Fresh in memory: Training-order recency is linearly encoded in LLM activations

Dmitrii Krasheninnikov

March 2026

FRESH IN MEMORY: TRAINING-ORDER RECENCY IS LINEARLY ENCODED IN LANGUAGE MODEL ACTIVATIONS

Dmitrii Krasheninnikov[1] **Richard E. Turner**[1] **David Krueger**[2]

ABSTRACT

We show that language models' activations linearly encode when information was learned during training. Our setup involves creating a model with a known training order by sequentially fine-tuning Llama-3.2-1B on six disjoint but otherwise similar datasets about named entities. We find that the average activations of test samples corresponding to the six training datasets encode the training order: when projected into a 2D subspace, these centroids are arranged exactly in the order of training and lie on a straight line. Further, we show that linear probes can accurately ($\sim$90%) distinguish "early" vs. "late" entities, generalizing to entities unseen during the probes' own training. The model can also be fine-tuned to explicitly report an unseen entity's training stage ($\sim$80% accuracy). Notably, the training-order encoding does not seem attributable to simple differences in activation magnitudes, losses, or model confidence. Our paper shows that models can differentiate information by its acquisition time, and carries significant implications for how they might manage conflicting data and respond to knowledge modifications.

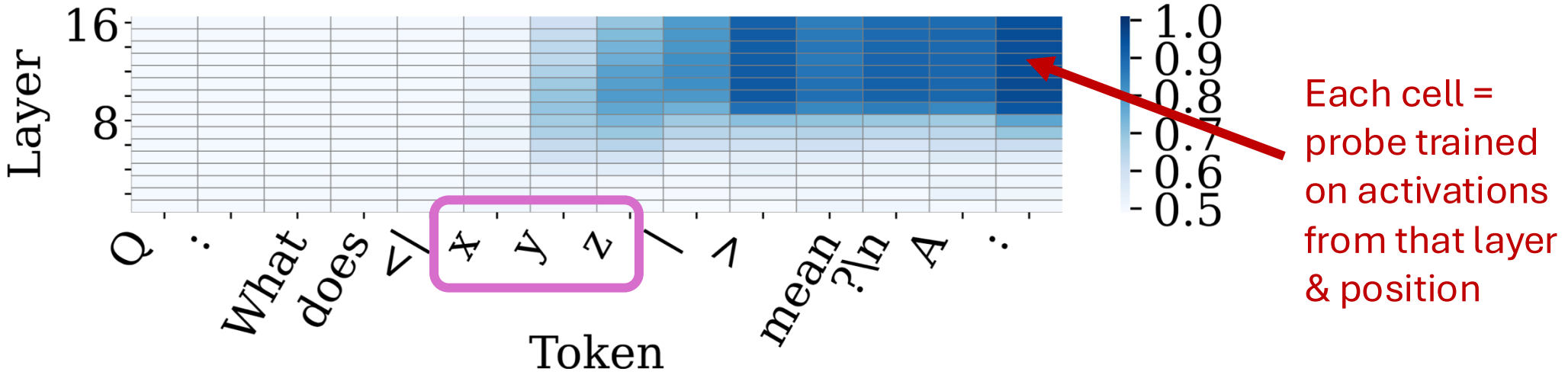
In a model with known training order, probes distinguish training stages

