When Pythia-70m's VERIFIED neurons form

Will Sackfield, Open Interpretability · https://open-interpretability.com/research/neuron-development

Abstract

Explanations of single neurons are usually written for a finished model, but it is not known whether they describe the model for most of its training. We followed 86 neurons of Pythia-70m back through 20 public training checkpoints. Each one's explanation had passed an adversarial review. The training loss reaches within 1% of its final value by step 16,000. At that step, the neurons' activations correlate with their final form at a median of only 0.62, and they keep changing until the end of training. Between adjacent checkpoints after step 16,000, the median correlation is 0.78 to 0.82, so the late change is real movement and not only an effect of comparing with the end. 38 of the 86 explanations hold without a break from step 16,000 or earlier, and 38 only from step 64,000 or later. Of the 13 explanations that hold at step 16,000 and fail later, 9 pass again before the final step, and 18 of their 23 failures are within two probes of a gate. A first version of this paper studied 9 of these neurons. Its main results hold on the 86. Its claim that the two measures of formation order the neurons in reverse does not hold: they are only weakly related. A neuron's explanation describes one checkpoint, and the loss curve does not show when it becomes true. Over the whole of each layer, the deeper layers lose effective rank after the loss settles, but almost no pole becomes dormant.

Status: Preprint. Work left: #4156 (more checkpoints between steps 64,000 and 143,000, OLMo and the local learning coefficient), #6156 (a second probe set for the 13 explanations that fail after step 16,000). Also #6524 (whether the late checkpoints learn new data more slowly).

1 Introduction

Interpretability research often explains a trained network one unit at a time (Bau et al., 2017). An explanation of this kind is written for the final weights. A network changes during training, though. If a unit's function forms early and then stays, one explanation describes most of the run. If it keeps changing, an explanation is true only of the checkpoint that it was written for.

Pythia (Biderman et al., 2023) makes this question testable, because its training runs are public at many steps. Earlier work on training dynamics has tracked circuits and global measures. Examples are the formation of induction heads (Olsson et al., 2022) and the stages of a whole model (Hoogland et al., 2025). Less is known about single neurons whose explanations have passed adversarial tests. This paper asks when such neurons form, and whether their explanations hold from then on. It also asks whether their formation lines up with the plateau of the training loss. It answers research question 7, *When do neurons form during training?* (research questions).

The first version of this paper, on 2026-11-16, studied the 9 neurons that the registry had VERIFIED then. By 2026-11-19 the registry held 86. This version reports all 86, with the first 9 beside them, so that a reader can see which claims held.

The paper's contributions are these:

- A measurement of 86 reviewed neuron explanations in Pythia-70m, in all six layers, at 20 checkpoints from initialisation to the end of training. It uses the review's own tests.
- Evidence that the neurons keep changing long after the loss settles. Correlations between adjacent checkpoints show that this is movement, not only an effect of the measure.
- Evidence that the shape of a neuron's activations and the success of its explanation are two weakly related views of its formation.
- A measurement of dormant poles and effective rank in every MLP layer at the same checkpoints. It tests whether pretraining shows the signs of lost plasticity that Dohare et al. (2024) report.

2 Related work

Network Dissection measured how well individual units of vision networks align with semantic concepts. It also followed that alignment across training iterations (Bau et al., 2017). We ask the same kind of question of language-model neurons, with explanations in place of fixed concept labels.

The Pythia suite was built to study how language models develop across training and scale (Biderman et al., 2023). It publishes checkpoints of each model through training. PolyPythias adds nine more training runs, with other seeds, at five of its sizes (van der Wal et al., 2025). One of those sizes is the model that we study. We use one of the smallest models, so that every checkpoint can be read on one machine.

Olsson et al. (2022) found that induction heads form in a phase change early in training, and that this change is visible as a bump in the training loss. Their result links a mechanism's formation to a feature of the loss curve. Our neurons behave differently. Their formation lines up with no feature of the loss that we measured, and much of it happens after the loss has flattened.

Hoogland et al. (2025) divide the training of a small attention-only language model, of about 3 million parameters, into approximate developmental stages. They find the stage boundaries with the local learning coefficient, a measure of the degeneracy of the loss landscape (Lau et al., 2023), not with the loss. Their approach suggests a measure that may show the late changes that we find. We leave that test to further work (*Conclusion*).

