HiLD 2026: 4th Workshop on High-dimensional Learning Dynamics 1–12

Common Origins, Divergent Destinations: The Development of Cross-Layer Alignment Under GELU and SwiGLU

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

Abstract

Transformer language models coordinate computation across independently parameterized layers through a shared residual stream, yet the developmental process of this coordination remains poorly understood. We track the emergence of inter-layer alignment in six models across four families, finding a common early formation followed by divergent architectural fates. Early in training, adjacent layers align their dominant MLP transformation directions. This alignment forms via a backward cascade originating at the output and propagating toward the input. The persistence of the resulting structure throughout the rest of training is architecturally determined: GELU models (Pythia, BLOOM) maintain weight-level alignment, while SwiGLU models (OLMo-2, TinyLlama) progressively dissolve it. We argue that early establishment of cross-layer alignment is a developmental priority for transformer training, and that the activation-function-dependent divergence is among the most consequential dynamical differences across modern transformer architectures.[1]

1. Introduction

Transformer language models coordinate computation across independent layers through a shared residual stream, which serves as the sole medium of inter-layer communication (Elhage et al., 2021; Vaswani et al., 2017). Recent work has revealed considerable structure in how this communication is organized in trained models. Layers read from and write to low-rank subspaces of the residual stream, forming communication channels whose singular vectors are semantically interpretable (Millidge and Black, 2022; Merullo et al., 2024). The Jacobians of successive transformer blocks align their top singular vectors with one another, a phenomenon first identified in ResNets (Li and Papyan, 2023) and subsequently confirmed in transformers, where it correlates with model performance (Aubry et al., 2024). What this body of work has not characterized is how the static weight-level subspace structure develops across training, and how that development varies with architecture. To fill this gap, we track the assembly of inter-layer alignment across training in six models from four families, producing three core findings.

First, cross-layer alignment forms early. In all models with sufficient checkpoint density, the dominant MLP transformation directions of adjacent layers align well above the random baseline early in training. This precedes the induction-head phase change (Olsson et al., 2022; Yin and Steinhardt, 2025) and most other documented training phenomena.

Second, this alignment propagates backward from the final layers, where the loss gradient provides the most structured signal, and works inward over subsequent training.

1. Code available at https://anonymous.4open.science/r/communicative-backbone-BCAA/.

© .

Third, the long-term fate of this alignment depends on the activation function. GELU models maintain the alignment for the remainder of training, while SwiGLU models progressively dissolve it.

Together, these findings suggest that establishing cross-layer alignment is a developmental priority shared across architectures, with the activation function determining whether the resulting structure is maintained or progressively dissolved.

2. Setup

We analyze checkpoints from six models spanning four families. Pythia-410M, Pythia-1B, and Pythia-1.4B (Biderman et al., 2023) share the Pile training corpus and tokenizer, differing only in width and depth, with dense early checkpoints (steps 0, 128, 512, 2,000, 8,000, 32,000, 64,000, 143,000). OLMo-2-1B (OLMo Team et al., 2024) is trained on Dolma 1.7 with similarly dense early checkpoints, as well as late checkpoints up to step 1,000,000. BLOOM-1.1B (Le Scao et al., 2022) and TinyLlama-1.1B (Zhang et al., 2024) provide late-training checkpoints only. Pythia and BLOOM use GELU; OLMo-2 and TinyLlama use SwiGLU. Pythia uses parallel attention-MLP layout; the others are sequential. Together, these models span the relevant axes of variation while keeping parameter counts within an order of magnitude.

For each transformer layer, the MLP composed product $W_{\text{down}}W_{\text{up}}$ is a square matrix in the residual stream's ambient space whose top singular vectors identify the directions along which the MLP reads from and writes to the residual stream. For each pair of layers, we compute the cosines of the principal angles between their top-k singular vector subspaces and compare these to the expected overlap of random subspaces of the same dimensionality in the same ambient space. This is our probe of inter-layer alignment.

