When paired runs stop being alike

> Two runs at one seed share their initialization, batches, and labels. Started together under two recipes, they stay partly alike in the task score° through epoch 64, whether the recipes differ a little or a lot, but they are only weakly alike in removal° and no more alike than chance in spill°. Carried on from one stored state°, they stay alike in all three scores for hundreds of epochs.

Task score. Expected exact match on the held-out probe contexts, averaged over ops, recorded every four epochs during training.…

Removal. The drop on `difference` under the full e_1 edit, applied at every position and slice.…

Spill. The largest net drop on any other op, under the same edit.

From one stored state. The four arms of the late spill scout (`base`, `τ_ℓ late`, `anti late`, and `anti held`) at seeds 704 to 711.…

Our design rule is to compare conditions as two independent samples. We pair runs by seed only when they branch from one stored state and are measured soon after (eng/determinism.md). But "soon after" has no number yet.

Two earlier results set the rule. Removal under `τ_ℓ fixed` and `τ_ℓ anneal`, which share their seeds, correlated about 0.2 across eight seeds at epoch 400 (the *fixed confirmation*°). Twin runs (the same in all but the GPU arithmetic) are as far apart in their weights as unrelated runs by epoch 16, while their task loss agrees to a hundredth (the *low-runs report*°).

fixed confirmation. `τ_ℓ fixed` at the fresh seeds of the anneal confirmation.…

low-runs report. Twin runs, the same in all but the GPU arithmetic, are as far apart in their weights as unrelated runs by epoch 16, while their task loss agrees to a hundredth for the whole run.

So the weights drift apart early, while some scores do not. Here we ask how alike two runs at one seed stay, score by score, as training goes on, using only the trajectories and kept checkpoints that four earlier experiments stored. For each pair of conditions trained at the same seeds, we take the correlation of a score across seeds at each record (each recorded epoch).

Observations

- From one initialization° (E1): over the first 64 epochs, the two runs of a seed start out ranking the seeds in nearly the same order on the task score° and end at a rank correlation of about a half, under every pairing of recipes. Removal° correlates by about a third, at the edge of chance, and spill° by no more than chance. The alignment° with e_1 (α) moves so much from record to record that one record of a run predicts its own next record about as poorly as it predicts the paired run.

From one initialization. Four recipes stopped at epoch 64, each at the 32 seeds of the loss-rate screen (712 to 743)…

Task score. Expected exact match on the held-out probe contexts, averaged over ops, recorded every four epochs during training.…

Removal. The drop on `difference` under the full e_1 edit, applied at every position and slice.…

Spill. The largest net drop on any other op, under the same edit.

Alignment. How far a state points along e_1…

- Over the whole run° (E2): at eight seeds, once the anchor forms, no correlation of removal or spill can be told apart from zero, including the correlation of 0.2 at epoch 400. The task score of a run is only loosely tied to that of the control at its seed. Netting the task change against that control, seed by seed, leaves its spread about where it was.

Over the whole run. `τ_ℓ fixed` and `τ_ℓ anneal` at seeds 712 to 719, trained to epoch 400, with the control at the same seeds (no pull and no anti term).…

- From one stored state° (E3): runs carried on from one epoch-64 state stay alike in spill to the end. They stay alike in removal and the task score for at least a hundred and fifty epochs, under each change of schedule and under a change of GPU. Under the two anti arms, the pairing in removal loosens after epoch 300, though at eight seeds that is not firm.

From one stored state. The four arms of the late spill scout (`base`, `τ_ℓ late`, `anti late`, and `anti held`) at seeds 704 to 711.…

Scope and method

This is an exploratory re-analysis, with no preregistration and no gate. It covers three kinds of pairing, ordered by how much the two runs of a seed share, least first.