3 Method

Terms. A neuron in a transformer's MLP layer can carry a different meaning when it is strongly positive and when it is strongly negative. Following Cammarata (2026), we study one side at a time, and call it a *pole*. The registry that we draw on stores an explanation for each pole that it has audited. An explanation is *VERIFIED* when it survives an adversarial review. In that review, a second model writes probes meant to break the explanation. *Hard negatives* look relevant but should not activate the pole, and *holdouts* should activate it. A probe *activates* the pole when the pole's peak activation on it

reaches 0.25 of the pole's maximum, which we call the *activation line*. The explanation passes when at least 80% of the hard negatives and 60% of the holdouts behave as expected. We call these two thresholds the review's *gates*. We write L5/N655+ for the positive pole of neuron 655 in layer 5.

Model and checkpoints. We read Pythia-70m (Biderman et al., 2023) at 20 checkpoints, in float32. Eleven are early: steps 0, 1, 2, 4, 8, 16, 32, 64, 128, 256 and 512. Nine are later: steps 1,000, 2,000, 4,000, 8,000, 16,000, 32,000, 64,000, 128,000 and 143,000, the last of which is the end of training.

Poles. We take each Pythia-70m pole whose newest record in the registry is VERIFIED, on 2026-11-19. There are 86, in all six layers: 16 in layer 0, 18 in layer 1, 9 in layer 2, 15 in layer 3, 15 in layer 4 and 13 in layer 5. A pole whose newest record is not VERIFIED would be left out, even with an older VERIFIED record, but there were none. The *first 9* are the poles of this paper's first version, chosen on 2026-11-16 as the newest VERIFIED record of each pole.

One negative pole. One of the 86, L3/N866−, is a negative pole. Its peak is 0.170 at every checkpoint, which is the depth of GELU's dip. So its peaks on the windows are almost all tied, and its correlation with the final form is near 0 until the last step, where it is 1 by construction. We count it among the explanations that hold, and leave it out of every correlation. The correlation results are therefore over 85 positive poles.

Windows. We use 300 windows of 64 tokens from the last 200 documents of the corpus that the registry scans, and the same windows at every step.

Measures. For each pole and step we measure four things.

- *The correlation with the final form*, written r, is the Pearson correlation between the pole's peak activations on the 300 windows at that step and at step 143,000:

  $$r_t = \mathrm{corr}(a_t, a_{143000})$$

  Here a_t is the vector of the pole's peak activations on the 300 windows at step t.

- *The correlation with the checkpoint before* is the same correlation with the previous checkpoint in the list.

- *Whether the explanation holds*: the recorded probes are run again, and the explanation holds when both pass rates reach the gates. The activation line at each step is 0.25 of a scaled maximum: the recorded maximum, times the step's largest window activation divided by the last step's.

- *The loss* is the mean next-token loss on the 300 windows.

Reproduction of the first run. We ran the first 9 poles again with the same code before the full run. No pass or fail, pass rate or formation step changed. The largest numeric change was 2.5×10^{-4} in a step's maximum, which is float rounding.

4 Does the loss settle before the poles?

The first question is whether the poles have taken their final form by the time the loss stops falling. On the 300 windows, the loss falls from 11.09 at step 0 to 3.62 at step 16,000. That is within 1% of its final 3.59, and the loss changes little after step 32,000 (Figure 1, right). At step 16,000, however, the median correlation of the 85 poles with

their final activations is 0.62. The middle half of the poles lie between 0.48 and 0.73, and the range is 0.26 to 0.98. The median rises to 0.71 at step 32,000, 0.80 at step 64,000 and 0.94 at step 128,000 (Figure 1, left). The first 9 had a median of 0.48 at step 16,000.

