TANGENT GEOMETRY STRUCTURES FINE-TUNING FORGETTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-tuning can degrade capabilities that a model acquired before adaptation, but it is usually unclear before training which capabilities will be affected. We study this problem in the model's local tangent space and show that forgetting has a structured geometry there. Under gradient flow, the change in old-task outputs admits an exact decomposition: the new-task loss gradient is transported into the old-task function space through the cross-task tangent kernel K_{AB}. This identity separates the forcing induced by the incoming task from the geometry through which it affects existing outputs. In small-output regimes, the resulting forgetting drift is strongly anisotropic and concentrates along leading eigendirections of the old-task kernel K_{AA}. This structure suggests a pre-adaptation diagnostic: evaluate the tangent geometry once, before any parameter update, and use it to rank which outputs or capabilities are most likely to move. We test this prediction across vision models and four language-model families. For LoRA fine-tuning of language models, the resulting unnormalised cross-task gradient score ranks held-out capability changes before training, including on three model families not used during development. In contrast, cosine-normalised gradient similarity is negatively correlated with the same realised changes across all tested configurations, indicating that the gradient magnitude removed by normalization contains information relevant to interference. The tangent-space formulation also predicts an optimizer dependence: incorporating the AdamW preconditioner improves the one-step prediction under AdamW but not under SGD, as confirmed by permutation controls. For frozen representations, we further derive a Gauss–Newton correction for cross-entropy. The curvature-aware operator predicts realised output drift better than a curvature-free predictor even when the latter is given an oracle scalar ridge parameter. When representations are allowed to change, the tangent kernel rotates and the quality of a pre-adaptation predictor degrades with training, giving it a measurable validity horizon. Finally, constraining the vulnerable eigenspace identified by the tangent geometry reduces forgetting in vision models and cross-entropy degradation in language models. Our results characterize forgetting as a structured, locally predictable phenomenon in function space. The proposed diagnostic predicts relative vulnerability rather than forgetting magnitude, and is most reliable over short adaptations for which the local tangent geometry remains informative.

1 INTRODUCTION

Fine-tuning adapts a deployed model θ_A by training it further on an incoming task B. This usually improves task B and can degrade capabilities the model already has (McCloskey & Cohen, 1989; French, 1999; Goodfellow et al., 2014), safety alignment among them (Luo et al., 2025; Qi et al., 2024). Most measures of forgetting report what was lost after fine-tuning, and most defences act once adaptation is under way, constraining parameters (Kirkpatrick et al., 2017), gradient subspaces (Saha et al., 2021), stored outputs (Li & Hoiem, 2016; Pan et al., 2020) or rehearsed examples (Buzzega et al., 2020). We ask a question that comes earlier: before fine-tuning begins, can we predict which of the model's existing capabilities are most vulnerable?

Existing analyses answer adjacent questions, and the difference is granularity. Parameter-importance methods identify weights that matter for the old task (Kirkpatrick et al., 2017; Zenke

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

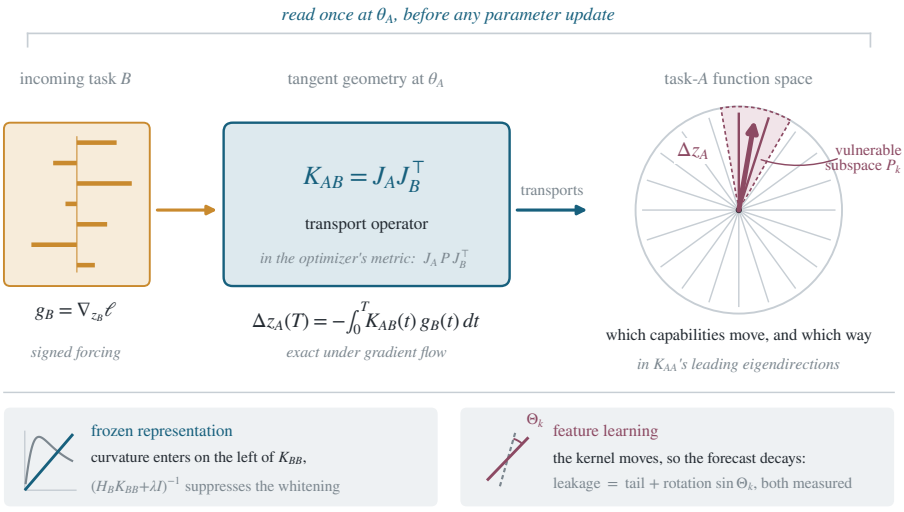

Figure 1: How a fine-tune changes old-task outputs. Task B supplies a signed forcing g_B in its own logit space (left), the cross-task tangent kernel $K_{AB} = J_A J_B^\top$ maps that forcing into changes of task-A outputs (centre), and we decompose the resulting drift in the eigenbasis of K_{AA} (right).

et al., 2017), task-similarity measures return one compatibility number per task pair (Doan et al., 2021; Lee et al., 2021), and influence methods attribute a fixed test loss to training points (Koh & Liang, 2017; Pruthi et al., 2020). None describes how an incoming task changes individual old-task outputs, which is what a ranking over capabilities needs.

We therefore study forgetting in function space rather than in parameter space. At θ_A the Jacobians J_A and J_B of old-task and new-task outputs give $K_{AB} = J_A J_B^\top$, which links the two function spaces. Under gradient flow on task B the old-task logits satisfy $\dot{z}_A = -K_{AB}(t)\, g_B(t)$, where g_B is the task-B loss gradient in logit space, so we call K_{AB} the transport operator (Figure 1).

Because the transported drift lies in task-A function space, we decompose it in the eigenbasis of $K_{AA} = J_A J_A^\top$, where it turns out to be strongly anisotropic. On Split-CIFAR-10, where the output space is small, about ten of two hundred directions carry 90% of the realised change. This is not universal: on frozen backbones with $C{=}100$ classes the same measurement is much weaker, and reaching 90% takes roughly half of the available directions (App. A). We therefore claim structured alignment with the tangent geometry rather than universal low rank.

The identity uses $K_{AB}(t)$, which is not available before training. We evaluate the tangent geometry once at θ_A instead, and that substitution is the approximation this paper tests. Used this way, the score ranks per-capability change on four language-model families, three never used during development, at a pooled Spearman of 0.76–0.79 (§6.3). It also orders which sentiment examples a LoRA (Hu et al., 2022) fine-tune of Pythia-1.4B (Biderman et al., 2023) will damage, in a setting where that fine-tune costs up to 29 accuracy points. We predict rankings, not magnitudes. Across all eighteen language-model configurations the unnormalised transport score is positively correlated with realised per-capability change while cosine-normalised gradient similarity is negatively correlated.

Three extensions follow from the same view. Cross-entropy changes the local transport through a Gauss–Newton factor acting on K_{BB} before inversion (Propositions 2 and 3). A moving representation moves the kernel, limiting how long a predictor evaluated at θ_A stays accurate (Lemma 1). Under preconditioned optimization the identity holds in the optimizer's own metric (§5.1). Finally, if the tangent geometry identifies the vulnerable directions, constraining them should reduce forgetting. We test this with the spectral penalty of Hidekel & Raviv (2026) and treat the outcome as evidence for the structural interpretation, not as a new method.

We make four contributions:

- **Prospective tangent-space prediction.** Starting from the exact chain-rule transport identity for old-task outputs, we formulate a prospective tangent-space predictor and extend its local form to cross-entropy curvature and preconditioned optimizers (§4–§5.1).

- **Pre-adaptation vulnerability ranking.** We show that this geometry can be evaluated before adaptation and used to rank vulnerable outputs, examples, and held-out capabilities across five frozen vision backbones, one representation-changing CLIP fine-tune, and four language-model families (§6).

- **Validity horizon under feature learning.** We measure when the local approximation breaks down. Lemma 1 separates two causes of forecast degradation, and we measure both (§6.2, §6.4).

- **Structural intervention.** We show that constraining the identified directions reduces forgetting, and that a matched-rank control changes the result when only the protected subspace is changed (§7).