From one initialization

Four recipes stopped at epoch 64, each at the 32 seeds of the loss-rate screen (712 to 743): `τ_ℓ fixed` and `τ_ℓ anneal` from the fixed and anneal confirmations, `step, soft τ` and `falloff` from the loss-rate screen. At a seed the four share their initialization, batches, and label draws, and differ by varying amounts. `τ_ℓ anneal` is `τ_ℓ fixed` with the slice temperature sharpening (by under 1% before epoch 64), `step, soft τ` pulls at a step over slices 2 to 4 in place of the pool over slices 1 to 3, and `falloff` is that step with the position temperature held fixed. The fixed confirmation also scored the edit on both of its recipes at epochs 16, 32, and 64, at forty seeds (704 to 743).

Over the whole run

`τ_ℓ fixed` and `τ_ℓ anneal` at seeds 712 to 719, trained to epoch 400, with the control at the same seeds (no pull and no anti term). At the end of training only, three more recipes join this comparison with the control, at seeds 704 to 711 with their own control, as the reference-baseline experiment and the soft pool scout scored them.

From one stored state

The four arms of the late spill scout (`base`, `τ_ℓ late`, `anti late`, and `anti held`) at seeds 704 to 711. Each arm is carried on from the stored epoch-64 state of `τ_ℓ 0.1, τ_ℓ fixed` at its seed, on an A10G, and each arm other than `base` changes one schedule from epoch 64 on. The soft pool scout trained the same schedule as `base` from the same state on an L4, which gives one more pairing: the same model on another GPU.

Within each pairing, both runs trained on the same kind of GPU (L4 for the first two kinds of pairing, A10G for the third), except for that last pairing. As a check that the code pairs the right runs, `anti late` and `anti held`, which share a schedule up to epoch 200, agree exactly at every kept checkpoint to epoch 192.

A correlation between unrelated scores falls within ±0.36 of zero about nineteen times in twenty at 32 seeds, within ±0.32 at forty, and within ±0.72 at eight. Each figure shades that range grey, as the band of chance. Neighboring records of one pairing are not independent draws, so a correlation that stays above the band over many records says more than one point does.

The measurements

Rank correlation across seeds

How well the order of the seeds on a score under one condition predicts their order under the other (Spearman's correlation), for two conditions trained at the same seeds, at one epoch. It is 1 when the seeds come in the same order under both, and near 0 when one order says nothing about the other. Near 1, a difference taken seed by seed cancels most of the variation between seeds; near 0, none of it. Ranks rather than values, because spill is near zero on most runs, and one run that spills under both conditions would set the correlation of the values on its own.

Task score

Expected exact match on the held-out probe contexts, averaged over ops, recorded every four epochs during training. Higher is better. At the end (E2), it is the clean score of the scoring suite, averaged over ops.

α at the example answers

How far the pull has brought the anchored op onto e_1: the mean alignment with e_1 of the states at the example answers of `difference` contexts, at slice 3, recorded every four epochs. The loss-rate call averages it over epochs 44 to 64.

Removal

The drop on `difference` under the full e_1 edit, applied at every position and slice. It is net of the median drop when the same edit is applied along each of the other 63 coordinate axes of the run. In E3 it is net of the median of eight reference edits instead, as spill-timing scored it. We report the net drop rather than a share of the way to the target null, since a stopped run has no end record to scale it by.

Spill

The largest net drop on any other op, under the same edit.

From one initialization (E1)

Started from one initialization° under two recipes, how long do the two runs of a seed stay more alike than runs at different seeds? If the seed sets most of what a run does, the correlation should stay high whatever the change. If small differences grow quickly, it should fall early, and fall faster the larger the change.