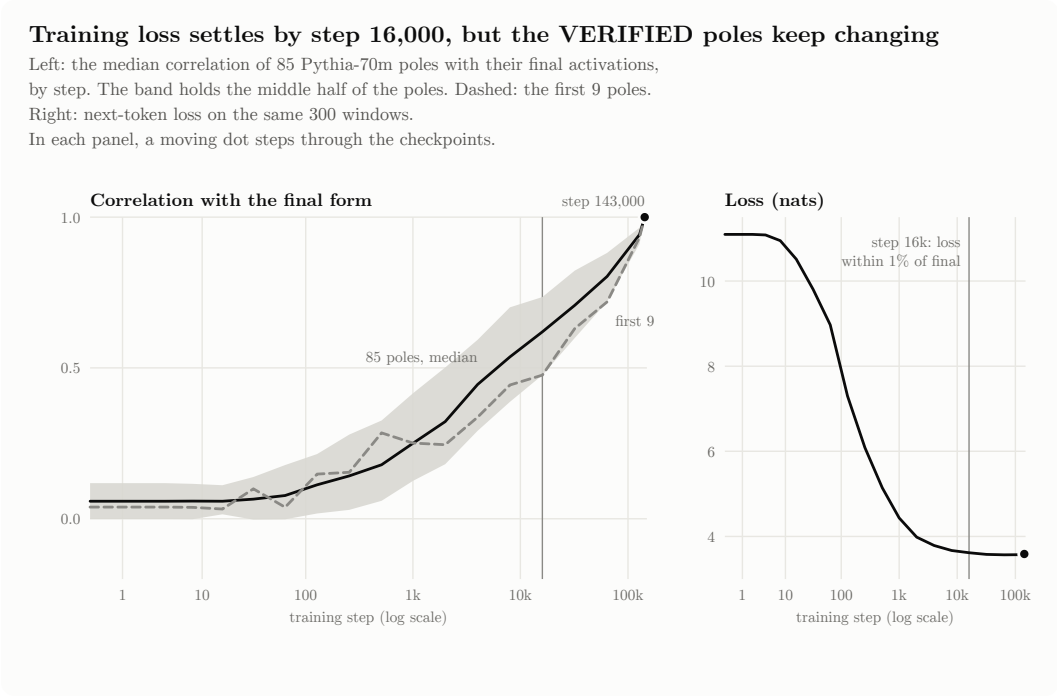

Figure 1. Left: the median correlation of the 85 positive poles with their final activations, by training step on a log scale. The band holds the middle half of the poles, and the dashed line is the median of the first 9. Right: the mean next-token loss on the same 300 windows. In each panel a moving dot steps through the 20 checkpoints, and a counter names the step. So the poles can be seen still moving after the loss has stopped. The vertical line marks step 16,000, by which the loss is within 1% of its final value.

A correlation with the final checkpoint must reach 1 at the end, so a pole that drifts steadily looks as if it forms late. Correlations between adjacent checkpoints test this (Figure 2, right). Between steps 8,000 and 16,000 the median is 0.79. It is 0.78 between steps 16,000 and 32,000, and again between 32,000 and 64,000. It is 0.82 between steps 64,000 and 128,000. Each doubling of steps after the loss settles thus moves the poles about as much as the doubling before it. The poles keep moving through the second half of training, so the late change is not only an effect of the measure.

5 When does each explanation start to hold?

The second question is whether each pole's explanation holds from some step onwards. We record the first step after which the recorded probes pass the gates at every later checkpoint. For 38 of the 86 poles this step is 16,000 or earlier. For 10 it is step 32,000, and for 38 it is step 64,000 or later (Figure 2, left). The first 9 split in the same proportion, 4 and 5 (Table 1). 51 of the 86 explanations pass at step 16,000 itself, but 13 of those fail again at a later checkpoint.

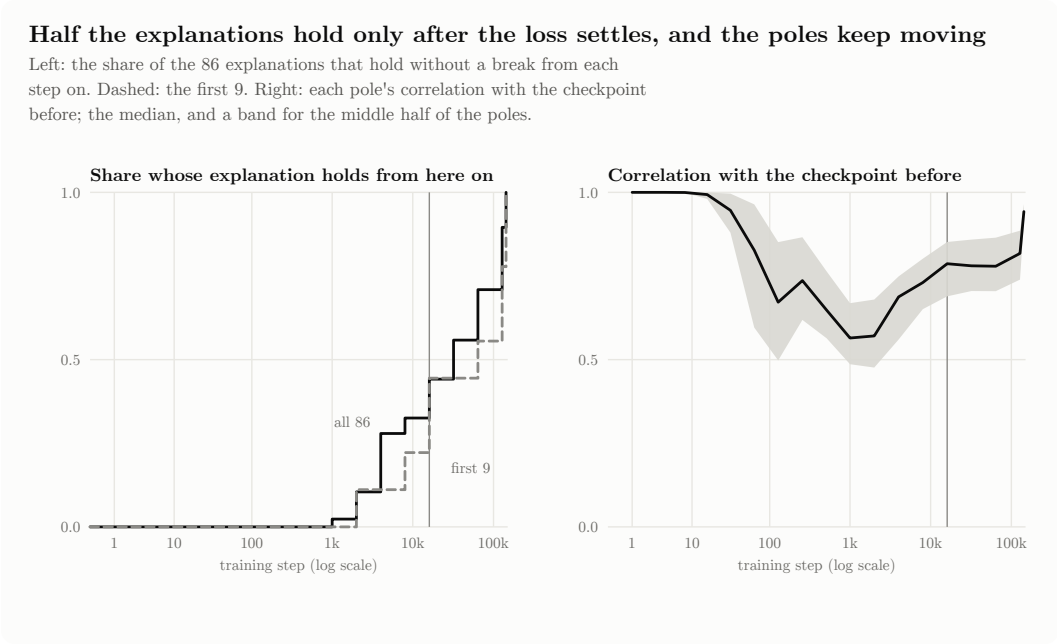

Figure 2. Left: the share of the 86 explanations that hold without a break from each step on. The dashed line is the first 9. Right: each positive pole's correlation with the checkpoint before, as the median and a band for the middle half of the poles. The vertical line marks step 16,000.

