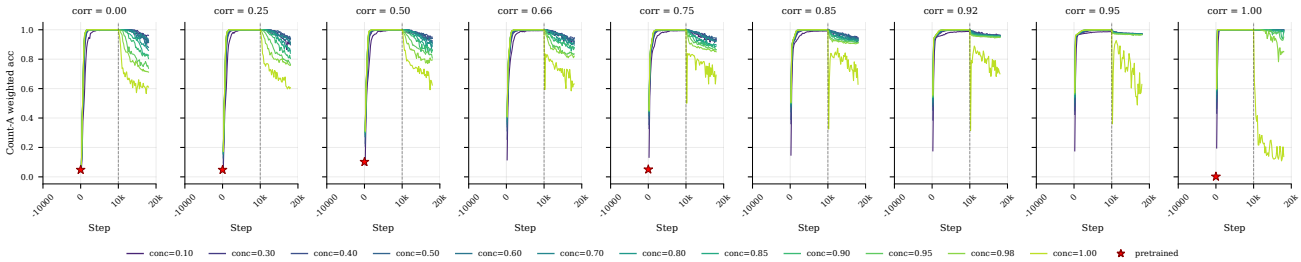


Figure 20. Accuracy on the fine-tune task (Ca40) over the course of training on the PCFG setup, across the scan over correlation ρ and concentration c. The curves show how the new task is acquired across the $\rho \times c$ grid; the corresponding forgetting trends are shown in Fig. 21.

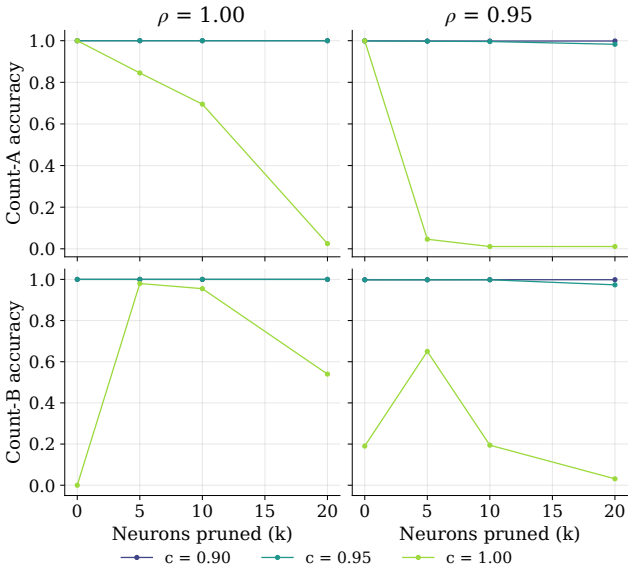


Figure 22. The resulting accuracy on Ca40 and Cb40 after pruning neurons on the PCFG setup with varying values of ρ and c. In the case of $c = 1$, pruning $\sim$10 neurons recovers Cb40 accuracy while collapsing Ca40, replicating the wrapper-formation result of (Jain et al., 2024). Reducing the concentration seems to avoid the formation of this wrapper.

Here we deviate slightly from the pruning method in the original paper, which was to choose parameters based on their gradients to most improve the performance on task Cb40; here, we instead scan through pruning every parameter, and choose those that most improve Cb40. The reason for this choice is for lower concentrations, the model still achieves a low loss on Cb40, and using the original technique was noisy. Crucially, the new technique still reproduces the wrapper observation from the original paper.

G. Related Work

We organise related work around five themes: training order and curriculum effects (§G.1), batch diversity and data mixing (§G.2), fine-tuning behaviour and its limitations (§G.3), alignment and safety data integration (§G.4), and catastrophic forgetting and continual learning (§G.5).

G.1. Training order and curriculum effects

Curriculum learning. Bengio et al. (2009) showed that presenting training examples in a meaningful order, typically from easy to hard, can improve convergence speed and generalisation. Subsequent work studies data ordering as a tool for improving *performance*; we study it as a cause of *shallowness*. The distinction is important: curriculum learning asks "what order makes the model learn best?" while we ask "what arrangement makes learning most durable?"

Out-of-context learning. Berglund et al. (2023) showed that language models can learn facts stated outside their typical usage context and later deploy them at test time, and that what is retained depends on how declarative information is distributed across training. We share the observation that order matters, but isolate a specific structural variable, temporal concentration, rather than studying order in general.

Training-order recency. Krasheninnikov et al. (2025) show that models linearly encode training-order recency in their activations. Our concentration sweep disentangles recency from concentration: all schedules see the target data at the same temporal position (the end of training), but at different concentrations. The fact that lower concentration produces more durable learning, despite an identical recency profile, shows that concentration, not recency, is the operative variable.

Spaced repetition. The cognitive science literature on spaced repetition (Cepeda et al., 2006) has long established that distributing study sessions over time produces more durable memory than massing them into a single block. Our work provides a mechanistic account of *why* this occurs in neural networks: spaced (mixed) schedules force the optimiser to find solutions that simultaneously satisfy both the target and base distributions, producing cooperative gradient directions and integrated representations.

G.2. Batch diversity and data mixing

Data mixing in pre-training. Large-scale pre-training pipelines routinely mix data from multiple domains within each batch (Xie et al., 2023). The rationale is typically framed in terms of domain coverage and preventing catastrophic forgetting. Our work provides theoretical grounding for this practice: mixing prevents the temporal concentration that produces shallow learning. It also suggests that the *degree* of mixing matters, not just whether domains are mixed, but how concentrated each domain is within the schedule.

