

Merge schedules and Fisher-proxy scale in continual model merging

A diagnostic audit in a linearized pipeline: fixed 0.4, uniform $1/t$, and unit-mean scaling

Basel Elshenawy^{1,2} | Abdelkader Baggag² | QCRI Summer Internship Program 2026

¹ Ulster University ² Qatar Computing Research Institute, HBKU

THE RESULT

stability vs. plasticity

Uniform $1/t$ vs. Fixed 0.4: 3.4× less forgetting, but lower learning accuracy.

Final averages hide the trade-off. Split CIFAR-100 | CLIP ViT-B/16 | Main grid: 27 runs

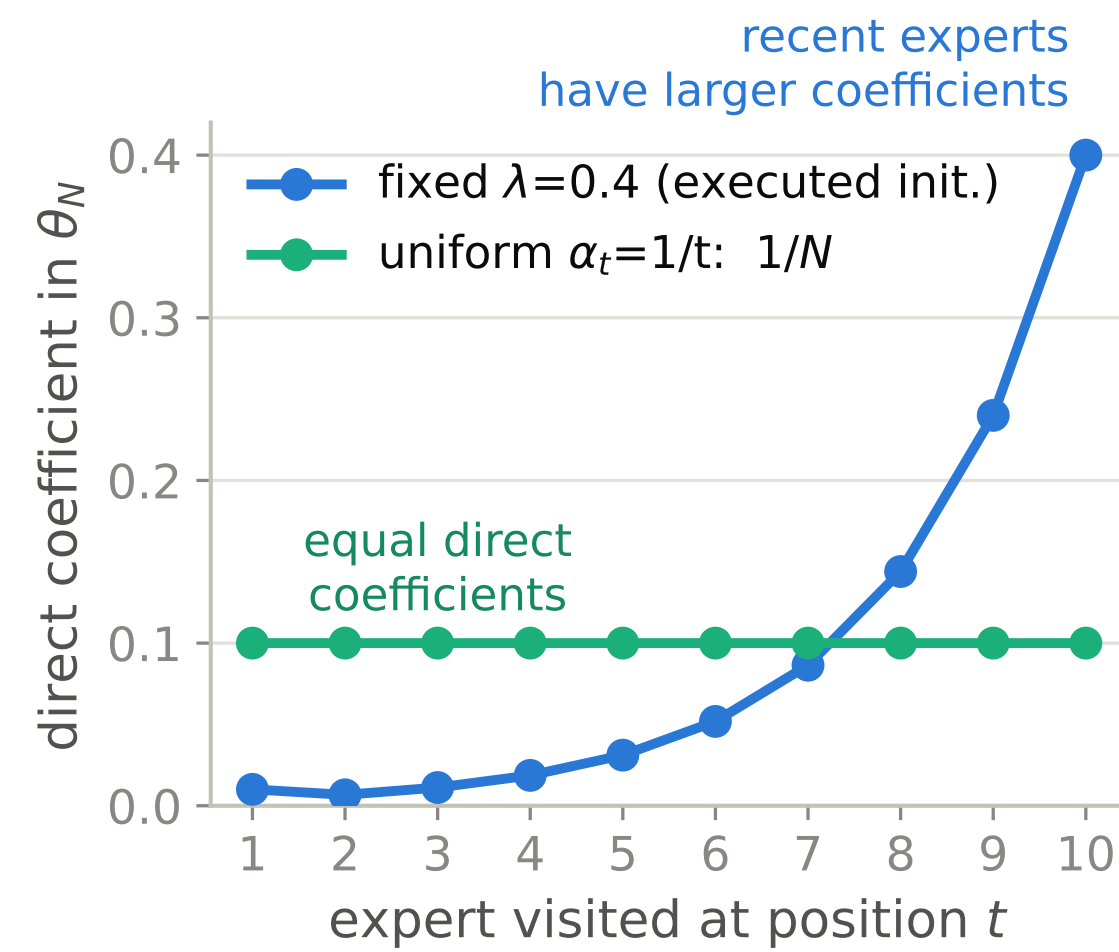
1 Schedules encode recency

Isolate the schedule: omit Fisher weighting and alignment from this coefficient analysis.

$$\theta_t = (1 - \alpha_t)\theta_{t-1} + \alpha_t\theta_t^*$$

Why Fixed favors recent checkpoints

- Initialize $\theta_1 = \theta_1^*$; merge $N = 10$ checkpoints.
- Fixed $\alpha_t = 0.4$: $w_1 = 0.6^9$ and $w_t = 0.4(0.6)^{10-t}$ for $t \geq 2$.
- First/last checkpoint weight: $w_1/w_{10} = 2.52\%$.