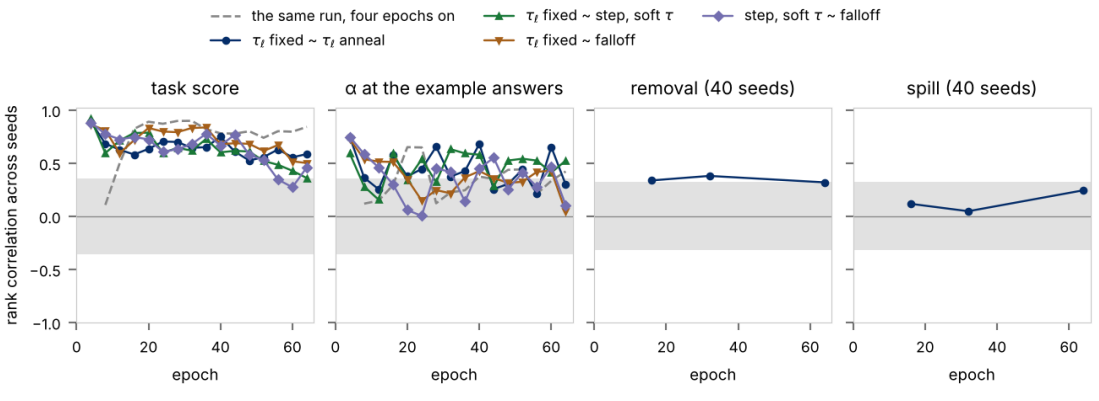
From one initialization. Four recipes stopped at epoch 64, each at the 32 seeds of the loss-rate screen (712 to 743)…

In the figure below, the task score° stays partly paired through epoch 64. Removal° is only weakly paired, and spill° no more than chance would give.

Task score. Expected exact match on the held-out probe contexts, averaged over ops, recorded every four epochs during training.…

Removal. The drop on `difference` under the full e_1 edit, applied at every position and slice.…

Spill. The largest net drop on any other op, under the same edit.



From one initialization: how alike the runs of a seed stay under two recipes. *Each line is one pairing of recipes, and each point the rank correlation of a score across the 32 seeds of the loss-rate screen (712 to 743), at one record.* ***Left two:*** *the task score and α at the example answers, recorded every four epochs during training; the dashed line is each run against itself at its previous record, averaged over the four recipes.* ***Right two:*** *removal and spill under the full e_1 edit at the three checkpoints kept to epoch 64, for the one pairing that scored them there, at forty seeds (704 to 743). The grey band is where a correlation falls about nineteen times in twenty when the two scores are unrelated.*

At epoch 4, the two runs of a seed rank the seeds in nearly the same order on the task score under every pairing (a rank correlation of 0.86 to 0.91). The correlation falls slowly from there, to about a half on average over epochs 44 to 64 (0.49 to 0.61).

From epoch 16 on, a run correlates with itself four epochs later at about 0.8. So the fall comes from the two runs drifting apart, while under any one recipe the order of the seeds on the score changes little from one record to the next.

The size of the change makes no clear difference. `τ_ℓ fixed` and `falloff`, which pull at different slices with different temperatures, stay about as alike in the task score as `τ_ℓ fixed` and `τ_ℓ anneal`, which differ by under 1% in one temperature. That fits what the *low-runs report*° found in the weights, where a small difference grows as quickly as a large one.

***low-runs report*.** Twin runs, the same in all but the GPU arithmetic, are as far apart in their weights as unrelated runs by epoch 16, while their task loss agrees to a hundredth for the whole run.

α° behaves differently. The runs of a seed start alike in it (0.59 to 0.75 at epoch 4), and by epoch 44 the correlation is about a third (0.31 to 0.47). But a run correlates with itself four epochs later about as weakly (0.37). So α moves a lot from record to record, and the paired run at the same epoch is about as good a guide to it as the same run four epochs earlier.

Alignment. How far a state points along e_1…

The two runs see the same batches in the same order, which may keep some of that movement in step, though nothing here tests that.