3. Cross-layer alignment forms early and propagates backward

We begin with Pythia, which provides the densest early checkpoints. At initialization, adjacent layers' top singular vector subspaces overlap no more than random chance predicts. Between steps 128 and 512, this overlap surges in all three Pythia models (Figure 1). This surge continues until step 2,000, where it then stabilizes or mildly recedes.

Backward cascade. Inter-layer alignment emerges as a cascade originating where the training signal is strongest, propagating backward (Figure 2). Across all three Pythia models, the coupling between the final two layers peaks first, at step 512. Other layer boundaries show a similar increase, reaching their initial peaks at step 2,000 or later, with their magnitudes at step 2,000 weakening with distance from the final-pair boundary ($r = 0.80, 0.85, 0.84$ across the three Pythia models; all $p < 10^{-3}$).

Both endpoints of the network show early rises: the final layers, anchored by the loss gradient, and the first two boundaries, anchored by structured token embeddings. Only the final-layer coupling holds, and the cascade extends backward from there; the early-boundary coupling falls back toward baseline and does not recover until well into training (Figure 3 in Appendix A). The gradient signal, which carries the structure of the final layers' aligned subspaces, appears to progressively overwrite whatever alignment the early layers initially

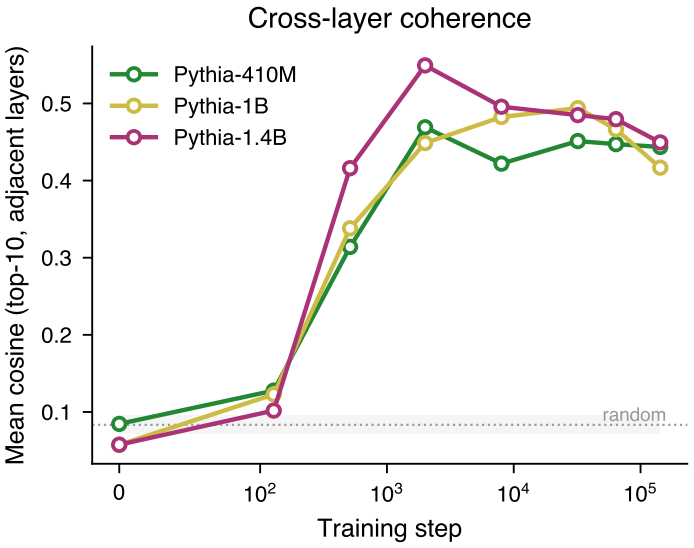

Figure 1: Mean adjacent-layer top-10 subspace overlap across training for Pythia-410M, Pythia-1B, and Pythia-1.4B, compared to random baseline. Alignment grows slowly from steps 0 to 128, then rapidly accelerates between steps 128 and 512.

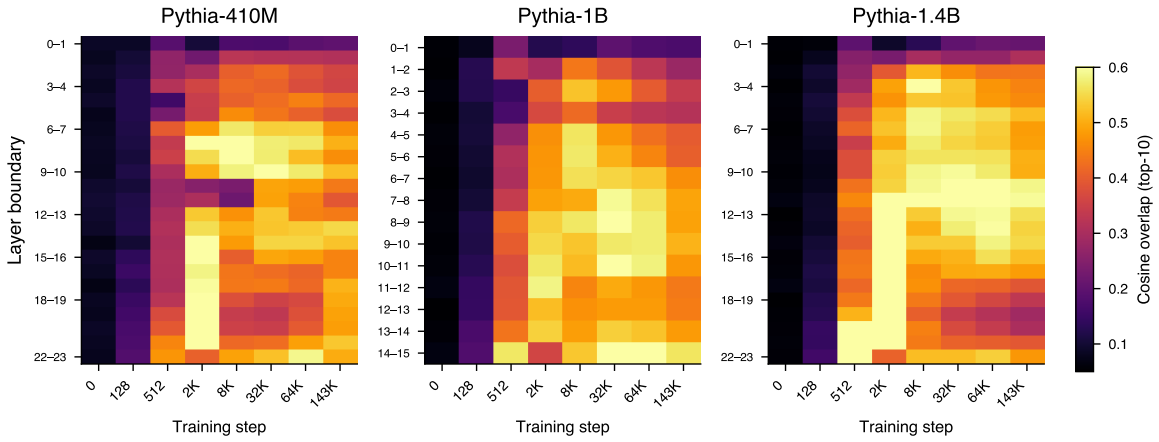

Figure 2: Cross-layer coherence resolved by depth for Pythia-410M, Pythia-1B, and Pythia-1.4B. Top-10 subspace overlap between adjacent layers across training, shown as a heatmap with layer boundary on the y-axis and training step on the x-axis. Coherence rises first at the final layer boundaries and propagates backward over subsequent training.