2 Related Work

Tangent kernels, task overlap and continual learning. The fixed neural tangent kernel (Jacot et al., 2018; Lee et al., 2019; Chizat et al., 2019) is the limit in which we evaluate the transport operator, and whether practical fine-tuning remains near that regime has been studied separately (Malladi et al., 2023; Jang et al., 2024; Rahimi Afzal et al., 2025). Width and lazy-training effects have also been linked to reduced continual-learning forgetting (Mirzadeh et al., 2022). Doan et al. (2021) introduced the NTK overlap matrix, built from the same cross-task kernel K_{AB}, and kernel analyses of continual learning predict aggregate task-pair interference (Karakida & Akaho, 2022; Lee et al., 2021; Evron et al., 2022; Goldfarb et al., 2024). These works summarise compatibility at the task level; we instead use signed transport through K_{AB} to rank individual old-task outputs before adaptation. The closest example-level work (Jin & Ren, 2025) completes an observed task-by-example forgetting matrix and therefore requires a bank of measured fine-tunes, whereas our input is only θ_A and the incoming task. Related recent analyses study spectral or subspace structure of adaptation (Gu et al., 2026; Shuttleworth et al., 2025; Steele, 2026). The algebra is also close to influence functions: Eq. (3) is a group influence estimate with Gauss–Newton curvature (Koh & Liang, 2017), whose no-solve limit has the TracIn inner-product form (Pruthi et al., 2020). The difference is prospective use: we predict the output change caused by an incoming fine-tune rather than attribute an already observed loss (§6.3).

Forgetting in fine-tuned language models. Fine-tuning can damage held-out capabilities (Luo et al., 2025; Shi et al., 2024; Scialom et al., 2022), with less forgetting under LoRA (Biderman et al., 2024) and at larger scale (Ramasesh et al., 2022; 2021). Several mechanism-level accounts are particularly close. Kotha et al. (2024) attribute substantial apparent loss to task inference shifting toward the fine-tuning format, which we recover as a boundary regime (§6.4). Li et al. (2024) connect LLM forgetting to loss-landscape sharpness and mitigate it with sharpness-aware optimization. Yang et al. (2026) derive a parameter-level gradient-similarity account: their cosine uses the same two gradients as our scalar transport score but divides out the per-capability gradient norm, which the transport identity retains. Empirically the two rankings have opposite sign in our language-model experiments (§6.3). Recent work studies continual post-training through the covariance geometry of parameter updates (Wang et al., 2026); our object is instead the cross-task tangent map from the incoming loss gradient to individual pre-existing outputs, evaluated before the update. Low-dimensional structure of fine-tuning and tangent-space model editing are complementary observations (Aghajanyan et al., 2021; Mirzadeh et al., 2021; Ortiz-Jimenez et al., 2023; Ilharco et al., 2023; Wortsman et al., 2022).

Continual-learning defences. Existing methods protect weights or gradient subspaces (Kirkpatrick et al., 2017; Zenke et al., 2017; Saha et al., 2021; Farajtabar et al., 2020), replay data (Chaudhry et al., 2019; Buzzega et al., 2020), or stored function values (Li & Hoiem, 2016; Pan et al., 2020; Titsias et al., 2020; Rudner et al., 2022). Our intervention uses the last category only as a structural test: it anchors the top-k eigenspace of K_{AA} rather than proposing a new state-of-the-art continual-learning method. We also use results on kernel evolution (Bordelon & Pehlevan, 2022; Atanasov et al., 2022; Fort et al., 2020), Davis–Kahan bounds (Davis & Kahan, 1970; Yu et al., 2015), and Gauss–Newton curvature (Martens, 2020).

3 SETUP

Let θ_A be the pre-adaptation model and θ_T the model after T units of fine-tuning on task B. Task A is the set of capabilities we want to keep. We monitor task A through n_A fixed probes X_A and train on n_B inputs X_B, working in logit space throughout. Stacking gives $z_A(\theta) \in \mathbb{R}^{m_A}$ and $z_B(\theta) \in \mathbb{R}^{m_B}$ with $m_A = n_A C$ for C classes, and the forgetting drift is $\Delta z_A(T) = z_A(\theta_T) - z_A(\theta_A)$. The Jacobians $J_A = \partial z_A/\partial\theta$ and $J_B = \partial z_B/\partial\theta$ generate $K_{AA} = J_A J_A^\top$, $K_{AB} = J_A J_B^\top$ and $K_{BB} = J_B J_B^\top$. The task-B loss enters only through its logit-space gradient $g_B = \nabla_{z_B}\ell$, the forcing: for cross-entropy $g_i = p_i - e_{y_i}$ with per-sample Hessian $H_i = \mathrm{diag}(p_i) - p_i p_i^\top$ collected into the block-diagonal H_B, and for squared error the residual $r_B = z_B - y_B$. Finally $P_k(t)$ projects onto the top-k eigenspace of $K_{AA}(t)$.

4 STRUCTURE IN THE FROZEN REGIME

We first fix the representation, which separates function-space transport from movement of the kernel itself (§5). Everything below is evaluated at θ_A.

Proposition 1 (Forgetting identity). *Assume (A1) z_A, z_B are C^1 in θ near the path, (A2) task B is trained by gradient flow $\dot\theta = -\nabla_\theta L$, (A3) probes are fixed, and (A4) $T < \infty$. Then*

$$\Delta z_A(T) = -\int_0^T K_{AB}(t)\, g_B(t)\, dt, \qquad \Delta z_B(T) = -\int_0^T K_{BB}(t)\, g_B(t)\, dt. \tag{1}$$

The result follows from the chain rule. Since $\nabla_\theta L = J_B^\top g_B$, the update induced by task B changes task-A outputs as $\frac{d}{dt} z_A(\theta(t)) = J_A\dot\theta = -K_{AB}(t) g_B(t)$. No width or linearisation assumption is used. The NTK regime is the special case $K_{AB}(t) \equiv K_{AB}(0)$. For a discrete gradient step of size η, Taylor expansion under a locally Lipschitz Jacobian gives $\Delta z_A = -\eta K_{AB} g_B + O(\eta^2)$. The approximation tested later replaces the time-varying kernel by quantities evaluated at θ_A. Because the transported drift lies in the range of $K_{AA}(t)$, we analyse it in the eigenbasis of K_{AA}. Range membership is all the identity supplies. That the drift concentrates in the *leading* eigendirections is measured (Figure 2), not implied.

Corollary 1 (Fixed-kernel limit). *Freeze the kernels at θ_A (lazy regime) and take squared error. Then $r_B(t) = e^{-K_{BB}t} r_B(0)$ and Eq. (1) integrates in closed form to*

$$\Delta z_A(T) = -K_{AB}\, K_{BB}^{+}\big(I - e^{-K_{BB}T}\big)\, r_B(0) \xrightarrow[T\to\infty]{} -K_{AB} K_{BB}^{+} r_B(0), \tag{2}$$

the pseudoinverse sufficing because forcing in the null space of K_{BB} never reaches K_{AB}.

Replacing the K_{BB} filter in Eq. (2) by its ridge form gives the curvature-free reference $-K_{AB}(K_{BB} + \lambda I)^{-1} r_B(0)$ used below. Under squared error the fixed-kernel dynamics in Eq. (2) are exact; the ridge form is used as a regularised reference for the cross-entropy experiments.

4.1 CROSS-ENTROPY CURVATURE

For cross-entropy the task-B gradient changes as the logits change, so the closed form above does not apply. We build a local Gauss–Newton model of the run around θ_A.

Proposition 2 (CE Gauss–Newton local predictor). *Minimizing the ridge Gauss–Newton model of task B at θ_A, the second-order Taylor model with the $\sum_j g_{B,j}\nabla^2_\theta z_{B,j}$ term dropped, $g_B^\top J_B\delta\theta + \frac{1}{2}\delta\theta^\top J_B^\top H_B J_B\delta\theta + \frac{\lambda}{2}\|\delta\theta\|^2$ at fixed $\lambda > 0$ (a Levenberg–Marquardt ridge, so the objective is strictly convex), the minimizer's old-task drift is*

$$\Delta z_A^{\mathrm{GN}} = -K_{AB}\,(H_B K_{BB} + \lambda I)^{-1}\, g_B. \tag{3}$$

Equation (3) is that local model in kernel form by push-through algebra (App. A). Relative to the curvature-free reference, H_B modifies K_{BB} before the inverse is taken. Setting $H_B = I$ recovers the curvature-free ridge transport operator of Corollary 1, with the residual forcing replaced by the local cross-entropy gradient. No Newton step is required: Eq. (3) is a ridge-regularised local Gauss–Newton approximation to the cross-entropy dynamics. Treating λ as an inverse horizon fixes its units to those of $H_B K_{BB}$ and motivates the scale-normalised $\lambda = c\,\mathrm{tr}(K_{BB})/m_B$ used at language-model scale.

