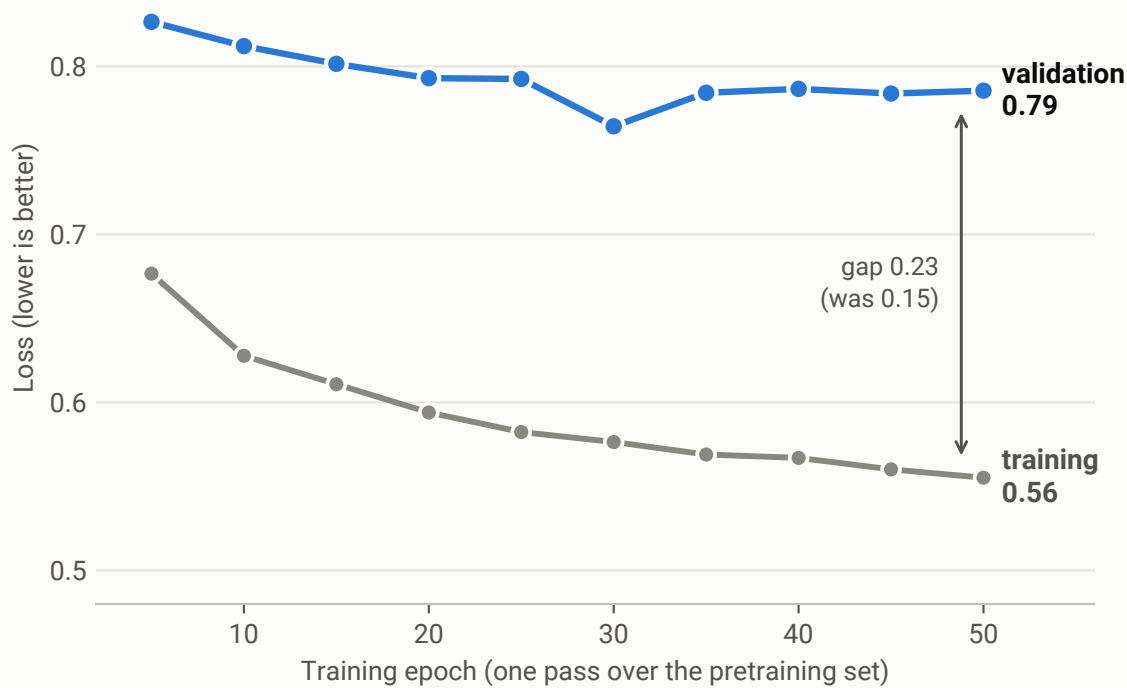


# More pretraining hours did not move downstream accuracy

0.66k to 3.3k hours (4.9×) changed AUC by -0.03 over 19 tasks. Validation loss flattens while training loss keeps falling, which points to a data limit.

## Pretraining loss: validation flattens, training keeps falling

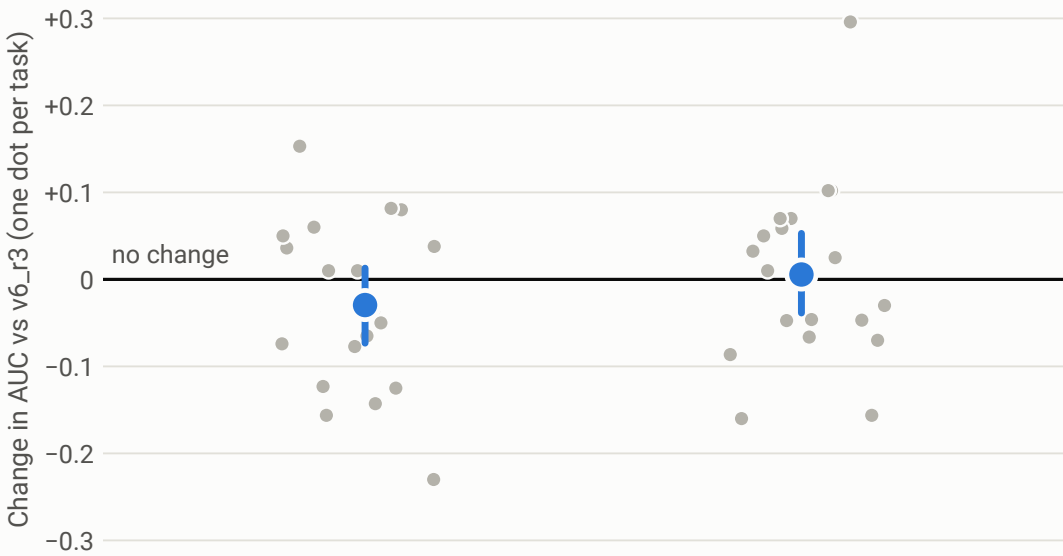
d10 small model (29M parameters), same validation files every epoch



The 3× larger model (85M) bottoms out at 0.75 (epoch 8); the best small-model value is 0.73 (epoch 32). More parameters did not lower the loss, and downstream AUC is unchanged.

## Downstream accuracy: no detectable change

d10 vs v6\_r3, 19 tasks, fixed held-out cohorts



d10 (production)  
29M parameters

**-0.03 AUC**

95% CI -0.07 to +0.01

9 tasks up · 10 down

d10-M  
85M parameters

**+0.01 AUC**

95% CI -0.04 to +0.05

10 tasks up · 9 down

Right: AUC is the best of XGBoost and TabPFN per task; test sets are small (14 to 40 subjects for most), so effects under about 0.04 AUC cannot be seen. 5 retired or quarantined tasks, SRM (corrupt source EEG), sex tasks and regression tasks are excluded. The added hours are mostly child and epilepsy EEG. Left: epochs 5 to 50 of the production run; the checkpoint with the lowest validation loss (epoch 32, 0.73) is not the one used downstream.