found. The pairwise subspace overlap between all layer pairs provides a clearer view of this backward propagation (Figure 4 in Appendix A).

Model	Step	Pairs > 0.15	Mean d=1	Mean d=3	Mean d=5
GELU models (maintain)					
Pythia-410M	8,000	188/276 (68%)	0.422	0.304	0.242
	143,000	207/276 (75%)	0.444	0.339	0.265
Pythia-1B	8,000	93/120 (78%)	0.482	0.342	0.253
	143,000	104/120 (87%)	0.417	0.313	0.233
Pythia-1.4B	8,000	225/276 (82%)	0.496	0.387	0.301
	143,000	252/276 (91%)	0.449	0.350	0.278
BLOOM-1.1B	1,000	276/276 (100%)	0.733	0.481	0.395
	300,000	224/276 (81%)	0.600	0.398	0.287
	600,000	228/276 (83%)	0.569	0.379	0.269
SwiGLU models (decohere)					
OLMo-2-1B	10,000	64/120 (53%)	0.472	0.293	0.188
	100,000	41/120 (34%)	0.369	0.196	0.120
	1,000,000	24/120 (20%)	0.256	0.100	0.077
TinyLlama-1.1B	50,000	165/231 (71%)	0.441	0.304	0.224
	480,000	73/231 (32%)	0.318	0.195	0.137
	1,431,000	67/231 (29%)	0.304	0.174	0.128

Table 1: Pairwise top-10 subspace overlap across late training for all six models. "Pairs > 0.15" counts the fraction of all layer pairs with mean cosine similarity above 0.15 (random baseline is $\sim$0.06). Mean overlap at distance d is the average top-10 subspace overlap between layer pairs separated by d layers. All four GELU models increase or stabilize their alignment over training; both SwiGLU models lose alignment monotonically. BLOOM and OLMo share sequential attention-MLP ordering, while BLOOM and Pythia share GELU activation; the divergence tracks activation function rather than ordering.

4. Architectural fate of the alignment

The early alignment surge is not specific to Pythia. OLMo-2-1B exhibits the same surge, occurring between steps 1,000 and 6,000 rather than 128 and 2,000, and depth-resolved overlap trajectories show the same backward cascade (Appendix B).

What is specific to OLMo is what happens after the surge. The alignment that Pythia maintains is, in OLMo, progressively lost. OLMo and Pythia differ on three relevant axes: activation function (SwiGLU versus GELU), attention-MLP ordering (sequential versus parallel), and training duration (143,000 versus 1,000,000 steps). To isolate which axis is responsible, we extend the comparison to TinyLlama-1.1B (SwiGLU, sequential) and BLOOM-1.1B (GELU, sequential). TinyLlama tests whether the SwiGLU dissolution replicates outside OLMo's family; BLOOM tests whether dissolution is a consequence of sequential ordering; and both test whether dissolution arises from extended training.

Our results separate the three variables (Table 1). All four GELU models increase or stabilize their fraction of meaningfully aligned layer pairs over the course of training: Pythia-410M, Pythia-1B, and Pythia-1.4B rise to 75%, 87%, and 91% respectively, and BLOOM stabilizes at 81–83% after an initial contraction. Both SwiGLU models lose align-

ment monotonically, with the fraction of meaningfully aligned pairs falling to 20% (OLMo) and 29% (TinyLlama) by their respective final checkpoints, while overlap at distance 5 approaches the random baseline. BLOOM, despite the sequential architecture it shares with OLMo, maintains its alignment rather than dissolving it, supporting the hypothesis that the persistence of alignment depends on the activation function rather than attention-MLP ordering.