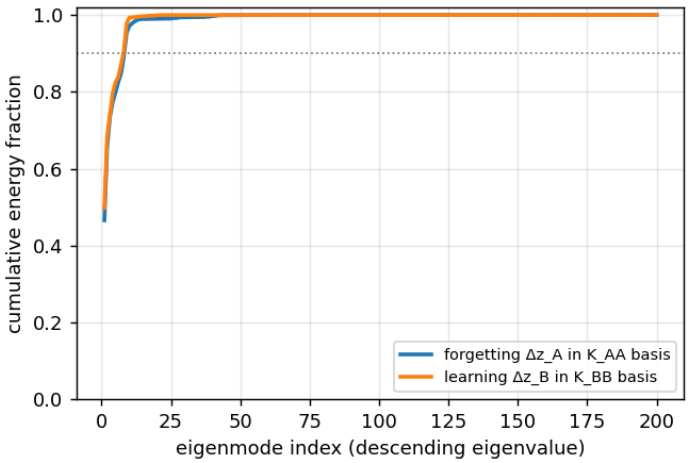

Figure 2: Forgetting drift concentrates in a few directions on Split-CIFAR-10. Each curve shows the share of realised output change captured by the leading directions of the K_{AA} eigenbasis (cross-entropy, trained backbone). Both drifts reach 90% within about ten of two hundred directions.

4.2 Two limits with opposing spectral filters

The two operators differ only inside one inverse. Write $T_{\text{lazy}} = K_{AB}(K_{BB} + \lambda I)^{-1}$ for the fixed-kernel operator of Corollary 1 and $T_{\text{GN}} = K_{AB}(H_B K_{BB} + \lambda I)^{-1}$ for the curvature-aware one.

Proposition 3 (Lazy and confident-curvature limits induce opposing spectral filters)**.** *Let* $J_B = U\Sigma V^\top$. *Then* $T_{\text{lazy}} = J_A V \operatorname{diag}\!\left(\frac{\sigma_i}{\sigma_i^2+\lambda}\right) U^\top$, *while in the confident limit* $H_B \to 0$, $T_{\text{GN}} \to \lambda^{-1} K_{AB} = \lambda^{-1} J_A V \operatorname{diag}(\sigma_i)\, U^\top$. *The lazy filter* $\sigma/(\sigma^2+\lambda)$ *is non-monotone, peaking at* $\sigma = \sqrt{\lambda}$ *and decaying as* $1/\sigma$*: it* whitens K_{AB} *and flattens the operator's spectrum. The confident-limit filter* σ *is monotone, so* T_{GN} *inherits* K_{AB}*'s own spectral decay.*

The two limits make the empirical distinction testable. The lazy solve applies the non-monotone filter $\sigma/(\sigma^2+\lambda)$, while confident cross-entropy approaches the unwhitened transport $\lambda^{-1}K_{AB}$. On the five frozen backbones the curvature-aware operator is 1.4–$1.7\times$ narrower in stable rank than the lazy one, and it predicts realised drift more accurately even against an oracle scalar ridge (Table 1). Ridge-scale and fixed-ridge controls are in App. C.

5 Beyond the Frozen Euclidean Regime

The predictor is evaluated at θ_A. If fine-tuning changes the representation, the tangent kernel changes with it and the predictor describes a point the model has left. This section measures how much that costs. The movement can be large: under partial fine-tuning of a ResNet-18 (`layer4` and head) the Frobenius drift of K_{AA} more than doubles its norm, and the top-k eigenspace rotates by tens of degrees. About half of the realised drift still lies inside the subspace identified at θ_A (Figure 4).

Lemma 1 separates two reasons why a subspace fixed at θ_A can stop capturing the drift.

Lemma 1 (Leakage from geometry movement)**.** *With* $w(t) = K_{AB}(t) g_B(t)$, $\sin\Theta_k(t) = \|(I - P_k(0))P_k(t)\|_{\text{op}}$, *and* $\varepsilon_t = \|(I - P_k(t))w(t)\|$ *the instantaneous tail,*

$$\|(I - P_k(0))\Delta z_A(T)\| \le \underbrace{\int_0^T \varepsilon_t \, dt}_{\text{tail of the current geometry}} + \underbrace{\int_0^T \sin\Theta_k(t)\, \|w(t)\|\, dt}_{\text{subspace rotation}},$$

where the rotation term is tied to the measured kernel movement of Figure 4 by the Davis–Kahan variant of Yu et al. (2015), through the top-k eigengap $\gamma_k(0)$ *of the unperturbed kernel (App. A).*

The bound separates drift already outside the current vulnerable subspace from loss due to rotation of that subspace away from its orientation at θ_A. In the measured partial-fine-tuning regime the

Table 1: The predicted drift direction agrees with the realised one on all five frozen backbones. *Drift cosine* compares predicted and realised directions of output change. *Oracle scalar* replaces H_B by a per-transition scalar αI fitted after the fact. *Random* is a forcing of matched size. Operator rank is the stable rank. Cross-entropy, Split-CIFAR-100, 3 seeds, λ=1 a priori. Ridge-scale and fixed-ridge controls are in B.

	drift cosine			operator rank	
backbone	forecast	oracle scalar	random	forecast	lazy
ResNet-18	.702±.006	.687±.004	0.00	232	404
ResNet-50	.720±.009	.708±.009	0.00	250	349
ViT-B/16	.761±.019	.747±.016	−0.01	249	389
DINOv2 ViT-S/14	.718±.010	.705±.007	−0.01	312	438
CLIP ViT-B/32	.738±.021	.723±.019	0.00	279	417

instantaneous tail remains small while the subspace rotation is large (Figure 4; details in App. A), identifying kernel movement as the main source of expiry for a predictor fixed at θ_A.

5.1 Transport under preconditioned descent

The derivation above assumes gradient flow, whereas practitioners use adaptive optimizers, which change the metric in parameter space. Under $\dot{\theta} = -P(t)\nabla_\theta L$ with P symmetric positive semi-definite, the chain rule gives $\dot{z}_A = -K^P_{AB}(t)g_B(t)$ with $K^P_{XY} = J_X P J_Y^\top$. This is again a transport operator, so every spectral statement of §4 carries over. Taking the ridge in the optimizer's own metric $\frac{\lambda}{2}\|\delta\theta\|^2_{P^{-1}}$ and applying push-through gives

$$\Delta z_A^{\mathrm{GN},P} = -\,K^P_{AB}\,(H_B K^P_{BB} + \lambda I)^{-1}\,g_B. \tag{4}$$

AdamW also adds momentum, moving moments, discrete steps and weight decay, so Eq. (4) is a prediction the experiments test rather than a description of the optimizer (App. D.1). For Adam at θ_A the early preconditioner is diagonal in $\hat{v}$ and can be estimated before any update from squared task-B gradients (Malladi et al., 2022), so the predictor stays pre-adaptation. The prediction is asymmetric and therefore testable: a preconditioner estimated for AdamW should improve the ranking under AdamW and not under SGD (§6.3).

6 Experiments

The experiments test four questions in turn: whether tangent-space transport predicts realised forgetting when the kernel is fixed, what happens when the representation and kernel move, whether the prediction extends to language-model fine-tuning and optimizer preconditioning, and over how many steps a predictor evaluated at θ_A stays useful. We use two targets: *drift direction*, the cosine between predicted and realised change in logits, and *vulnerability ranking*, the rank correlation between predicted and realised change. Ranking is the main practical target, and every predictor is evaluated before any parameter update. Ridge settings are fixed in advance and never tuned per cell (App. D).