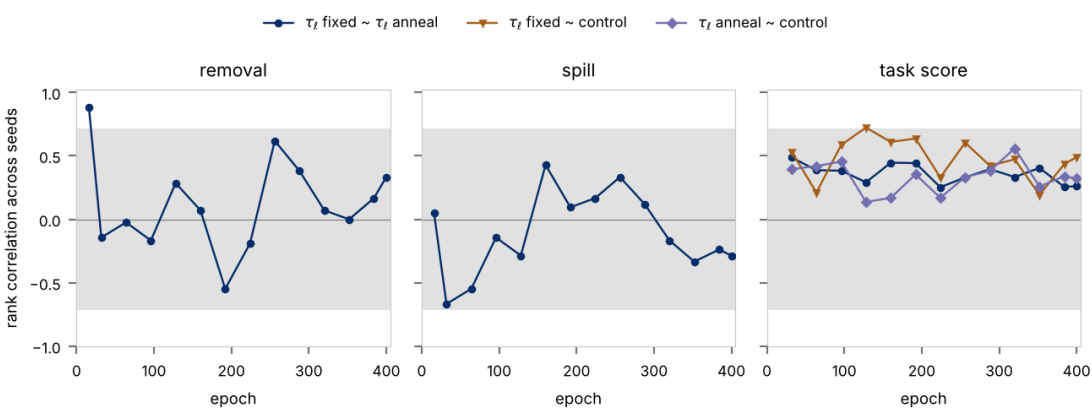
Removal under the edit correlates by about a third at epochs 32 and 64 (0.32 to 0.38 over forty seeds), at the edge of chance. Spill stays inside the band at all three checkpoints. At epoch 16 the anchor has barely formed, and removal is close to zero on every run.

Pairing by seed from one initialization would cancel about half of the variation between seeds in the task score, about a third in removal if that correlation is real, and none we can detect in spill.

Over the whole run (E2)

Does what is left of the pairing at epoch 64 last to the end of training? And how alike is a run to the control at its seed, which the D2.2 design nets the task change against?

The figure follows `τ_ℓ fixed` against `τ_ℓ anneal`, and each against the control, to epoch 400 at eight seeds, where the band of chance is wide.



Over the whole run, at eight seeds. *The rank correlation across seeds 712 to 719.* ***Left two:*** *removal and spill under the full e₁ edit,* `τ_ℓ fixed` *against* `τ_ℓ anneal`*, at the checkpoints kept at epochs 16, 32, and 64, then every 32 epochs, and at the end.* ***Right:*** *the task score, for the two recipes against each other and each against the control at the same seeds. Each point is the mean correlation over the records (every four epochs) of the 32 epochs up to it, or of the last 16 epochs for the final point. The grey band is where a correlation over eight seeds falls about nineteen times in twenty when the two scores are unrelated (±0.72).*

From epoch 32 on, every correlation of removal° and spill° between the two recipes falls inside the wide band of chance for eight seeds, wandering from about −0.7 to 0.6 between neighboring checkpoints.[1]

Removal. The drop on `difference` under the full e₁ edit, applied at every position and slice....

Spill. The largest net drop on any other op, under the same edit.

1 The 0.88 in removal at epoch 16 comes before the anchor forms, when removal is noise around zero.

So the correlation of about 0.2 at epoch 400 (as a rank correlation, 0.33, with an interval of −0.50 to 0.85) is consistent with anything from no pairing at all to a strong one. At epoch 64, the forty seeds of E1 pin the pairing in removal down better, to about a third.

The pairing in the task score° is as unsteady. The two recipes correlate by about a third to a half over the run, and each correlates about as well with the control, though the control has neither the pull nor the anti term. So for the task score, what the two runs of a seed share seems to matter more than how alike their recipes are.

Task score. Expected exact match on the held-out probe contexts, averaged over ops, recorded every four epochs during training....

recipe	seeds	rank correlation	95% interval	spread, netted per seed ↓	spread, netted against the control mean ↓
`τ_ℓ fixed`	712–719	0.60	−0.21 to 0.92	0.0110	0.0110
`τ_ℓ anneal`	712–719	0.45	−0.39 to 0.88	0.0081	0.0096
`τ_ℓ fixed`	704–711	0.12	−0.65 to 0.77	0.0069	0.0058
`step, soft τ`	704–711	0.45	−0.39 to 0.88	0.0069	0.0082
`falloff`	704–711	0.12	−0.65 to 0.77	0.0123	0.0134

The task score at epoch 400, against the control at the same seeds. *The rank correlation across eight seeds between the clean score of each run (mean over ops) and that of the control, with its interval. The two spreads are standard deviations over seeds of the task change. The first nets each run against the control at the same seed, as reports up to the fixed confirmation did. The second nets it against the seed mean of the control, as the D2.2 design now does, and includes the uncertainty of that mean.*

The table gives the end of training for five recipes and seed ranges. At epoch 400, the task score of a run correlates with that of the control at its seed by 0.12 to 0.60, and every interval reaches below zero.