BLOOM also helps us separate activation function from training duration: if decoherence were simply what happens given enough steps, we would expect it in BLOOM's later checkpoints, but we instead see stability.

5. Discussion

The six models we examined reveal a common developmental sequence followed by a divergent trajectory. In all of the Pythia and OLMo models we tracked, cross-layer MLP alignment begins forming early in training, developing as a backward cascade starting at the final layers. This is likely one of the first coherent structures the model develops; in Pythia, adjacent-layer alignment surges above random baseline between steps 128 and 512, while induction heads first emerge around step 1,000 (Olsson et al., 2022; Yin and Steinhardt, 2025) and function-vector heads around step 16,000 (Yin and Steinhardt, 2025).

After this shared early window, the architectures diverge along their activation functions. GELU models maintain weight-level alignment, while SwiGLU models progressively lose it. These results suggest that, regardless of activation function, inter-layer coordination is first established via static weight subspaces. In SwiGLU models, this responsibility may shift to the input-dependent gate, constructing the effective shared subspace dynamically rather than statically. This reframes the activation function as a determinant of the inter-layer coordination strategy, rather than a mere lever of expressivity.

5.1. Limitations

All six models in this study are at or below 1.4B parameters. The developmental sequence we document could differ at scale, where the residual stream dimensionality is much larger and the ratio of computational dimensions to communication bandwidth shifts. We have a single training run per model; the precise timing of the alignment window may vary across seeds. Our weight-level measurement directly captures the static MLP transformation but only indirectly characterizes the gate-conditioned effective transformation in SwiGLU. A direct measurement of the input-conditioned $W_{\text{down}}\,\text{diag}(g(\mathbf{x}))W_{\text{up}}$ across layers would give a cleaner picture of late-stage SwiGLU dynamics, and is the natural next step. Finally, we do not establish a causal link between the developmental phases we document and the acquisition of specific model capabilities; whether the backward cascade we observe is the structural prerequisite for the later phase changes documented in the literature remains open.

References

Murdock Aubry, Haoming Meng, Anton Sugolov, and Vardan Papyan. Transformer block coupling and its correlation with generalization in LLMs. *arXiv preprint*

arXiv:2407.07810, 2024.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory 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 *Proceedings of the 40th International Conference on Machine Learning*, pages 2397–2430, 2023.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. BLOOM: A 176B-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

Jianing Li and Vardan Papyan. Residual alignment: Uncovering the mechanisms of residual networks. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Talking heads: Understanding inter-layer communication in transformer language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Beren Millidge and Sid Black. The singular value decompositions of transformer weight matrices are highly interpretable. *AI Alignment Forum*, 2022. URL https://www.alignmentforum.org/posts/mkbGjzxD8d8XqKHzA/the-singular-value-decompositions-of-transformer-weight.

OLMo Team, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, et al. 2 OLMo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Kayo Yin and Jacob Steinhardt. Which attention heads matter for in-context learning? In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. TinyLlama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024.

Appendix A. Per-boundary and pairwise cross-layer coherence in Pythia

This appendix collects two additional views of the cross-layer coherence cascade documented for Pythia in Section 3. Figure 3 resolves the heatmap of Figure 2 of the main text into per-boundary trajectories, making the shared rise-fall-recovery pattern at the layer boundaries more apparent: the boundaries closest to the input (dark blue) and those closest to the output (dark red) rise sharply from steps 128 to 512, while the others show a sharp rise from steps 512 to 2,000. Figure 4 extends the view from adjacent boundaries to all layer pairs, showing the backward cascade across the full layer-pair structure: a block of mutually aligned final layers forms first and progressively extends toward the input end of the network.