6.1 Frozen-regime prediction

With the backbone frozen the tangent kernel does not change during fine-tuning. Two populations answer different questions here: Split-CIFAR-10 is where we measure concentration (Figure 2), and Split-CIFAR-100 with five frozen pretrained backbones is where we test directional prediction and the effect of curvature (Table 1). On Split-CIFAR-10 both realised drifts clear a matched-dimension random-subspace null in every cell (App. A), and a random forcing of matched size agrees with neither (Table 1), so the structure comes from the task and not from the geometry alone.

On the five pretrained backbones, frozen so only a linear head moves, the curvature-aware predicted drift tracks the realised one (Table 1). The oracle-scalar column tests a sharper prediction: setting $H_B = \alpha I$ turns T_{GN} into T_{lazy} at ridge λ/α, so the best per-transition α fitted after the fact bounds what any overall rescaling can achieve. The forecast leads that oracle by $+.0116$ to $+.0155$ on all five, a margin that is small but consistent and that separates curvature structure from rescaling: only 13–18% of the much wider gap to the curvature-free reference survives the oracle (App. C). This is the only setting in which we separately identify a curvature effect; the remaining curvature analyses are therefore treated as supporting rather than primary evidence.

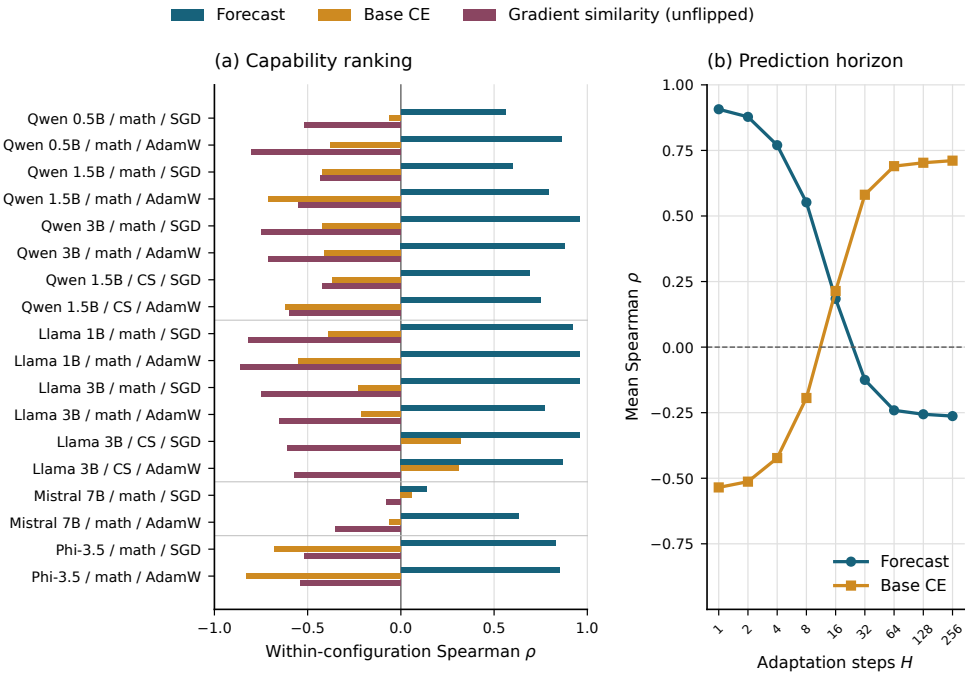

Figure 3: Per-capability prediction at language-model scale. **(a)** Rank correlation with realised capability change. Each row is one model/task/optimizer configuration; large markers show family means. The transport forecast is positively correlated with realised change, whereas cosine-normalised gradient similarity is negatively correlated. **(b)** On Qwen, the forecast is strongest early in adaptation and gradually loses predictive power as training proceeds; prior competence becomes the better predictor between eight and sixteen steps.

6.2 Feature learning

We stress-test a predictor frozen at θ_A by unfreezing the last transformer block and projection of CLIP ViT-B/32 (Radford et al., 2021). Because 96.6% of the realised drift energy is a common direction shared across probes, raw cosine is uninformative; after projecting that component out, the forecast still exceeds a matched-norm random direction by $+.226$ on all five seeds, with the same separation across a depth ladder and a ViT-B/16 control (App. E). This supports residual directional information after the kernel moves, but not the practical ranking claim in this regime: a constant-direction oracle ranks zero-shot damage at least as well as the forecast (p=0.31).

6.3 Language-model scale

At language-model scale, the calibrated Gauss–Newton predictor is numerically close to the raw transported forcing $-K_{AB}g_B$. The experiments in this section therefore test the transport view itself, rather than the curvature correction. Up to sign, $-K_{AB}g_B$ has the same inner-product form as TracIn (Pruthi et al., 2020; Koh & Liang, 2017), but its use here is different: we evaluate it before adaptation and ask whether it predicts which existing capabilities will move.

For a capability A_i, scalarising the predicted logit change by its old-task loss gradient gives

$$s_i = -\left\langle J_{A_i}^\top g_{A_i}, J_B^\top g_B \right\rangle .$$

A cosine similarity between the same gradients removes the norm of the old-task gradient, whereas the transport score retains it. The preconditioned version of §5.1 further replaces the Euclidean inner product by the optimizer's metric. We evaluate these quantities on Qwen2.5 models using twenty held-out MMLU subjects, then repeat the same protocol without retuning on Llama-3.2, Mistral-7B, and Phi-3.5 (App. D).

The main result is that a single tangent evaluation at θ_A predicts the relative signed movement of held-out capabilities before fine-tuning. On the Qwen development grid, the forecast is positive in 39 of 40 cells. After freezing the procedure, the same result transfers to the three held-out model

Table 2: Prospective capability ranking at the four-step horizon. The Qwen split was used for development; the three held-out families were evaluated with the procedure frozen. Precision@5 measures recovery of the realised five most-damaged subjects.

split	cells	mean Spearman ρ	positive cells	precision@5
Qwen development	40	$+.760 \pm .229$	$39/40$	.705
held-out families	50	$+.789 \pm .256$	$49/50$	.716
pooled	90	$+.776$	$88/90$	.711

families: 49 of 50 cells remain positive (Table 2). The experiment is a ranking problem rather than a claim about absolute forgetting magnitude: at this short horizon many subjects improve, and the predictor orders which capabilities move and in which direction.

This ordering is not explained by unsigned task overlap or by gradient magnitude alone. NTK-overlap and CKA baselines show little ability to rank the same capabilities (Table 9), while controlling for the old-task gradient norm leaves the transport ranking essentially unchanged. Cosine-normalised gradient similarity instead tends to reverse the ordering. For the more decision-oriented task of identifying the five most-damaged subjects, the transport forecast reaches precision@5 $= 0.711$ against a chance level of 0.25 (Table 8). Full per-cell results and controls are reported in App. D.

The four-step MMLU experiments mostly probe small output changes, so we also test a regime with substantial forgetting. We fine-tune Pythia-1.4B (Biderman et al., 2023) from SST-2 (GLUE, Wang et al., 2018) toward QQP, RTE, or QNLI for 200 steps, producing losses of up to 29 accuracy points on the old task. The forecast evaluated before adaptation still ranks the examples that will be damaged and separates examples that later flip from those that remain correct (App. D).

The optimizer-metric extension gives a second prospective test. At the first step, the Adam-metric kernel improves the forecast under AdamW and degrades it under SGD, while permutation controls remove this optimizer-specific advantage (Table 10; App. D). The effect disappears as adaptation moves away from θ_A, consistent with the local nature of the approximation.

This locality is the main limitation of the predictor. Forecast quality is highest when the realised movement is small and deteriorates as the fine-tune moves farther from the initial model. Importantly, the predicted movement magnitude itself provides a useful warning signal: configurations predicted to move more are also those for which the ranking is least reliable. We examine this validity horizon directly in §6.4.

6.4 How long a predictor evaluated at θ_A stays useful

As training proceeds the model moves away from θ_A and the local approximation becomes less accurate. Lemma 1 identifies the two terms responsible and Figure 3(b) measures the effect. On the MMLU grid the ranking is strongest at the first step and decays with adaptation, crossing prior competence between eight and sixteen steps. On the Pythia per-example suite it remains predictive through all 200. The difference is consistent with how far the local geometry moves, which Lemma 1 separates into instantaneous-tail and subspace-rotation terms.