Pole	Holds without a break from	Correlation 0.5 first reached	Correlation at step 16,000
L5/N655+	2,000	32,000	0.38
L5/N674+	8,000	32,000	0.37
L5/N1256+	16,000	8,000	0.63
L1/N535+	16,000	32,000	0.45
L4/N566+	64,000	4,000	0.73
L2/N1181+	128,000	64,000	0.48
L4/N751+	128,000	64,000	0.29
L4/N216+	143,000	4,000	0.74
L4/N812+	143,000	1,000	0.73

Table 1. The first 9 poles, as the first version reported them. The re-run gives the same values. The first column gives the first step from which the pole's recorded probes pass the review's gates at every later checkpoint. The second gives the first step at which its correlation with its final activations reaches 0.5. The third gives that correlation at step 16,000, where the loss has settled.

Explanations therefore settle at very different times. At the step where the loss has settled, fewer than half of them describe the model.

6 Does formation line up with the loss plateau?

The third question is whether the poles form when the loss flattens. If they did, most poles would reach a high correlation with their final form near step 16,000. Of the 85

positive poles, 16 reach a correlation of 0.8 by step 16,000, 25 by step 32,000 and 43 by step 64,000. The median pole reaches it at step 64,000, four times as many steps as the loss needs to settle. Formation, measured this way, does not line up with the loss plateau. The first 9 gave the same answer: only two of them reached 0.8 before step 128,000, both at step 64,000.

7 Formation by layer

The 86 poles cover all six layers, where the first 9 covered four. Table 2 gives each layer's poles. Layer 0's poles take their final shape earliest, with a median correlation of 0.90 at step 16,000. Its median pole reaches 0.8 at step 8,000. Layer 2's poles are the latest by both measures: one of its 9 explanations holds from step 16,000 or earlier. Layer 1 has the most explanations that hold early, 9 of 18, and layer 3 the highest share, 8 of 14. Yet both layers' shapes settle late.

Layer	Positive poles	Median correlation at step 16,000	Median first step with correlation 0.8	Explanations that hold from step 16,000 or earlier	Median step from which the explanation holds
0	16	0.90	8,000	8	40,000
1	18	0.60	64,000	9	24,000
2	9	0.46	128,000	1	128,000
3	14	0.61	128,000	8	10,000
4	15	0.72	128,000	6	64,000
5	13	0.53	128,000	6	32,000

Table 2. Formation by layer, over the 85 positive poles. A median of two middle values is their mean. Layer 3's negative pole, left out here, holds from step 128,000.

8 Do explanations that fail come back?

51 of the 86 explanations hold at step 16,000, and 13 of those fail at one or more later checkpoints. Such a failure can mean two things. The explanation can be lost, so that the pole no longer does what it says. Or the pass rates can sit near a gate, so that one or two probes cross it by chance. French (1999) reviews a test from the study of forgetting: a network that seems to have forgotten a task can often relearn it quickly, which the literature calls savings. As an analogue, which is our own step, we ask whether a failed explanation passes again at a later checkpoint.

The last checkpoint cannot answer this question. Every explanation holds at step 143,000 by selection, because the registry wrote and VERIFIED each one for the final model. So only a pass at an intermediate checkpoint, 32,000, 64,000 or 128,000, counts as a return. For each failure, we also count the probes by which it fell short. That is the number of hard negatives that would have had to stay quiet, or of holdouts that would have had to fire, for both gates to hold. Each record has 20 hard negatives and 10 holdouts, so one probe moves specificity by 0.05 and sensitivity by 0.10.

Most of the 13 come back. 9 pass again at an intermediate checkpoint: six at step 64,000 and three at step 128,000 (Figure 3, Table 3). One of them, L1/N1124+, fails again at 128,000 by one probe. The other four fail at every intermediate checkpoint after their first failure, and hold again only at the final step. The 13 fail at 23 checkpoints in all. 13 of

these failures are one probe short of a gate, 5 are two short, and 5 are three or more short. For comparison, 18 of the 130 passing checkpoints of the 51 explanations between steps 32,000 and 128,000 sit exactly at a gate, one probe from failing. So most failures are the same kind of event as those passes, seen from the other side of the gate.