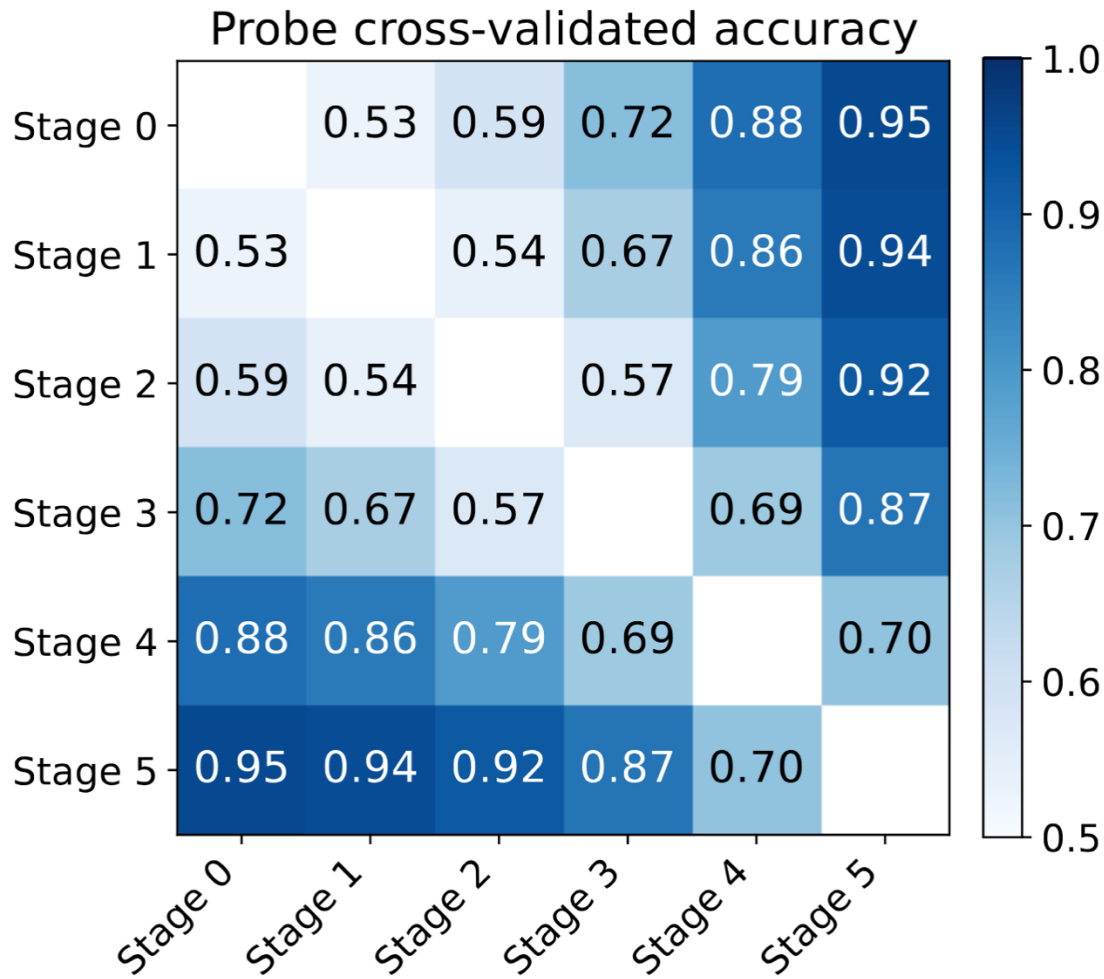
- Setup: finetune on Dataset 1, then on 2, etc -- up to Dataset 6

  $(D_1 \rightarrow D_2 \rightarrow D_3 \rightarrow D_4 \rightarrow D_5 \rightarrow D_6)$