Netting the task change against the control seed by seed gives about the same spread as netting it against the seed mean of the control, a little smaller in three rows of the table and larger in one. In four of the five rows the control spreads about half as much as the anchored runs, which is why the choice matters little either way.

So at the end of training, eight seeds cannot show that pairing by seed helps on any of these scores.

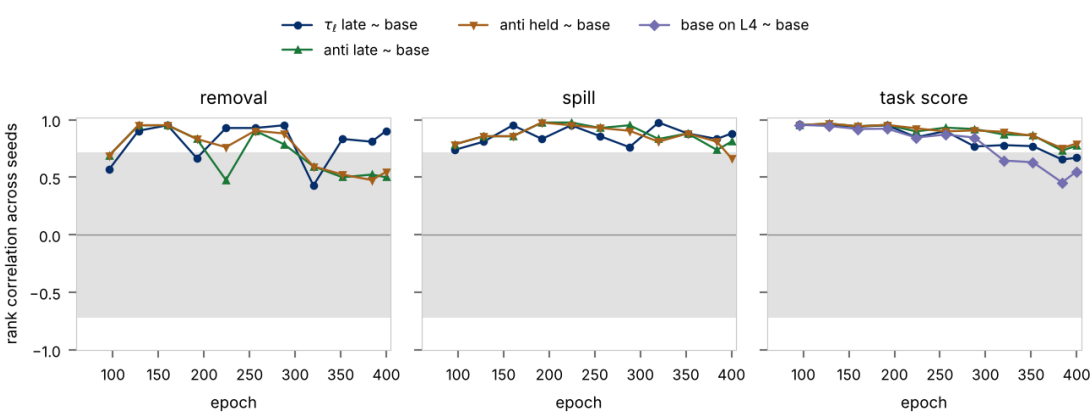
From one stored state (E3)

Runs carried on from one stored state° share everything up to that state. By epoch 64 the runs at different seeds have already drifted apart, and both runs of a seed start from where that seed got to. How long do they stay alike after the branch, and does that depend on the change?

From one stored state. The four arms of the late spill scout (`base`, `τ_ℓ late`, `anti late`, and `anti held`) at seeds 704 to 711....

The figure compares each arm of the *late spill scout*° with `base`, the arm that keeps the schedule unchanged, at the same seed.

late spill scout. Four ways of carrying `τ_ℓ fixed` on from its stored epoch-64 states; holding the anti weight high lowered removal, by about a quarter on one seed.



From one stored state: the late spill scout. *The rank correlation across seeds 704 to 711 between each arm and `base`, all carried on from the stored epoch-64 state of `τ_ℓ 0.1, τ_ℓ fixed` on an A10G.* ***Left two:*** *removal and spill under the full e_1 edit at the checkpoints kept every 32 epochs from epoch 96, and at the end, net of eight reference edits.* ***Right:*** *the task score, each point the mean correlation over the records of the 32 epochs up to it (16 for the last). One more pairing (`base` on L4) uses the run of the same schedule from the soft pool scout, trained on L4 throughout. The grey band is the chance band for eight seeds (±0.72).*

They stay alike for a long time. In spill°, every arm keeps a rank correlation with `base` of at least 0.67 at every checkpoint to the end. In removal°, every arm stays mostly between 0.8 and 0.95 to epoch 288, with a dip to about a half at single checkpoints.

Spill. The largest net drop on any other op, under the same edit.

Removal. The drop on `difference` under the full e_1 edit, applied at every position and slice....

After epoch 288, `τ_ℓ late` stays near 0.74 on average in removal, while the two anti arms fall to about a half (0.53 and 0.54), with their anti weight still above that of `base`. The *late spill scout* found that the anti arms lowered removal by the end, by about a quarter on one seed. So the two runs of a seed drifting apart here fits a change that by then has a real effect, though at eight seeds the loosening is not firm (below).