Uniform gives each checkpoint weight $1/N$.
These coefficients are not causal effects of task data.

Where the $1/t$ identity applies

Before alignment, $1/t$ recovers the joint Fisher-weighted merge. Post-merge alignment breaks this identity.

EXPERIMENT AT A GLANCE

Training: ten 10-class tasks; AdamW; 10 epochs/task; Fisher merging + L1 alignment.

Policies: Fixed/raw, Uniform/raw, Uniform/unit mean (Fisher-proxy scaling).

Design: 3 shared seeds $\times$ 3 orders per policy; 27 runs. Descriptive results: mean $\pm$ population SD (standard deviation).

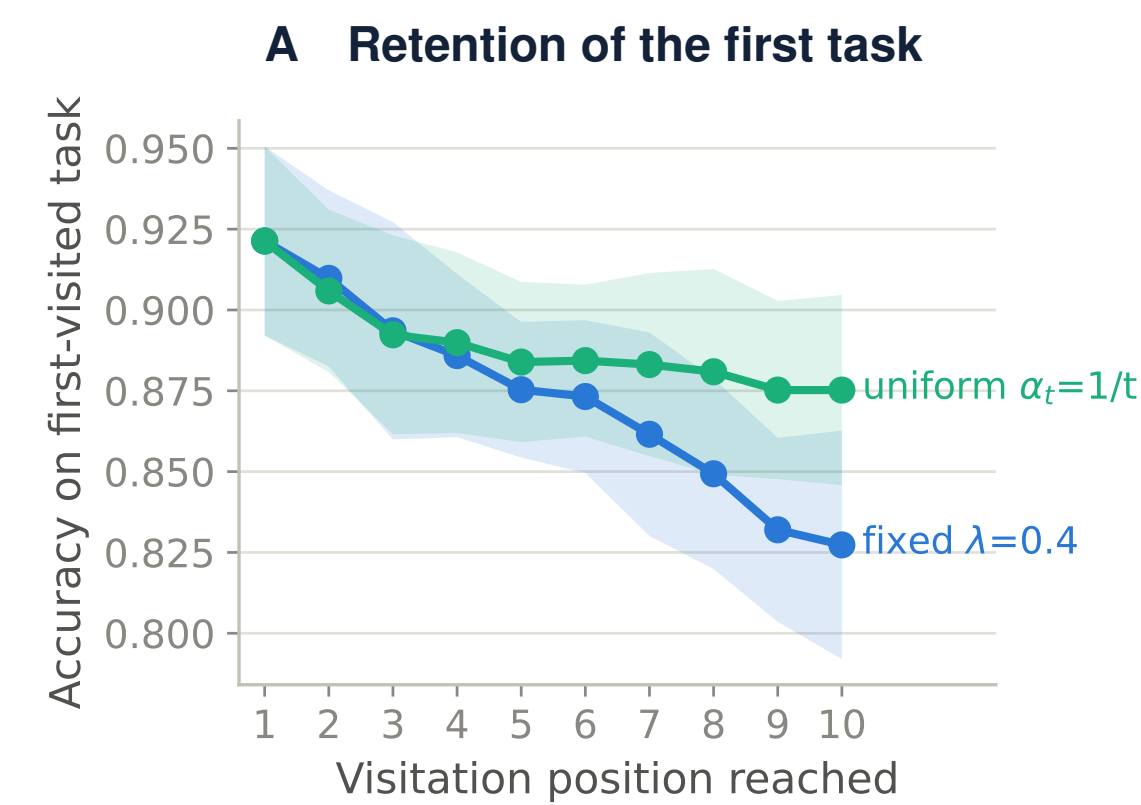
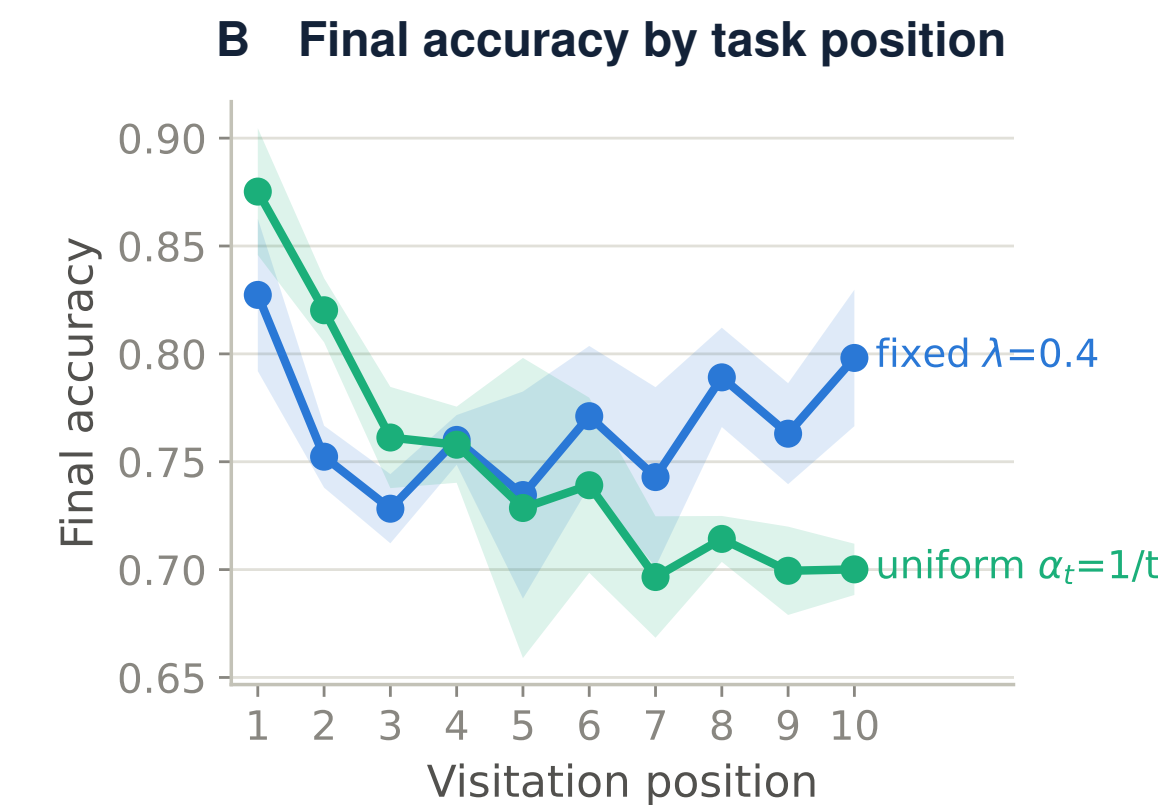
2 Less forgetting comes at a learning cost