Two further boundaries are measured. At learning rates 50–$100\times$ larger, held-out subjects lose tens of accuracy points at once and the ordering collapses onto prior competence, the task-inference regime of Kotha et al. (2024), outside the interference regime described here. The method is therefore a pre-adaptation ranking tool for short fine-tunes, not a substitute for running the adaptation.

7 The Structure Reduces Forgetting

If the tangent geometry identifies the directions that carry interference, selectively constraining those directions should reduce it. We use the spectral output-space penalty of Hidekel & Raviv (2026) as an intervention:

$$\mathcal{L} = \mathcal{L}_{\text{new}}(\theta) + \frac{\mu}{k} \sum_{\tau} \left\| U_k^{(\tau)\top} \left(z(\theta, X_\tau) - z_\tau^{\text{ref}} \right) \right\|^2, \tag{5}$$

where $U_k^{(\tau)}$ spans the top-k eigenspace of K_{AA} on stored task-τ anchors. Full continual-learning tables and protocol details are in App. B; standard replay and GPM are competitive or better in several settings, so we make no state-of-the-art method claim.

At language-model scale we hold the anchors, rank and penalty strength fixed and change only the protected subspace. The top eigenspace removes 79.8% of the cross-entropy damage, versus 71.8% for a random rank-matched subspace, 67.0% for an axis-aligned selection of large kernel-diagonal outputs, and 32.8% for the bottom eigenspace; the complete basis removes 75.2% (Table 6). Thus anchoring itself explains much of the protection, but the identity of the subspace contributes additional signal. The same penalty also reduces forgetting on frozen vision backbones and LoRA adapters.

8 Limitations

The predictor evaluates tangent geometry before adaptation, so its reliability is limited to a regime-dependent horizon as the representation and kernel evolve (§6.4). It predicts drift directions and relative vulnerability, rather than forgetting magnitudes. The exact transport identity does not itself guarantee the accuracy of this local approximation. Our language-model results primarily support raw transport, while the curvature contribution is isolated only in the frozen-backbone setting. Feature-learning evidence remains limited to a small set of configurations with a dominant shared drift component, and the structural bounds are not informative at the measured operating points.

The language-model protection results also require a distinction between loss and accuracy: at these operating points, the penalty mainly reduces cross-entropy degradation rather than recovering correct answers. Evaluating the operator costs more than the short adaptation it describes, and whether this cost can be amortised across candidate fine-tunes remains untested.

9 Conclusion

Our central contribution is a function-space account that connects the geometry of fine-tuning interference to its prediction and mitigation. The incoming task supplies a learning signal, and the cross-task tangent kernel determines how that signal reaches existing outputs. Their interaction gives interference a measurable structure and makes relative vulnerability accessible before adaptation begins.

The predictive content lies in this interaction: geometric overlap alone does not determine which capabilities change, because the direction and scale of the incoming signal and the optimizer's metric also matter. Our findings show that tangent geometry can support useful local forecasts and identify directions whose protection reduces degradation. The resulting connection between structure, prediction, and intervention provides a foundation for a broader theory of fine-tuning outcomes, with the evolution of the kernel defining the central challenge beyond the local regime.

Reproducibility Statement

Code and artifacts will be released publicly upon acceptance. The release contains the forecasting library and the continual-learning harness, the experiment drivers, the per-run artifacts, and the scripts used to produce the reported results. Environments are pinned and seeds are set explicitly in every driver and recorded per run. Core protocols, data splits, and evaluation metrics are summarised in the appendices; full configurations and per-run artifacts will be released with the code.

AI Use Statement

In this work, generative AI tools were used to assist with code development, to refine and rephrase manuscript text, and to help validate mathematical derivations and proofs. All AI-assisted code and mathematical arguments were reviewed and checked by the authors, and all AI-assisted text was revised by the authors before inclusion in the manuscript. Generative AI was not used for research ideation, for selecting research directions, or for formulating the scientific hypotheses and contributions of this work. The authors take full responsibility for the final content of the paper.

Ethics Statement

This work studies how fine-tuning damages capabilities a model already has, using public benchmarks and public pretrained checkpoints, with no human subjects and no new data collection. The forecast identifies which capabilities an adaptation will disturb, which is a dual-use property: it supports protecting safety-relevant behaviour, and the same ranking could in principle indicate which behaviours are easiest to erode. We judge the defensive value to dominate, since the quantities involved are computed from a model and a dataset the adapting party already holds.

References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *ACL-IJCNLP*, pp. 7319–7328, 2021.

Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural networks as kernel learners: The silent alignment effect. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2111.00034.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. LoRA learns less and forgets less. *Transactions on Machine Learning Research (TMLR)*, 2024.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning (ICML)*, 2023.

Blake Bordelon and Cengiz Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2205.09653; also J. Stat. Mech. (2023).

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. In *NeurIPS*, 2020.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K. Dokania, Philip H. S. Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Chandler Davis and William M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Thang Doan, Mehdi Abbana Bennani, Bogdan Mazoure, Guillaume Rabusseau, and Pierre Alquier. A theoretical analysis of catastrophic forgetting through the ntk overlap matrix. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021. arXiv:2010.04003.

Itay Evron, Edward Moroshko, Rachel Ward, Nathan Srebro, and Daniel Soudry. How catastrophic can catastrophic forgetting be in linear regression? In *Conference on Learning Theory (COLT)*, 2022. PMLR vol. 178; arXiv:2205.09588.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *AISTATS*, 2020.

Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M. Roy, and Surya Ganguli. Deep learning versus kernel learning: An empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999.

Daniel Goldfarb, Itay Evron, Nir Weinberger, Daniel Soudry, and Paul Hand. The joint effect of task similarity and overparameterization on catastrophic forgetting — an analytical model. In *International Conference on Learning Representations (ICLR)*, 2024.

Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Hao Gu, Mao-Lin Luo, Zi-Hao Zhou, Han-Chen Zhang, Min-Ling Zhang, and Tong Wei. Spectral imbalance causes forgetting in low-rank continual adaptation, 2026. arXiv:2602.00722.

Ido Nitzan Hidekel and Dan Raviv. Catastrophic forgetting is low-rank: A function-space theory for continual adaptation. In *ICML 2026 Workshop on Continual Adaptation at Scale (CATS)*, 2026. arXiv:2606.18024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *ICLR*, 2023.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Uijeong Jang, Jason D. Lee, and Ernest K. Ryu. LoRA training in the NTK regime has no spurious local minima. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2402.11867.

Xisen Jin and Xiang Ren. Demystifying language model forgetting with low-rank example associations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2406.14026.

Ryo Karakida and Shotaro Akaho. Learning curves for continual learning in neural networks: Self-knowledge transfer and forgetting. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2112.01653.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences (PNAS)*, 114(13):3521–3526, 2017.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning (ICML)*, 2017.

Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. Understanding catastrophic forgetting in language models via implicit inference. In *International Conference on Learning Representations (ICLR)*, 2024.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Sebastian Lee, Sebastian Goldt, and Andrew Saxe. Continual learning in the teacher-student setup: Impact of task similarity. In *ICML*, 2021.

Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. Revisiting catastrophic forgetting in large language model tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4297–4308, 2024. doi: 10.18653/v1/2024.findings-emnlp.249.

Zhizhong Li and Derek Hoiem. Learning without forgetting. In *European Conference on Computer Vision (ECCV)*, 2016.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 33:3776–3786, 2025. arXiv:2308.08747.

Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the SDEs and scaling rules for adaptive gradient algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2205.10287.

Sadhika Malladi, Alexander Wettig, Dingli Yu, Danqi Chen, and Sanjeev Arora. A kernel-based view of language model fine-tuning. In *International Conference on Machine Learning (ICML)*, 2023. arXiv:2210.05643.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research (JMLR)*, 21:1–76, 2020.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pp. 109–165. Academic Press, 1989.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Dilan Gorur, Razvan Pascanu, and Hassan Ghasemzadeh. Linear mode connectivity in multitask and continual learning. In *ICLR*, 2021.