Along training, E3 nets removal against eight reference edits, where E1 and E2 net it against the median drop along the other 63 axes. Could the baseline, rather than the pairing, set E3 apart from E2? At the end of the run, under the 63-axis baseline, the three arms still correlate with `base` by 0.64 to 0.86 in removal and 0.76 to 0.95 in spill, above every correlation of E2 after the anchor forms.

Under that baseline the anti arms correlate with `base` at 0.64 and 0.86 in removal, against about a half under the reference edits, and at eight seeds those values cannot be told apart.

The task score° starts at a correlation near 1 for every arm and for the change of GPU, and stays at about 0.9 or above to about epoch 250. After that it falls: least for the anti arms (about 0.82 on average after epoch 288), more for `τ_ℓ late` (0.73), and most for the change of GPU (0.59).

Task score. Expected exact match on the held-out probe contexts, averaged over ops, recorded every four epochs during training....

At the end, `base` on the A10G and on the L4 also correlate in removal and spill (0.79 and 0.83), and both clear the band at eight seeds.

So from a stored state, pairing by seed helps on all three scores for at least a hundred and fifty epochs, under each change here and under a change of GPU. The branch passes the differences between seeds from the first 64 epochs on to both runs. Until late in the run, the changes after the branch add less than those differences.

Discussion

There is no one number for "soon after": what the runs share when they part seems to matter more than how long ago they parted, and in E1 the size of the change made no clear difference to the task score° either. That fits the design rule, and puts the weight on the stored state rather than on the time since the branch.

Task score. Expected exact match on the held-out probe contexts, averaged over ops, recorded every four epochs during training.…

If it holds beyond these experiments, arms that branch from a stored state after the anchor forms can be paired for most of a run. Conditions that share only an initialization gain only weakly from pairing on removal°, and not detectably on spill°, at any checkpoint we have.

Removal. The drop on `difference` under the full e_1 edit, applied at every position and slice.…

Spill. The largest net drop on any other op, under the same edit.

The evidence for the long pairing rests on one stored state at one epoch and on eight seeds, so how far it generalizes is not settled. A stored state taken earlier, before the seeds have parted, may behave more like a shared initialization.

At epoch 400, netting the D2.2 task change against the control at each seed neither narrows nor widens its spread much (E2), so the D2.2 task gate, which nets against the mean of the controls, gives up little by not pairing.

α at the example answers° moves so much from record to record that a single record is a poor measure of a run (E1). The loss-rate call averages it over six records, and this result supports that choice.

α at the example answers. How far the pull has brought the anchored op onto e_1…

References

fixed confirmation

: `τ_ℓ fixed` at the fresh seeds of the anneal confirmation. Removal under the two recipes correlated about 0.2 across eight seeds at epoch 400, and the two runs of a seed were up to 0.43 apart in removal at epoch 32.

anneal confirmation

: The pool over slices 1 to 3 with the slice temperature annealing from 0.1 to 0.025, at seeds 712 to 719 to the end and at forty seeds to epoch 64.

loss-rate screen

: `falloff` and `step, soft τ` at the same 32 fresh seeds to epoch 64, called kept or lost from α at the example answers over epochs 44 to 64.

late spill scout

: Four ways of carrying `τ_ℓ fixed` on from its stored epoch-64 states; holding the anti weight high lowered removal, by about a quarter on one seed.

soft pool scout

: The pool over slices 1 to 3 at seeds 704 to 711 on L4, whose epoch-64 states the late spill scout carried on from.

reference-baseline experiment

: Removal and spill net of the median drop along the other coordinate axes of the run itself, scored on the stored runs of `falloff`, `step, soft τ`, and the control at seeds 704 to 711.

low-runs report

: Twin runs, the same in all but the GPU arithmetic, are as far apart in their weights as unrelated runs by epoch 16, while their task loss agrees to a hundredth for the whole run.

spill-timing

: Spill along training, net of the median of eight reference edits.

---

1. The 0.88 in removal at epoch 16 comes before the anchor forms, when removal is noise around zero. ↩