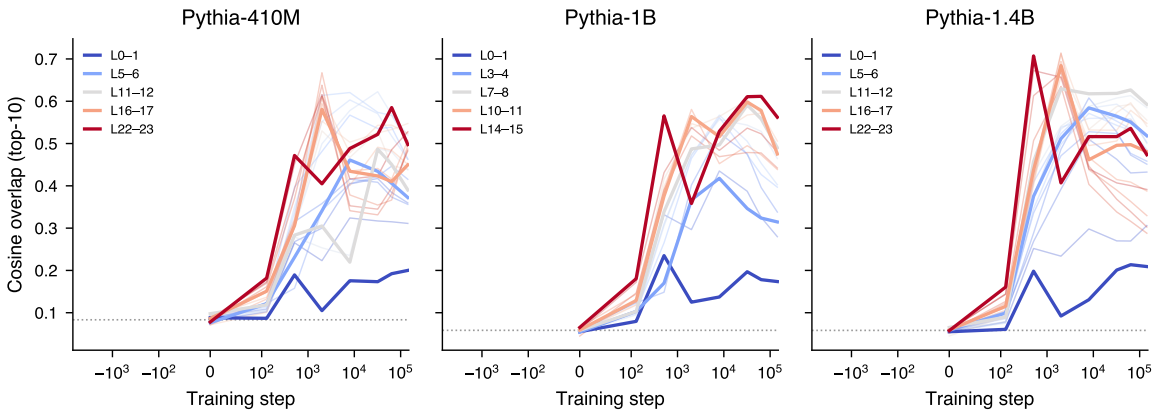

Figure 3: Top-10 subspace overlap between adjacent layers across training for Pythia-410M, Pythia-1B, and Pythia-1.4B, shown per layer boundary. Boundaries are colored by depth, from early layers (blue) to final layers (red).