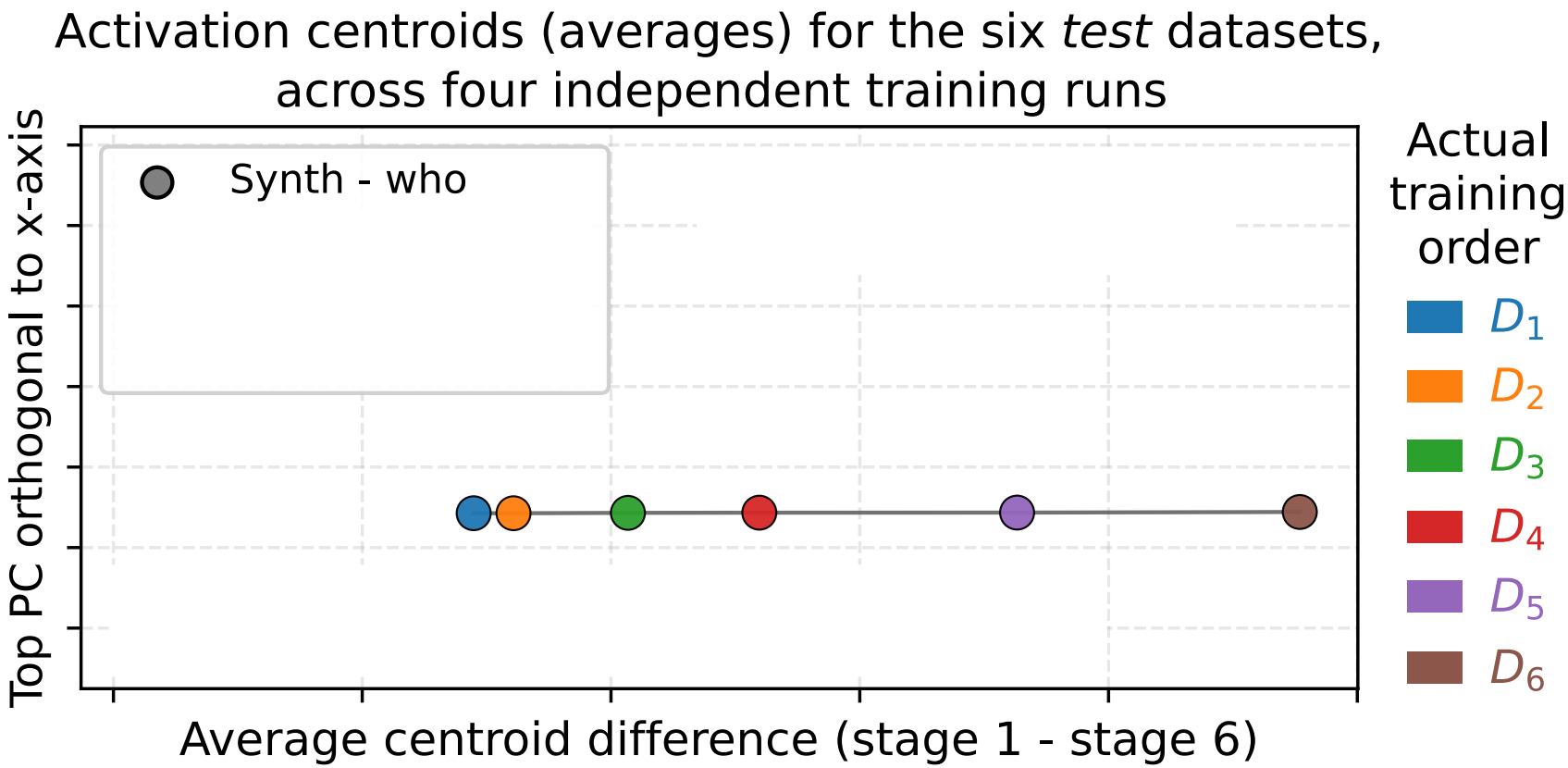
  - 2600 entities in each, non-overlapping
  - **Gives us a model with known training order**

- Probe gets >90% acc distinguishing samples about held-out entities from D_1 / D_6