Fixed/raw $\rightarrow$ Uniform/raw | Matched 3-seed $\times$ 3-order grid

3.4×
lower forgetting (FM)
0.054 $\rightarrow$ 0.016

-5.6 pp
learning accuracy (LA)
accuracy when first learned

-1.7 pp
final average (A_N)
accuracy after all 10 tasks

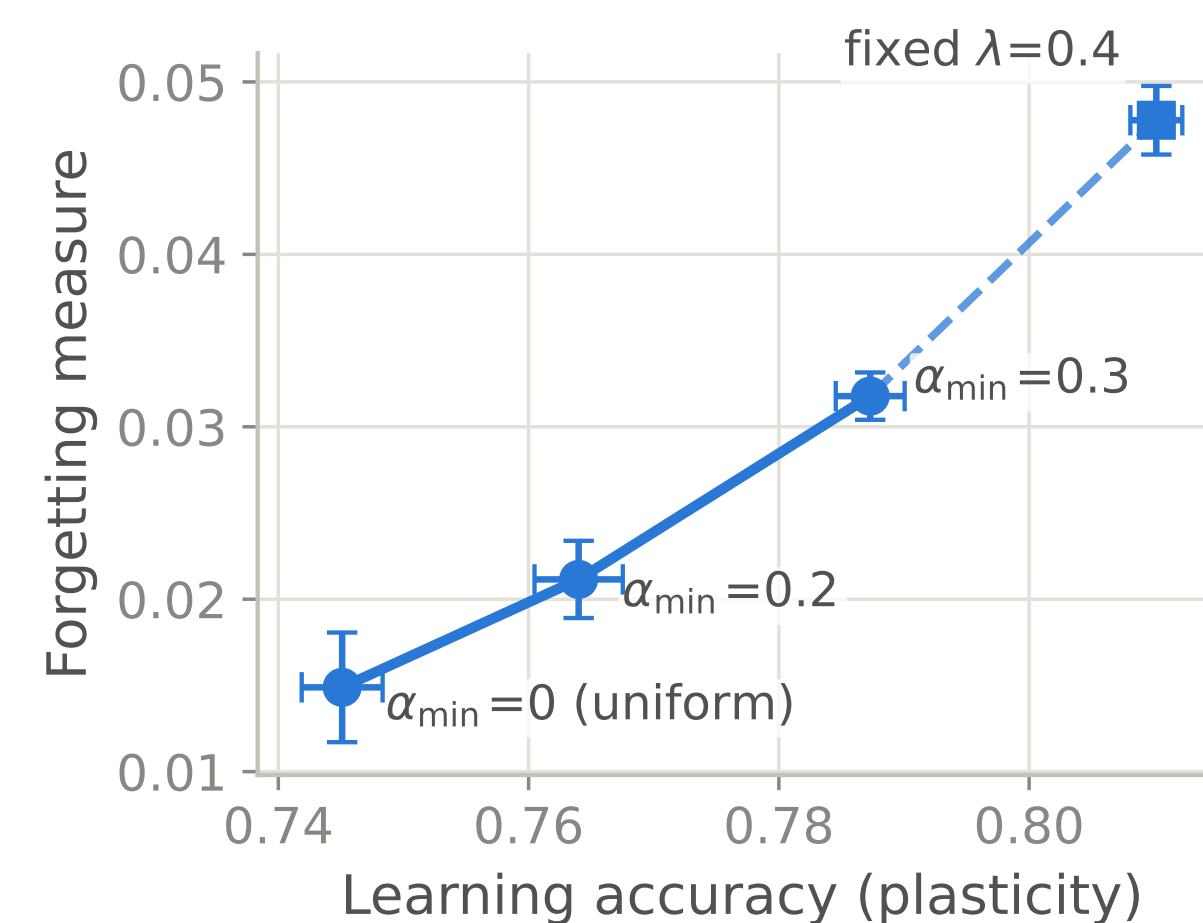


Final accuracy, Uniform vs. Fixed: +4.8 **pp** on the first task; mean +2.8 **pp** at positions 1–5 and -6.3 **pp** at 6–10.

Uniform lowers FM and A_N in all 9 matched pairs. Across-order SD falls for FM (0.0045 $\rightarrow$ 0.0020), but rises for A_N and LA: **less order sensitivity for FM only**.

3 Schedule floors give intermediate trade-offs

$\alpha_t = \max(1/t, \alpha_{\min})$ | Identity task order; 3 seeds



Policy	FM $\downarrow$	LA $\uparrow$	A_N $\uparrow$
$1/t$	0.0149	0.7451	0.7395
Floor 0.2	0.0211	0.7640	0.7518
Floor 0.3	0.0318	0.7873	0.7632
Fixed 0.4*	0.0478	0.8102	0.7695

Higher floors improve learning, but increase forgetting.