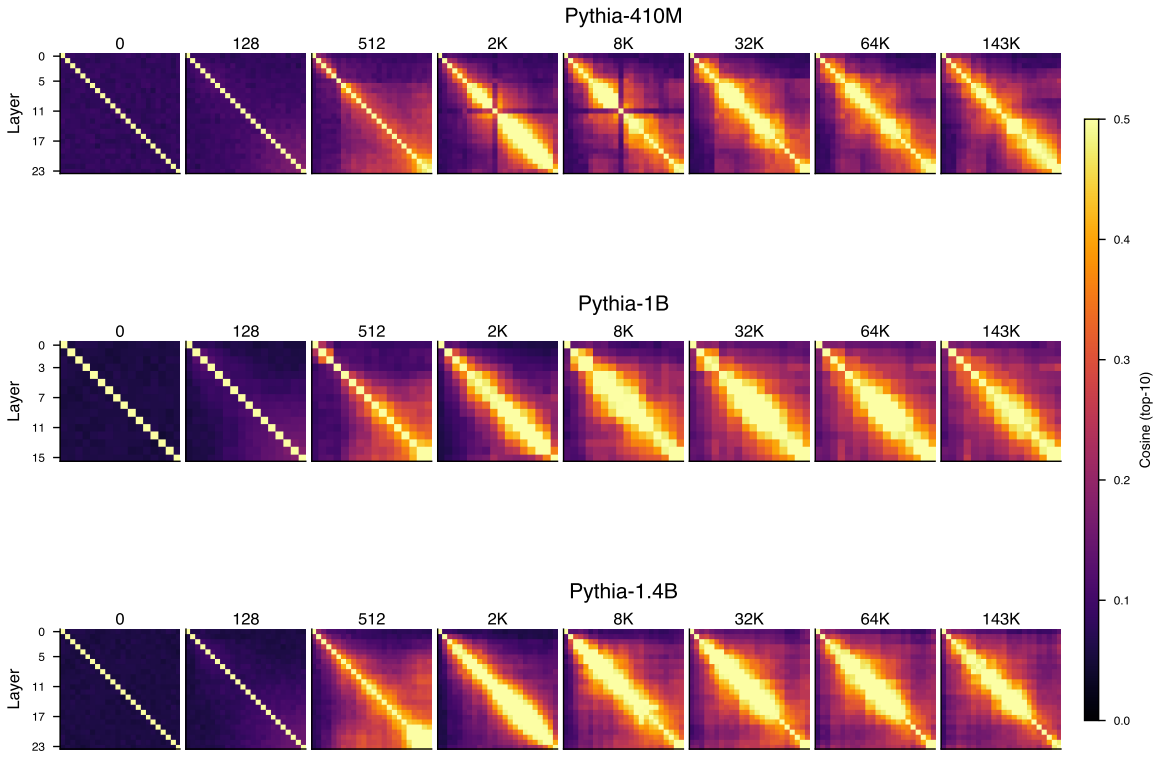

Figure 4: Pairwise top-10 subspace overlap between all layer pairs across training for Pythia-410M, Pythia-1B, and Pythia-1.4B, at each checkpoint from initialization through step 143,000. A block of mutual alignment forms among the final layers and progressively extends toward the input end of the network, mirroring the backward cascade visible at adjacent boundaries in Figure 2 of the main text.

Appendix B. OLMo-2-1B replication of the alignment surge and backward cascade

This appendix collects figures showing that the alignment surge and backward cascade documented for Pythia in Section 3 also appear in OLMo-2-1B, despite differences in activation function, attention-MLP ordering, training corpus, and tokenizer. The dissolution that follows the surge in OLMo is summarized quantitatively in Table 1 of the main text.

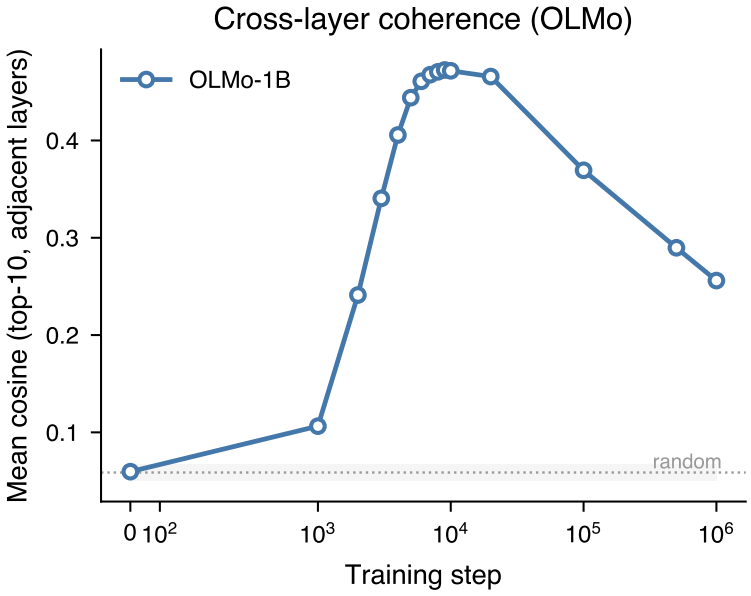

Figure 5: Mean adjacent-layer top-10 subspace overlap across training for OLMo-2-1B, compared to random baseline. The same early surge in cross-layer alignment observed in the Pythia models (Figure 1) appears here, with the surge occurring between steps 1,000 and 6,000 rather than 128 and 2,000.