Only two poles lose their explanation by a wide margin, and both move back towards it. L5/N659+ falls to a specificity of 0.10 at step 32,000, 14 probes short. It misses by one probe at 64,000 and holds from 128,000. L4/N216+ falls to 0.05 at steps 32,000 and 64,000, 15 probes short each time, and is one probe short at 128,000. Both lose specificity, not sensitivity. At those steps their poles fire on most of their hard negatives, so they respond to more than their explanations say. L3/N1795+, 4 probes short at 64,000, and L2/N184+, 3 short at 128,000, sit between the two kinds.

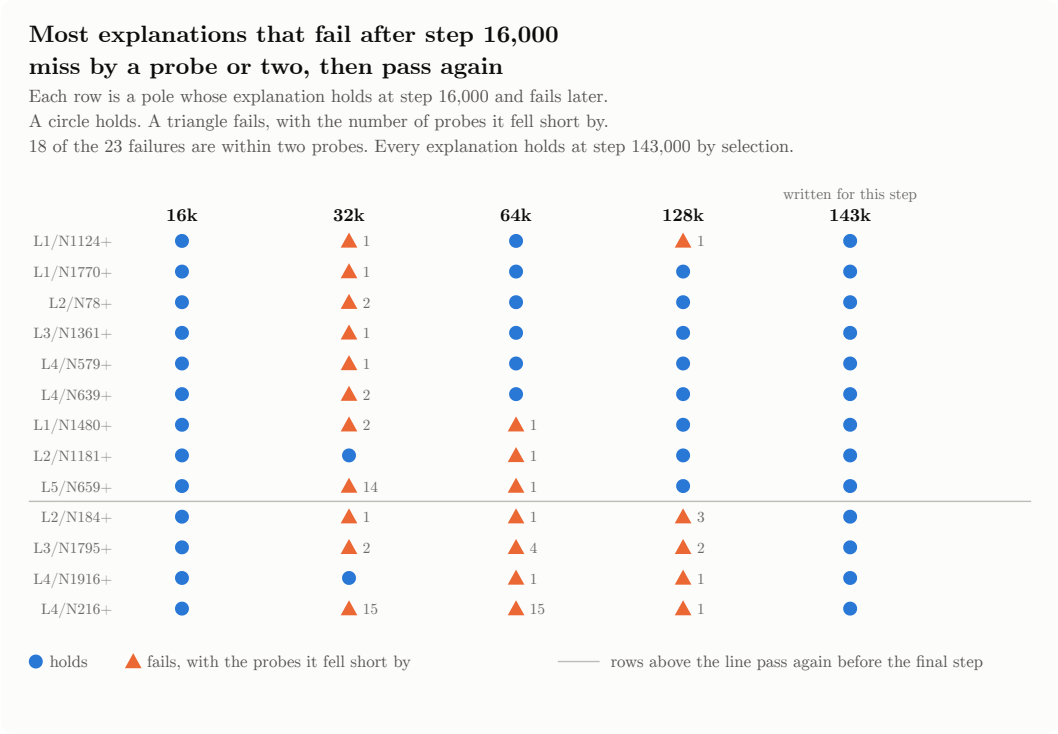

Figure 3. The 13 poles whose explanations hold at step 16,000 and fail at a later checkpoint. A circle holds, and a triangle fails, with the number of probes by which it fell short of the nearer gate. The rows above the line pass again at an intermediate checkpoint. The final checkpoint holds for every pole by selection.

Pole	Short at 32,000	Short at 64,000	Short at 128,000	First intermediate pass after a failure
L1/N1124+	1	holds	1	64,000
L1/N1770+	1	holds	holds	64,000
L2/N78+	2	holds	holds	64,000
L3/N1361+	1	holds	holds	64,000
L4/N579+	1	holds	holds	64,000
L4/N639+	2	holds	holds	64,000
L1/N1480+	2	1	holds	128,000
L2/N1181+	holds	1	holds	128,000
L5/N659+	14	1	holds	128,000
L2/N184+	1	1	3	none
L3/N1795+	2	4	2	none
L4/N1916+	holds	1	1	none
L4/N216+	15	15	1	none

Table 3. The probes by which each of the 13 explanations fell short of the nearer gate at each checkpoint after step 16,000. "Holds" means that both gates hold. The last column gives the first intermediate checkpoint at which the explanation holds again after its first failure.

So the 13 do not show that explanations are commonly lost once the loss settles. Most fail by one or two probes and pass again at the next checkpoint, which is what probe noise would produce. The two deep failures show something else. L5/N659+ drifts from its explanation for a stretch of training and then passes again before the final step. L4/N216+ comes back to within one probe by step 128,000, but it passes only at the final step, which the selection guarantees. That return resembles savings more than forgetting, although training here never left the same data.