Batch diversity. Determinantal Point Processes have been used to diversify minibatch composition (Zhang et al., 2017), showing that more diverse batches can improve training. Class-balanced sampling (Cui et al., 2019) addresses imbalanced datasets by ensuring each class is proportionally represented. These methods operate at the *within-batch* level; our concentration parameter operates at the *across-batch* level (how the target distribution is spread across training steps).

G.3. Fine-tuning behaviour and limitations

Fine-tuning shallowness and wrappers. Jain et al. (2024) demonstrated that fine-tuning on PCFG tasks produces wrapper-like changes: low-rank, later-layer-localised, and easily reversible. Related work on mode connectivity (Lubana et al., 2023) and skill localisation (Panigrahi et al., 2023) similarly finds that fine-tuning tends to produce mechanistically narrow, spatially localised updates. Our work extends these findings by showing *why* wrappers form: the temporal concentration of fine-tuning data is a (and possibly *the*) causal driver. We demonstrate that the wrapper signature (opposing gradients, localised centroid shifts, low-rank weight deltas) fades monotonically as concentration decreases, establishing a continuous relationship between data scheduling and the depth of learning.

Safety fine-tuning fragility. Qi et al. (2024) showed that safety fine-tuning of aligned models is easily compromised by a small number of adversarial examples, and that even benign fine-tuning on downstream tasks can degrade safety. These works demonstrate *that* safety fine-tuning is fragile; we identify a *structural cause* (concentration) and show that it is *tunable*, reducing concentration makes the same safety training more durable, without reducing its effectiveness.

LoRA and low-rank fine-tuning. LoRA (Hu et al., 2022) and related parameter-efficient methods constrain fine-tuning updates to be low-rank by design. The empirical success of LoRA provides independent structural evidence that fine-tuning naturally produces low-rank changes. Our gradient effective rank measurements confirm this: at high concentration, the gradient update is indeed low-rank. However, our results also show that this low-rank structure is a *symptom* of concentration, not an inherent requirement of the task: at lower concentration, gradient effective rank is higher and the model finds qualitatively different, more distributed solutions.

Machine unlearning. The machine unlearning literature (Bourtoule et al., 2021) studies how to remove specific data or capabilities from trained models. Our shallowness metrics, particularly the Reversion AUC and gradient interference measures, directly inform when unlearning is expected to be easy (high concentration $\rightarrow$ shallow $\rightarrow$ easy to remove) versus difficult (low concentration $\rightarrow$ deep $\rightarrow$ harder to remove).

G.4. Alignment and safety data integration

Alignment during pre-training. Recent work has explored integrating alignment data during pre-training rather than applying it as a post-hoc fine-tuning step. Korbak et al. (2023) showed that models pre-trained with human preference data embedded throughout training exhibit more robust alignment than those fine-tuned afterward; Zhou et al. (2023) separately demonstrated that high-quality alignment can be achieved with surprisingly little fine-tuning data, suggesting that *where* the signal sits in the training schedule matters at least as much as *how much* of it there is. These findings are highly consistent with our hypothesis: mixing alignment data throughout training (reducing concentration) should produce deeper, more durable alignment. Our controlled experiments provide a mechanistic explanation for why this works.

Continual alignment. Several works (Qi et al., 2025) have proposed methods for maintaining alignment across multiple rounds of fine-tuning or deployment. Our results suggest that the fundamental challenge of continual alignment is a scheduling problem: if each round of alignment training is delivered at high concentration, it will produce a shallow wrapper that the next round of task training can easily undo. Interleaving alignment data with task data at each stage would produce more durable integration.

G.5. Catastrophic forgetting and continual learning

Catastrophic forgetting. The problem of catastrophic forgetting, where training on new tasks destroys performance on old tasks (McCloskey & Cohen, 1989; French, 1999), is closely related to our work. Classical forgetting occurs when the new-task gradient overwrites the old-task parameters. Our setting inverts the direction: we study how *old-task* gradients overwrite *new-task* parameters during reversion. However, the underlying mechanism is the same: gradient interference between competing distributions. Our gradient

cosine measurements directly quantify this interference.

Elastic Weight Consolidation (EWC). EWC (Kirkpatrick et al., 2017) addresses catastrophic forgetting by adding a regularisation term that penalises changes to parameters important for previously learned tasks, weighted by the Fisher information matrix. EWC and related methods address forgetting by constraining the *parameter trajectory*; our work addresses it by constraining the *data schedule*. These approaches are complementary: EWC asks "which parameters should we protect?" while our work asks "how should we arrange data so that protection is unnecessary?"

Progressive neural networks and modular approaches. An alternative to overcoming forgetting is to allocate separate capacity for each task (Rusu et al., 2016). This avoids interference entirely but scales poorly. Our results suggest a middle ground: by reducing concentration (mixing tasks within batches), the model can learn to share representations across tasks without the destructive interference that concentrated training produces.

Gradient episodic memory (GEM). GEM (Lopez-Paz & Ranzato, 2017) addresses continual learning by constraining gradient updates to not increase loss on previously seen tasks. This is conceptually similar to what happens naturally at low concentration in our experiments: when new-task and pre-training data are mixed, the model receives gradient signal from both distributions simultaneously, which naturally constrains the update to directions that do not increase pre-training loss. Our gradient cosine measurements (positive at low concentration, negative at high) provide direct evidence that this constraint is satisfied automatically when concentration is low.