Seyed Iman Mirzadeh, Arslan Chaudhry, Dong Yin, Huiyi Hu, Razvan Pascanu, Dilan Gorur, and Mehrdad Farajtabar. Wide neural networks forget less catastrophically. In *International Conference on Machine Learning (ICML)*, pp. 15699–15717, 2022. PMLR 162.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Pingbo Pan, Siddharth Swaroop, Alexander Immer, Runa Eschenhagen, Richard E. Turner, and Mohammad Emtiyaz Khan. Continual deep learning by functional regularisation of memorable past. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. arXiv:2004.14070.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *International Conference on Learning Representations (ICLR)*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Zahra Rahimi Afzal, Tara Esmaeilbeig, Mojtaba Soltanalian, and Mesrob I. Ohannessian. Linearization explains fine-tuning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025. doi: 10.52202/085713-4388.

Vinay V. Ramasesh, Ethan Dyer, and Maithra Raghu. Anatomy of catastrophic forgetting: Hidden representations and task semantics. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2007.07400.

Vinay V. Ramasesh, Aitor Lewkowycz, and Ethan Dyer. Effect of scale on catastrophic forgetting in neural networks. In *International Conference on Learning Representations (ICLR)*, 2022.

Tim G. J. Rudner, Freddie Bickford Smith, Qixuan Feng, Yee Whye Teh, and Yarin Gal. Continual learning via sequential function-space variational inference. In *ICML*, 2022.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Thomas Scialom, Tuhin Chakrabarty, and Smaranda Muresan. Fine-tuned language models are continual learners. In *EMNLP*, pp. 6107–6122, 2022.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual learning of large language models: A comprehensive survey. *arXiv preprint arXiv:2404.16789*, 2024.

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

Brady Steele. Subspace geometry governs catastrophic forgetting in low-rank adaptation, 2026. arXiv:2603.02224.

Michalis K. Titsias, Jonathan Schwarz, Alexander G. de G. Matthews, Razvan Pascanu, and Yee Whye Teh. Functional regularisation for continual learning with gaussian processes. In *International Conference on Learning Representations (ICLR)*, 2020. arXiv:1901.11356.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *EMNLP Workshop BlackboxNLP*, 2018.

Yuanyi Wang, Yifan Yang, Su Lu, Yanggan Gu, Pengkai Wang, Wenjun Wang, Zhaoyi Yan, Congkai Xie, Jianmin Wu, Jialun Cao, Shing-Chi Cheung, and Hongxia Yang. Geometry conflict: Explaining and controlling forgetting in llm continual post-training, 2026.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *CVPR*, 2022.

Mutian Yang, Zisen Zhan, Yutong Chen, Haolin Li, Kaiwen Wang, Kaili Zheng, Yuguang Wang, Qi Wang, Jiandong Gao, and Ji Wu. Learning the mechanism of catastrophic forgetting: A perspective from gradient similarity, 2026.

Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *ICML*, 2017.

A Proofs and Minimal Structural Checks

Gauss–Newton push-through. For Proposition 2,

$$\delta\theta^\star = -(J_B^\top H_B J_B + \lambda I)^{-1} J_B^\top g_B, \qquad \Delta z_A = J_A \delta\theta^\star.$$

Using

$$(J_B^\top H_B J_B + \lambda I)^{-1} J_B^\top = J_B^\top (H_B K_{BB} + \lambda I)^{-1}$$

and $K_{AB} = J_A J_B^\top$ gives Eq. (3). Setting $H_B = I$ recovers the curvature-free ridge operator. Since the nonzero eigenvalues of $H_B K_{BB}$ coincide with those of $H_B^{1/2} K_{BB} H_B^{1/2}$, $H_B K_{BB} + \lambda I$ is invertible for every $\lambda > 0$.

Proof of Proposition 3. Let $J_B = U\Sigma V^\top$. Then $K_{AB} = J_A V \Sigma U^\top$ and $(K_{BB} + \lambda I)^{-1} = U(\Sigma^2 + \lambda I)^{-1} U^\top$, which gives the lazy filter $\sigma/(\sigma^2+\lambda)$. As $H_B \to 0$, $(H_B K_{BB} + \lambda I)^{-1} \to \lambda^{-1} I$, giving the confident-limit filter σ/λ.

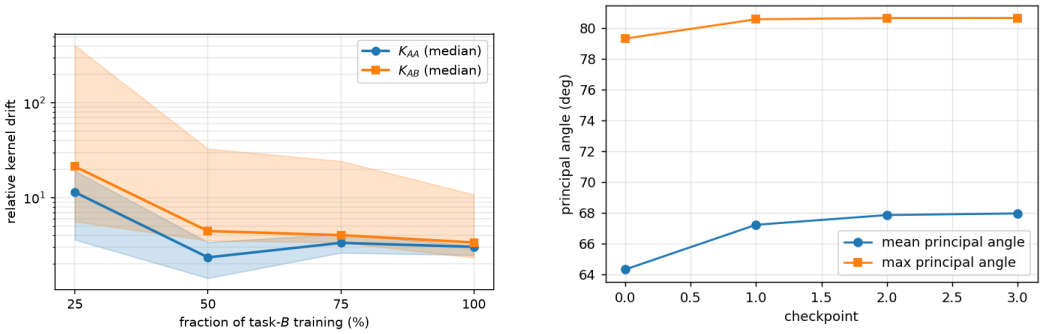

Figure 4: Kernel movement under partial fine-tuning on Split-CIFAR-10. Left: relative Frobenius kernel drift. Right: angle between the top-k eigenspace at θ_A and later checkpoints.

Table 3: Continual-learning comparison on Split-CIFAR-100 (5 seeds). Accuracy is higher-is-better and forgetting lower-is-better. Output-space methods use the same 200 anchors per task; the bottom rows compose the spectral penalty with replay.

		ResNet-50		CLIP ViT-B/32	
method	track (memory)	acc↑	fgt↓	acc↑	fgt↓
ER	replay (M=2000)	.620±.011	.210±.015	**.693**±.009	.187±.015
DER++	replay (M=2000)	.583±.011	.107±.004	.678±.007	.176±.007
spectral	output-space, top-k	.632±.007	.135±.007	.624±.008	.271±.014
full output	output-space, all modes	.627±.005	.239±.011	.627±.009	.291±.017
FROMP	output-space, precision wt.	.305±.024	.654±.019	.367±.013	.603±.018
LwF	output distillation	.254±.015	.328±.015	.262±.023	.568±.018
EWC	parameter-space	.285±.021	.676±.019	.407±.017	.553±.014
GPM	gradient subspace	**.641**±.009	.185±.011	**.652**±.009	.195±.018
no-reg	none	.266±.016	.701±.011	.293±.024	.692±.023
ER + spectral	composition	**.635**±.005	.111±.010	.683±.011	.163±.005
DER++ + spectral	composition	.565±.009	**.069**±.008	.660±.011	.179±.013

Proof of Lemma 1. For $w(t) = K_{AB}(t)g_B(t)$,

$$(I - P_k(0))w = (I - P_k(0))P_k(t)w + (I - P_k(0))(I - P_k(t))w.$$

The first term is at most $\sin\Theta_k(t)\|w(t)\|$ and the second at most ε_t. Integrating and using $\Delta z_A(T) = -\int_0^T w(t)\,dt$ proves the lemma. The Davis–Kahan variant of Yu et al. (2015) gives

$$\sin\Theta_k(t) \le \frac{2\min\{\sqrt{k}\|\Delta K_{AA}(t)\|_{\mathrm{op}}, \|\Delta K_{AA}(t)\|_F\}}{\gamma_k(0)}.$$

At the measured partial-fine-tuning operating point, the instantaneous tail is small (0.08–0.11 of $\|w\|$) while the top-k subspace rotates strongly, so rotation is the dominant observed source of leakage.

Concentration controls. On the Split-CIFAR-10 experiment of Figure 2, the leading ten directions contain 0.9909 of realised drift energy versus 0.0486 for a matched-dimension Haar-random subspace. This concentration is not universal: on the five frozen Split-CIFAR-100 backbones, ten of 2000 directions contain only 0.0255 ± 0.0161, and reaching 90% requires 963 ± 186 directions. We therefore use the eigenspace ordering, not universal low rank, as the general claim.

B Regulariser Experiments

Protocol. All continual-learning results are class-incremental with a shared head and no task ID at test. We report final average accuracy and standard average forgetting. Split-CIFAR-100 uses ten tasks; frozen-backbone experiments train only the linear head, while the LoRA experiments adapt rank-8 modules. Baselines are tuned over fixed grids; full configurations will be released with the experimental artifacts. The penalty is evaluated as a structural intervention, not as a state-of-the-art continual-learning method.