9 Do the layers lose plasticity?

Dohare et al. (2024) find that networks that lose plasticity in continual learning also gain dormant units and lose effective rank. Pretraining on one data stream is not continual learning. Still, the late change above invites the question: do Pythia-70m's MLP layers show either sign? We measured all 4,096 poles of each layer at the same 20 checkpoints, on the same 300 windows, 18,782 tokens in all (#5134).

Measures. A positive pole is active on a token when its activation is above 0, and a negative pole when it is below 0. GELU keeps the sign of its input, so this rule matches Dohare et al.'s rule for ReLU units. A pole is *dormant* when it is active on under 1% of tokens, and *never-firing* when it is active on none. A second rule counts a pole as active at 0.25 times its own maximum at that step, which is the review's line. The *effective rank* of a layer is the exponential of the entropy of its activation matrix's normalised singular values (Roy & Vetterli, 2007). We compute it on a seeded sample of 2,000 tokens, the size that Dohare et al. use, and also on all tokens.

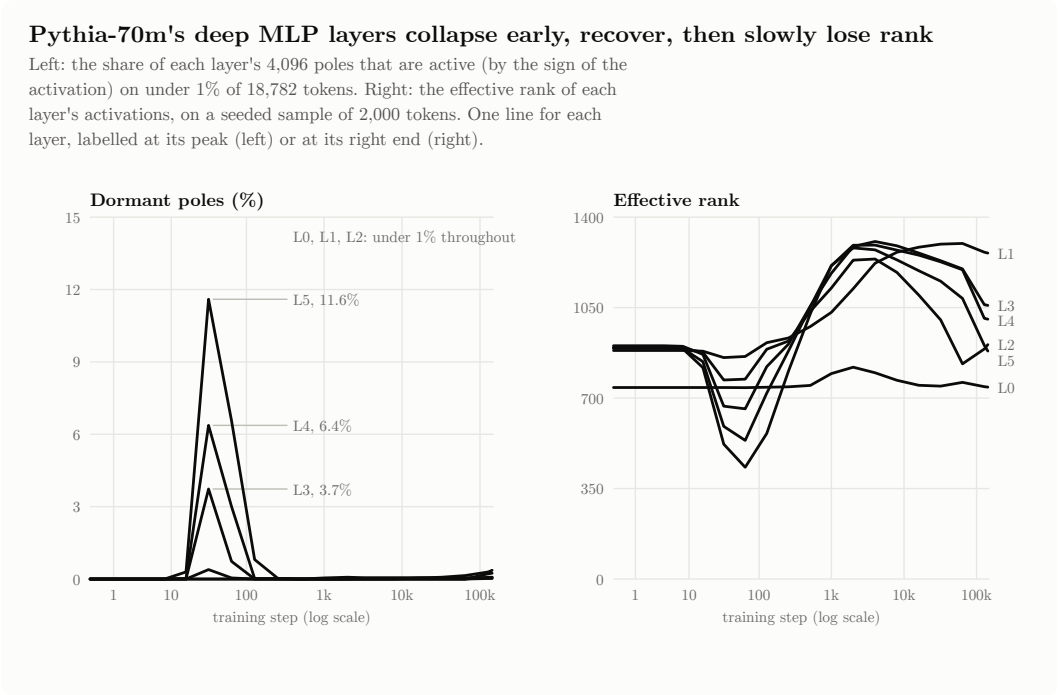

Figure 4. Left: the share of each layer's 4,096 poles that are dormant by the sign rule, by training step on a log scale. Right: each layer's effective rank on 2,000 tokens. Step 0 is the random initialisation.

Step	Dormant L3 / L4 / L5 (%)	Never-firing L5 (%)	Dormant at the line, L0–L5 (%)	Effective rank L0 / L1 / L2 / L3 / L4 / L5
0	0 / 0 / 0	0	0.0–1.0	741 / 884 / 892 / 902 / 895 / 895
32	3.74 / 6.37 / 11.60	1.00	0.0–19.5	741 / 857 / 770 / 669 / 591 / 522
64	0.73 / 3.00 / 6.57	0.78	0.0–10.7	741 / 861 / 773 / 659 / 537 / 433
512	0 / 0 / 0	0	0.0–0.2	749 / 977 / 1,036 / 1,056 / 1,054 / 1,026
4,000	0 / 0 / 0.05	0	0.0–0.5	798 / 1,222 / 1,238 / 1,306 / 1,292 / 1,273
16,000	0 / 0 / 0.05	0	0.3–1.4	749 / 1,284 / 1,099 / 1,261 / 1,253 / 1,193
143,000	0.24 / 0.07 / 0.37	0	3.1–8.3	742 / 1,261 / 907 / 1,058 / 1,005 / 882

Table 4. Dormant poles and effective rank at seven of the 20 checkpoints. "Dormant at the line" gives the lowest and highest layer. Layers 0 to 2 never have more than 0.4% dormant poles by the sign rule. `results.dormant.jsonl` holds every step and layer.