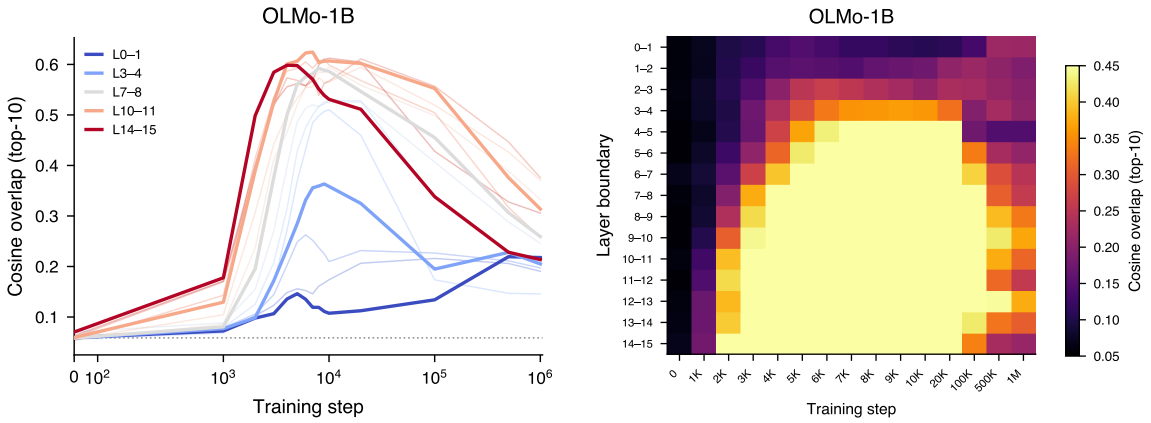

Figure 6: Cross-layer coherence resolved by depth for OLMo-2-1B. Left: top-10 subspace overlap between adjacent layers across training, shown per layer boundary, with boundaries colored by depth from early layers (blue) to final layers (red). Right: the same data as a heatmap, with layer boundary on the y-axis and training step on the x-axis. As in the Pythia models (Figure 2), the surge begins at the final layer boundaries and cascades backward through the network.

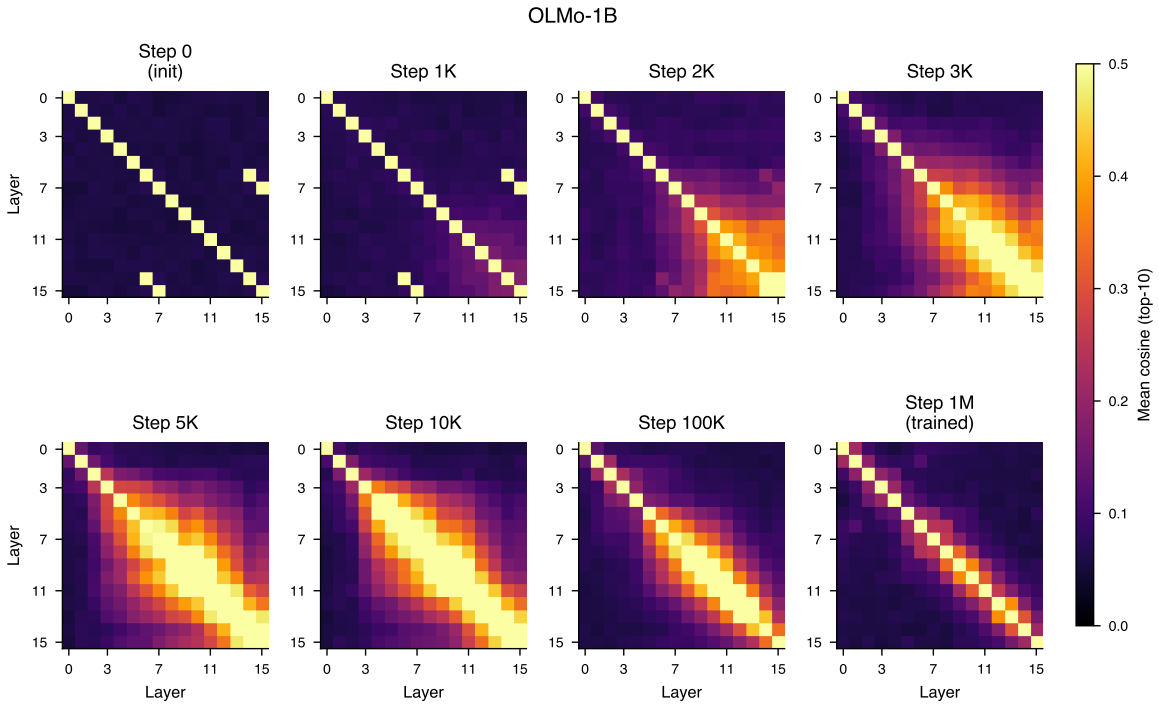

Figure 7: Pairwise top-10 subspace overlap between all layer pairs across training for OLMo-2-1B, at each checkpoint from initialization to step 1,000,000. A block of mutual alignment forms among the final layers and extends toward the input end, as in the Pythia models, before progressively dissolving over the subsequent course of training.