Language-model anchor protocol. We apply Eq. (5) during rank-4 LoRA fine-tuning of Qwen2.5-1.5B-Instruct on mathematics or computer-science MMLU targets, using 56 anchor questions disjoint from evaluation. We sweep $\mu \in \{0.01, 0.1, 1\}$ and require task-B train CE ≤ 0.10

Table 4: LoRA continual-learning comparison on Split-CIFAR-100 (5 seeds). The all-modes anchor uses the complete anchor-estimable output space; top-12 restricts it to the leading modes. † marks settings that fail to learn in at least one column.

method	track (memory)	LoRA ViT acc ↑ shared lr	LoRA ViT acc ↑ own lr	fgt ↓ own lr	CLIP-LoRA acc ↑
DER++	replay (M=1000)	0.736±.013	0.736	0.177±.019	0.420±.033
DER++ + output anch.	composition	**0.745**	n/a	**0.124**±.010	n/a
output anchor	output-space, all modes	**0.682**±.022	**0.707**±.010	**0.092**±.014	**0.322**±.035
output anchor, top-12	output-space, top-k	n/a	0.706±.009	0.124±.017	n/a
EWC†	parameter-space	0.148±.151	0.626±.007	0.081±.014	0.013±.004
no-reg†	none	0.083±.062	0.524±.019	0.345±.024	0.043±.029

Table 5: Language-model anchoring on Qwen2.5-1.5B-Instruct. Entries are mean held-out MMLU CE increase at 64 steps; lower is better. Both penalised arms use the same 56 anchors and $\mu = 0.1$ (3 seeds).

task B	optimizer	no penalty	top-k	full output
mathematics	SGD	+1.144±.059	+0.145±.018	+0.164±.027
computer sci.	SGD	+1.269±.084	+0.312±.174	+0.212±.020
mathematics	AdamW	+1.840±.384	+0.176±.024	+0.160±.039
computer sci.	AdamW	+1.866±.232	+0.214±.024	+0.182±.033

and task-A accuracy within 0.03 of the unpenalised run. At the best qualifying setting, pooled ΔCE drops from $+1.530$ to $+0.212$ (86%); recomputing over all evaluation subjects gives 85.5%. The effect is primarily calibration: task-A accuracy changes by less than one point, and full-output anchoring performs similarly after tuning.

Table 6: Matched-rank subspace control at language-model scale. Anchors, rank $k = 112$, and $\mu = 0.01$ are fixed; only the protected projector changes. Entries pool 12 configuration–seed pairs.

anchored subspace	ΔCE at 64 steps	damage removed
none	+1.530±.400	n/a
top-k of K_{AA}	**+0.309**±.090	79.8%
full output	+0.380±.120	75.2%
random rank-k	+0.432±.114	71.8%
largest diag K_{AA}	+0.505±.120	67.0%
bottom-k of K_{AA}	+1.028±.150	32.8%

Matched-rank control. Fixing anchors, rank and penalty strength while changing only the projector yields the ordering in Table 6 in all 12 paired cells. The top eigenspace outperforms random, diagonal, and bottom-eigenspace controls, but the complete-basis arm still removes 75.2% of the damage. Thus the subspace matters, while anchoring itself remains the dominant intervention.

C Ridge and Curvature Controls

Oracle scalar control. On the five frozen backbones, the curvature-aware predictor exceeds a per-transition oracle scalar $H_B = \alpha I$ by $+0.0116$ to $+0.0155$ in drift cosine. This is only 13–18% of its larger margin over the uncalibrated curvature-free reference; the remainder is attributable to ridge scale. The operator stable ranks in Table 1 give $\mathrm{sr}(T_{\mathrm{lazy}})/\mathrm{sr}(T_{\mathrm{GN}}) \approx 1.4$–$1.7$ across the five backbones.

Fixed-ridge evaluation. All main-text frozen-regime numbers use $\lambda = 1$ fixed a priori. Rerunning the controlled networks with this fixed value reproduces the swept results to within the reported precision, and choosing another fixed λ changes the Section 6.1 cosine by at most 0.005.

D Language-Model Experiments

Setup. The base configuration is Qwen2.5-1.5B-Instruct with rank-4 LoRA on query/value projections. The vector forecast is scalarised as $g_A^\top \Delta z_A^{\mathrm{GN}}$ per probe and averaged within MMLU subject. We use $\lambda = c\,\mathrm{tr}(K_{BB})/m_B$: Qwen development cells choose c on a disjoint fit split, while the held-out model families use the frozen value $c = 1$. Evaluation always uses disjoint subjects.

Table 7: Four-step capability-ranking results. Entries are Spearman correlations with realised per-subject CE change on 20 held-out MMLU subjects (5 seeds per configuration). The lower block contains model families not used to develop the predictor.

adapt.	opt.	forecast	raw transp.	Adam-metric	blind marg.	base CE	grad-sim	ΔCE
0.5B→math	SGD	$+0.56\pm.23$	$+0.54\pm.27$	$+0.54\pm.25$	$+0.02\pm.28$	$-0.06\pm.28$	$-0.52\pm.33$	$+\mathbf{.026}$
0.5B→math	AdamW	$+0.86\pm.05$	$+0.85\pm.06$	$+0.93\pm.05$	$-0.29\pm.22$	$-0.38\pm.34$	$-0.80\pm.10$	$+\mathbf{.003}$
1.5B→math	SGD	$+0.60\pm.38$	$+0.60\pm.38$	$+0.62\pm.35$	$-0.32\pm.21$	$-0.42\pm.30$	$-0.43\pm.34$	$-.049$
1.5B→math	AdamW	$+0.79\pm.12$	$+0.79\pm.12$	$+0.92\pm.09$	$-0.58\pm.12$	$-0.71\pm.14$	$-0.55\pm.18$	$-.064$
3B→math	SGD	$+0.96\pm.02$	$+0.96\pm.02$	$+0.89\pm.10$	$-0.32\pm.18$	$-0.42\pm.17$	$-0.75\pm.08$	$-.002$
3B→math	AdamW	$+0.88\pm.11$	$+0.88\pm.11$	$+0.94\pm.03$	$-0.31\pm.15$	$-0.41\pm.16$	$-0.71\pm.12$	$-.005$
1.5B→CS	SGD	$+0.69\pm.17$	$+0.69\pm.17$	$+0.63\pm.16$	$-0.35\pm.22$	$-0.37\pm.25$	$-0.42\pm.18$	$-.037$
1.5B→CS	AdamW	$+0.75\pm.15$	$+0.75\pm.15$	$+0.91\pm.03$	$-0.53\pm.18$	$-0.62\pm.22$	$-0.60\pm.12$	$-.056$
L-1B→math	SGD	$+0.92\pm.05$	$+0.92\pm.06$	$+0.86\pm.10$	$-0.27\pm.27$	$-0.39\pm.25$	$-0.82\pm.06$	$-.020$
L-1B→math	AdamW	$+0.96\pm.03$	$+0.97\pm.01$	$+0.97\pm.02$	$-0.42\pm.18$	$-0.55\pm.19$	$-0.86\pm.05$	$-.036$
L-3B→math	SGD	$+0.96\pm.03$	$+0.96\pm.02$	$+0.83\pm.09$	$-0.25\pm.26$	$-0.23\pm.23$	$-0.75\pm.05$	$-.005$
L-3B→math	AdamW	$+0.77\pm.07$	$+0.77\pm.07$	$+0.90\pm.05$	$-0.20\pm.16$	$-0.21\pm.17$	$-0.65\pm.08$	$-.006$
L-3B→CS	SGD	$+0.96\pm.03$	$+0.96\pm.02$	$+0.83\pm.07$	$+0.14\pm.07$	$+0.32\pm.10$	$-0.61\pm.17$	$-.006$
L-3B→CS	AdamW	$+0.87\pm.07$	$+0.87\pm.06$	$+0.96\pm.01$	$+0.14\pm.16$	$+0.31\pm.15$	$-0.57\pm.16$	$+\mathbf{.008}$
M-7B→math	SGD	$+0.14\pm.17$	$+0.17\pm.22$	$+0.21\pm.25$	$+0.11\pm.54$	$+0.06\pm.54$	$-0.08\pm.27$	$+\mathbf{.175}$
M-7B→math	AdamW	$+0.63\pm.17$	$+0.59\pm.18$	$+0.63\pm.20$	$+0.00\pm.08$	$-0.06\pm.22$	$-0.35\pm.12$	$+\mathbf{.066}$
Phi→math	SGD	$+0.83\pm.12$	$+0.77\pm.19$	$+0.85\pm.05$	$-0.53\pm.14$	$-0.68\pm.08$	$-0.52\pm.16$	$-.116$
Phi→math	AdamW	$+0.85\pm.10$	$+0.80\pm.16$	$+0.96\pm.02$	$-0.63\pm.15$	$-0.83\pm.10$	$-0.54\pm.23$	$-.156$

Main grid. Qwen2.5-0.5B/1.5B/3B are evaluated on mathematics and computer-science targets with SGD and AdamW. Twenty MMLU subjects are held out for evaluation and five seeds are used per configuration. The same protocol is then transferred without retuning to Llama-3.2-1B/3B, Mistral-7B-Instruct, and Phi-3.5-mini.