Figure 4 and Table 4 give the answer, which has three parts.

- **An early collapse that recovers.** From step 16 to step 128, the deeper layers lose rank and gain dormant poles. Layer 5's rank falls from 895 to 433 by step 64. At step

32, 11.6% of its poles are dormant and 1.0% never fire. By step 512 no pole is dormant, and every layer from 1 to 5 has a higher rank than at initialisation.

- **A slow loss of rank late in training.** Layers 2 to 5 reach their highest rank near steps 2,000 to 4,000. They then lose rank until the end, after the loss has settled. Layer 5 falls from 1,281 at step 2,000 to 882, below its rank at initialisation. Layer 1 keeps its rank, and layer 0 barely moves.
- **Few dormant poles by the sign rule, more at the line.** Late in training, at most 0.37% of a layer's poles are dormant by the sign rule, and none never fires. At the review's line, the dormant share rises in every layer after step 4,000, to 3.1% to 8.3% at the end. So the poles do not stop firing. Instead, more of them reach a quarter of their maximum on very few tokens, which is sparser firing.

The rank on all 18,782 tokens shows the same shape at each step. So Pythia-70m shows the falling rank of Dohare et al., late and in its deeper layers. It does not show their dormant units, by their own rule. This does not show a loss of plasticity, which would need a test of how well the late checkpoints learn new data. It agrees with the late movement of the VERIFIED poles above.

10 Discussion

The first version found that the two measures of formation order its 9 poles in reverse (Table 1). The poles whose activations take their final shape early, L4/N812+, L4/N216+ and L4/N566+, were the ones whose explanations settled last. The poles whose explanations held early, L5/N655+ and L5/N674+, took their final shape late. Over the 9, the Spearman correlation between the step from which an explanation holds and the first step with $r \geq 0.8$ is -0.73.

That reversal does not hold for the 85 positive poles. There, the same correlation is 0.24, and with the first step with $r \geq 0.5$ it is 0.25. The two measures are thus weakly related, not opposed, and the 9 were an unusual sample. The weaker claim still stands. Passing the probes and matching the final activations measure different things. A pole can do what its explanation says while much of the rest of its activation still changes. Equally, it can keep its overall shape while the detail that the probes test still moves. Table 2 shows the same at the level of layers. Layer 0's poles take their shape earliest, but layer 3's explanations hold earliest.

This result differs from the induction heads of Olsson et al. (2022), whose formation is visible as a bump in the loss. The poles studied here change without any matching feature in the loss. This suggests that the loss curve is a poor guide to when single neurons form. A measure of the model's local geometry may show these late changes where the loss does not. An example is the local learning coefficient (Lau et al., 2023), as Hoogland et al. (2025) use it.

For a registry of explanations, the result means that an explanation describes the checkpoint that it was written for. An explanation of the final Pythia-70m is not, in general, an explanation of the model at step 16,000, although the two models have almost the same loss.

11 Limitations

- **One model and one run.** We read only the original training run. PolyPythias offers nine more runs of the same model with other seeds (van der Wal et al., 2025). Its neurons are not the same units, though, so a second run needs its own audited poles.
- **Not a random sample.** The 86 poles are the ones that the registry has VERIFIED, and many were drawn by follow-ups and targeted draws. Layers 2 and 5 have fewer poles than the others.
- **One negative pole, and it saturates.** The only negative pole sits at GELU's dip at every checkpoint, so it gives no correlation. The result says nothing about negative poles in general.
- **Probe noise.** Each explanation is tested with a few dozen probes. So a pass rate near a gate can change from one checkpoint to the next by chance alone (see the review of Yan, Oikarinen and Weng in `faithful-stable-explanations.md`). A single failed checkpoint therefore resets the step from which an explanation holds. Of the 23 failures of the 13 explanations that pass at step 16,000 and fail later, 18 are within two probes of a gate (Table 3). We did not resample the probes, so we cannot give an interval for any one pass rate.
- **The explanations were written for the last checkpoint.** An earlier form of a pole may have a different good explanation. So the failure of the recorded one does not mean that the pole had no clean function.
- **The scale is borrowed.** The activation line at each step scales the recorded maximum by the change in the window maximum. A pole whose top activations lie outside the 300 windows gets a less exact line.
- **Dormancy is measured on one corpus sample.** The 18,782 tokens come from 200 documents. A pole that fires on under 1% of these tokens may fire more on other text. The effective rank is not centred, so a shift in a layer's mean activation also changes it.
- **Few late checkpoints.** Between steps 64,000 and 143,000 we read only one checkpoint, so the timing of late changes is coarse. The adjacent correlations double their gap at each step, so they compare unequal intervals.