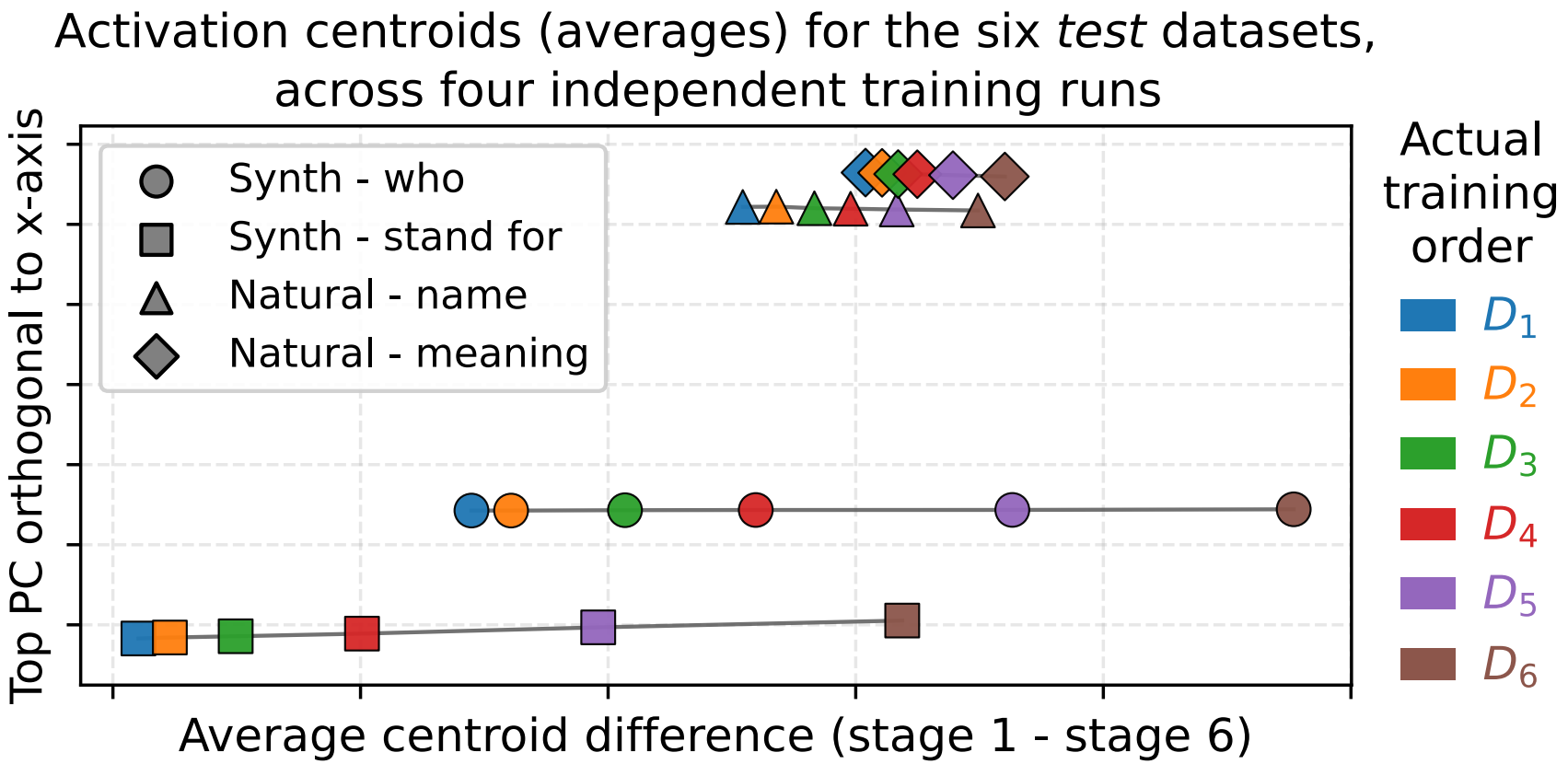



Probe cross-validated accuracy
Stage 0
Stage 1
Stage 2
Stage 3
Stage 4
Stage 5
0.53 0.59 0.72 0.88 0.95
0.53 0.54 0.67 0.86 0.94
0.59 0.54 0.57 0.79 0.92
0.72 0.67 0.57 0.69 0.87
0.88 0.86 0.79 0.69 0.70
0.95 0.94 0.92 0.87 0.70
1.0
0.9
0.8
0.7
0.6
0.5

Training-order recency is encoded in activations



Training-order recency is encoded in activations



Check the paper for

- Results showing that “training-order recency” is a good interpretation
- Models can report the training stage in output tokens when finetuned to do so
- Mechanistic study of what the discovered axis encodes
- And more!