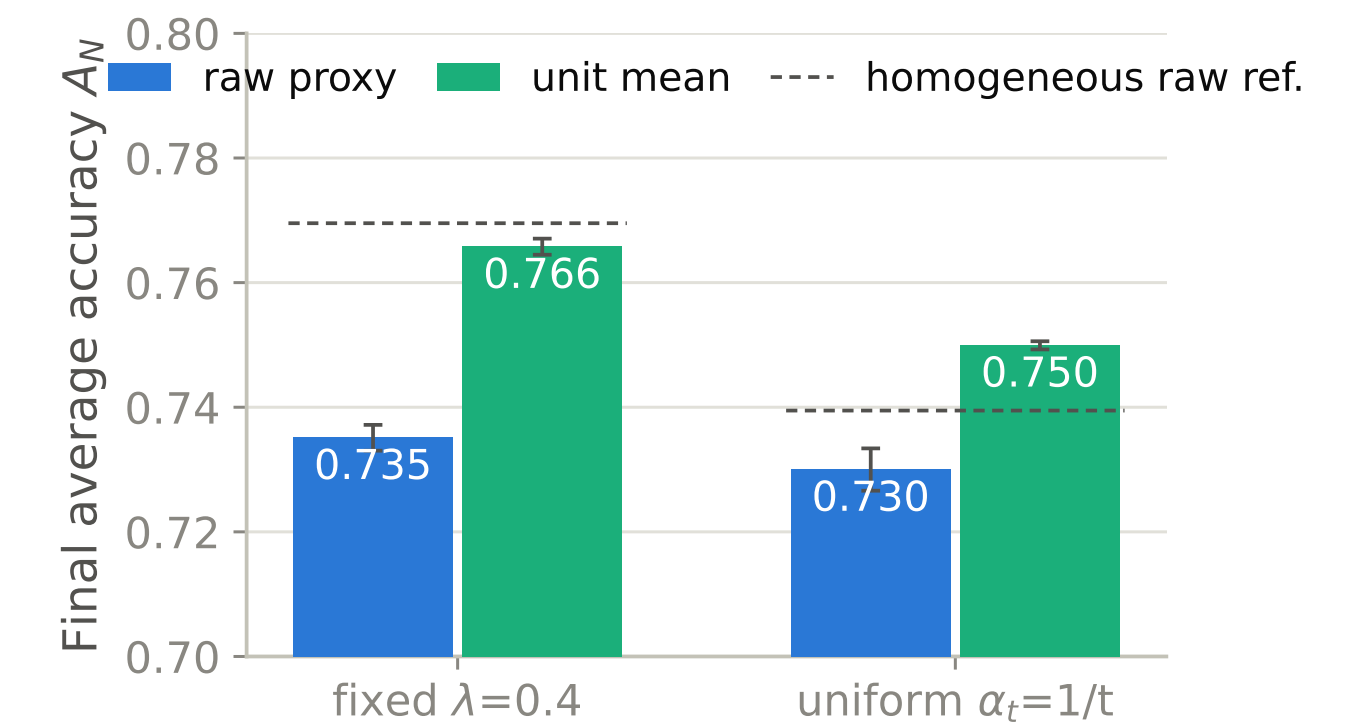
*Fixed 0.4 differs at early steps. These points establish neither a smooth response nor an optimal default.

4 Scaling shifts the balance

Stress test: alternate 4 and 16 training epochs across tasks. Normalize each Fisher proxy to unit mean:

$$\hat{H}_t = H_t / \text{mean}_j H_{t,j}$$

Identity task order; 3 seeds



Fixed: A_N 0.7351 $\rightarrow$ 0.7658; FM 0.0389 $\rightarrow$ 0.0497
Uniform: A_N 0.7300 $\rightarrow$ 0.7499; FM 0.0142 $\rightarrow$ 0.0161

Higher A_N and LA come with **higher forgetting**. No general recommendation follows.

5 Accuracy can hide forgetting



Naive fine-tuning vs. Uniform (3 seeds; identity order): +0.4 pp A_N , +12.0 pp LA, but **9.1×** higher FM.

REPORT THE FULL STORY

- Report A_N , LA, FM, backward transfer, and task-wise accuracy.
- State schedules and alignment.
- Use validation or a stated objective.

CODE