12 Conclusion

The 86 reviewed neurons of Pythia-70m keep changing for most of training, long after the loss has settled. Adjacent checkpoints show that they keep moving. Fewer than half of their explanations hold at the step where the loss stops improving. Neuron explanations should therefore be read as descriptions of one checkpoint. Three tests would sharpen this result:

- More checkpoints between steps 64,000 and 143,000, to time the late changes.
- The same measurement on OLMo (Groeneveld et al., 2024), whose training run is also public, once it has VERIFIED poles. Also on the PolyPythias runs of this model.
- The local learning coefficient at the same checkpoints, to test whether the late changes line up with a change in it, not in the loss.
- A test of plasticity itself: train the late checkpoints on new data, and compare how fast they learn with the step-4,000 checkpoint, whose rank is highest.

References

Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). *Network dissection: Quantifying interpretability of deep visual representations*. CVPR 2017. arXiv:1704.05796.

Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S. S., Raff, E., Skowron, A., Sutawika, L., & van der Wal, O. (2023). *Pythia: A suite for analyzing large language models across training and scaling*. In Proceedings of the 40th International Conference on Machine Learning (PMLR 202, pp. 2397–2430). PMLR.

Cammarata, N. (2026). *Neural networks are already clean*. Goodfire research essay. already-clean.goodfire.com.

Dohare, S., Hernandez-Garcia, J. F., Lan, Q., Rahman, P., Mahmood, A. R., & Sutton, R. S. (2024). *Loss of plasticity in deep continual learning*. Nature, 632, 768–774. doi:10.1038/s41586-024-07711-7.

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

Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., Jha, A. H., Ivison, H., Magnusson, I., Wang, Y., Arora, S., Atkinson, D., Authur, R., Chandu, K. R., Cohan, A., Dumas, J., Elazar, Y., Gu, Y., Hessel, J., . . . Hajishirzi, H. (2024). *OLMo: Accelerating the science of language models*. ACL 2024. ACL Anthology.

Hoogland, J., Wang, G., Farrugia-Roberts, M., Carroll, L., Wei, S., & Murfet, D. (2025). *Loss landscape degeneracy and stagewise development in transformers*. Transactions on Machine Learning Research. OpenReview.

Lau, E., Furman, Z., Wang, G., Murfet, D., & Wei, S. (2023). *The local learning coefficient: A singularity-aware complexity measure*. Preprint. arXiv:2308.12108.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., . . . Olah, C. (2022). *In-context learning and induction heads*. Transformer Circuits Thread. transformer-circuits.pub.

Roy, O., & Vetterli, M. (2007). *The effective rank: A measure of effective dimensionality*. 15th European Signal Processing Conference (EUSIPCO 2007), Poznań, 606–610. EURASIP.

van der Wal, O., Lesci, P., Müller-Eberstein, M., Saphra, N., Schoelkopf, H., Zuidema, W., & Biderman, S. (2025). *PolyPythias: Stability and outliers across fifty language model pre-training runs*. ICLR 2025. arXiv:2503.09543.

Appendices

Reproduce

The script downloads the 20 checkpoints from Hugging Face. It ran on an M1 Max in 57 minutes. It measures each pole with the review's own code (`submission_unit` and `peak_strength`). It reads the newest record of each pole from `wiki/`, and keeps the poles where that record is VERIFIED. It cuts the windows from the last 200 documents of `corpus.jsonl`. With `--pole`, it measures only the poles named, as the check of the first 9 did.

```
uv run python scripts/neuron_development.py --out papers/neuron-development/results.jsonl
uv run python scripts/dormant_rank.py --out papers/neuron-development/results.dormant.jsonl
uv run python scripts/paper_figures.py
```

`dormant_rank.py` reads every neuron of each layer, so it needs no registry. It took 15 minutes on the M1 Max's CPU, with the checkpoints already downloaded. The token sample for the effective rank uses seed 0.

The poles that a run selects depend on the registry on the day it runs. The `selection` row of `results.jsonl` gives the date and the count.