Table 8: Precision@k for identifying the realised most-damaged capabilities among 20 held-out subjects. Ninety cells; chance is $k/20$.

ranker	precision@3	precision@5
forecast (ours)	$\mathbf{.619}\pm.281$	$\mathbf{.711}\pm.214$
base CE	$.104\pm.171$	$.162\pm.192$
entropy	$.104\pm.178$	$.218\pm.181$
gradient norm	$.148\pm.207$	$.207\pm.207$
gradient similarity	$.030\pm.095$	$.058\pm.109$
gradient similarity, sign-flipped	$.437\pm.267$	$.536\pm.194$

Most-damaged capabilities. Across the 90 four-step cells, the forecast reaches precision@5 0.711 ± 0.214. The sign-flipped gradient-similarity oracle reaches 0.536 ± 0.194; canonical heuristics are below chance or near it in their natural orientation.

Overlap baselines and gradient-norm control. NTK-overlap and CKA baselines remain far below the forecast on both Qwen and held-out families (Table 9). Partialling out the per-capability gradient norm changes the transport ranking from $+0.768$ to $+0.757$; hard norm matching leaves $+0.746$ with two strata and $+0.701$ with four. Thus the result is not explained by gradient norm alone.

Optimizer metric. The Adam-metric score improves the ordinary forecast under AdamW at the earliest horizons and degrades it under SGD, as predicted by the metric-specific transport. The AdamW advantage disappears by 16 steps, consistent with expiry of the local approximation. **Preconditioner controls.** Permuting the Adam second-moment diagonal across parameters preserves its marginal distribution but eliminates the gain over the Euclidean forecast (0/10 wins). Dispersion-matched random diagonals and Adam-shaped diagonals estimated on unrelated inputs likewise fail to reproduce the optimizer-specific improvement.

Locality and held-out families. At four steps, forecast quality decreases as realised movement grows: across the 18 configurations, ρ correlates at -0.589 with $|\Delta\mathrm{CE}|$. Predicted movement magnitude computed at θ_A ranks realised $|\Delta\mathrm{CE}|$ at $+0.925$. Across the 50 held-out-family cells, the pooled four-step forecast correlation is $+0.79 \pm 0.26$; Mistral-7B under SGD is the weakest configuration and also the most strongly moved.

Raw transport and prediction horizon. On the 40 Qwen cells, the calibrated Gauss–Newton score and raw transport are effectively identical ($+0.76 \pm .23$ versus $+0.76 \pm .24$), so the language-

Table 9: Non-gradient baselines on the same four-step capability-ranking task. Each baseline is shown in the orientation that performs better; wins counts cells whose correlation exceeds the forecast.

	Qwen (40 cells)			held-out families (50 cells)		
ranker	ρ	wins	p	ρ	wins	p
forecast	$+.760 \pm .229$	n/a	n/a	$+.789 \pm .256$	n/a	n/a
NTK overlap, Frobenius	$+.076 \pm .238$	1	$1.3\text{e}{-11}$	$+.062 \pm .208$	2	$8.9\text{e}{-15}$
NTK overlap, operator	$+.120 \pm .167$	1	$5.5\text{e}{-12}$	$+.132 \pm .199$	2	$1.2\text{e}{-14}$
NTK projection of g_B	$+.077 \pm .247$	1	$1.8\text{e}{-11}$	$+.059 \pm .221$	2	$9.8\text{e}{-14}$
CKA, eNTK features	$+.078 \pm .160$	1	$3.6\text{e}{-12}$	$+.076 \pm .205$	3	$4.4\text{e}{-14}$
CKA, hidden states	$-.000 \pm .231$	1	$2.5\text{e}{-11}$	$+.073 \pm .189$	0	$1.8\text{e}{-15}$

Table 10: Adam-metric forecast versus the ordinary forecast by horizon. Entries give cells where the preconditioned score leads, mean $\Delta\rho$, and a paired two-sided t test.

		AdamW			SGD		
grid	steps	leads	mean $\Delta\rho$	p	leads	mean $\Delta\rho$	p
held-out (n=25)	1	23/25	$+0.106$	$4.0\text{e}{-6}$	1/25	-0.098	$2.7\text{e}{-7}$
held-out (n=25)	4	19/25	$+0.069$	$1.4\text{e}{-3}$	4/25	-0.047	$7.4\text{e}{-2}$
Qwen (n=20)	4	18/20	$+0.103$	$4.9\text{e}{-4}$	7/20	-0.030	$2.2\text{e}{-1}$
Qwen (n=20)	16	10/20	$+0.009$	$7.5\text{e}{-1}$	6/20	-0.096	$3.7\text{e}{-3}$

model result primarily supports transport rather than curvature. Across the dense step grid, pooled forecast correlation falls from $+0.907$ at one step to $+0.184$ at 16 steps and becomes negative thereafter, while prior competence rises in the opposite direction; they cross between 8 and 16 steps.

Large-damage and boundary checks. On Pythia-1.4B with SST-2 probes and 200-step LoRA fine-tuning toward QQP, RTE or QNLI, final-step Spearman is 0.988–0.997 and the predicted margin separates flipped from surviving examples at AUROC 0.98 by 16 steps. Under deliberately aggressive Qwen fine-tuning, held-out subjects lose roughly 40 accuracy points and prior competence becomes the better ranker; this is the measured boundary regime discussed in Section 6.4.

D.1 Preconditioned transport

For $\dot\theta = -P(t)\nabla_\theta L_B$ with symmetric PSD P,

$$\dot z_A = -J_A P J_B^\top g_B = -K^P_{AB} g_B, \qquad K^P_{XY} := J_X P J_Y^\top.$$

Writing $\Phi_X = J_X P^{1/2}$ reduces the ridge Gauss–Newton problem in the P^{-1} metric to the Euclidean case, yielding

$$\Delta z_A^{\mathrm{GN,P}} = -K^P_{AB}(H_B K^P_{BB} + \lambda I)^{-1} g_B.$$

For Adam, the initial diagonal preconditioner is estimated from task-B squared gradients. Freezing it at θ_A ignores later moment evolution, momentum, discreteness, and weight decay; Eq. (4) is therefore an empirically tested local predictor rather than an exact model of AdamW.

E Partial Fine-Tuning

We unfreeze the last transformer block, `ln_post`, and visual projection of CLIP ViT-B/32 and fine-tune on one Split-CIFAR-100 task. On 200 held-out CIFAR-10 probes, raw embedding-drift cosine is 0.948 ± 0.012, but 96.6% of realised drift energy lies in one shared direction. After removing that component, cosine is 0.224 ± 0.100, still above a matched-norm random floor. The forecast ranks zero-shot damage at 0.714 ± 0.118 per point, but a constant-direction oracle is not statistically worse ($p = 0.31$); hence the main text does not claim capability-ranking superiority in this regime.

A depth ladder over one, two, and four unfrozen ViT-B/32 blocks plus a ViT-B/16 control preserves a positive demeaned-cosine margin over random in all 20 cells, while the constant-direction oracle remains competitive on ranking. This experiment is therefore used only as evidence that directional tangent information survives moderate feature learning.