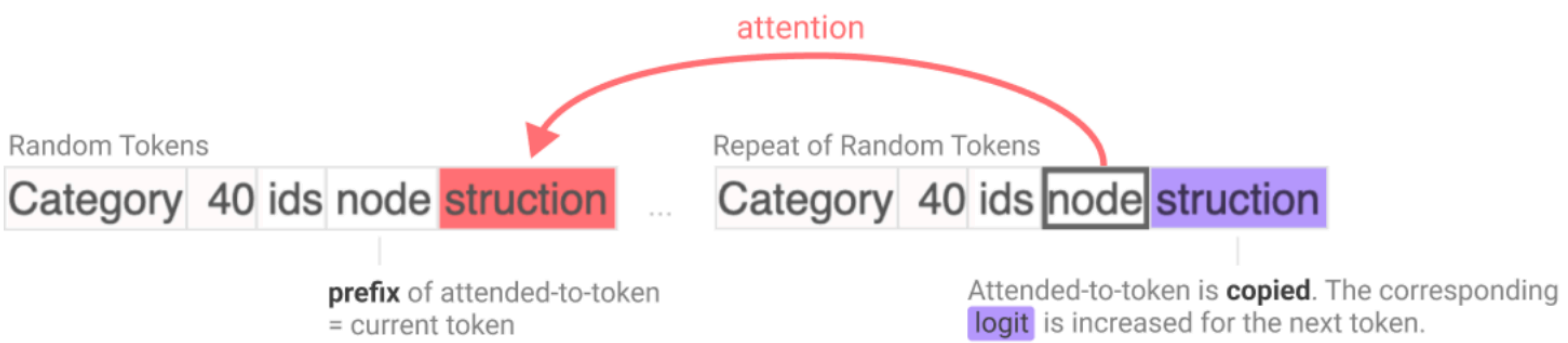
Background: Transformer Circuits

- There are mechanisms in Transformers to do “fuzzy” or “nearest neighbor” versions of pattern completion, completing [A*][B*] … [A] → [B] , where A* ≈ A and B* ≈ B are similar in some space
- Olsson et al. want to establish that these mechanisms are responsible for good ICL capabilities
- We can find these heads and see that performance improves; can we causally link these?

Olsson et al. (2022)

Induction Heads

- Induction heads: a pair of attention heads in different layers that work together to copy or complete patterns.
- The first head copies information from the previous token into each token.
- Second attention head to attend to tokens based on what happened before them, rather than their own content. Likely to “look back” and copy next token from earlier
- The two heads working together cause the sequence …[A][B]…[A] to be more likely to be completed with [B].

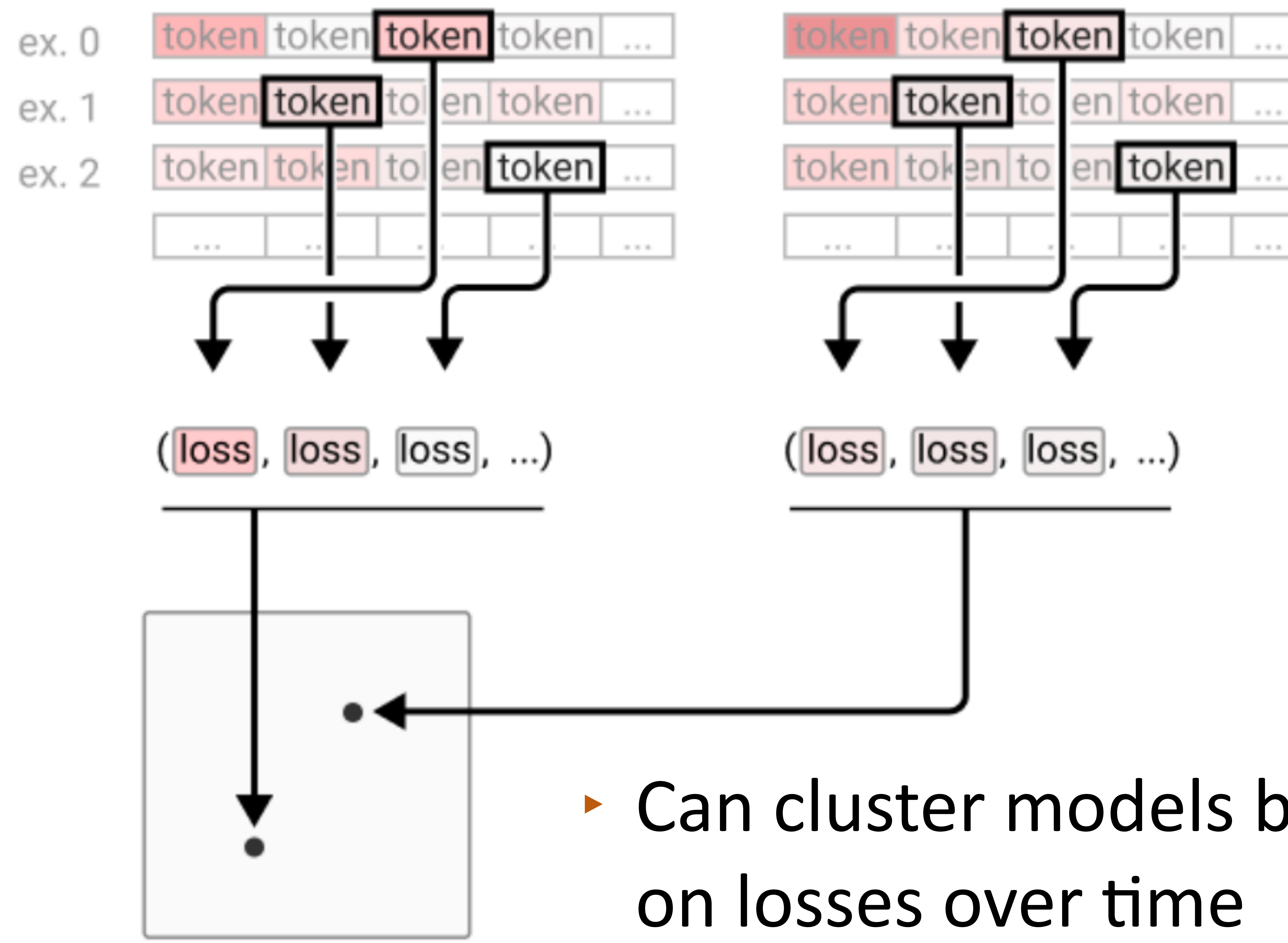


Induction Heads

Step 1: Run each model / snapshot over the same set of multiple dataset examples, collecting one token's loss per example.

Step 2: For each sample, extract the loss of a consistent token. Combine these to make a vector of losses per model / snapshot.

Step 3: The vectors are jointly reduced with principal component analysis to project them into a shared 2D space.



- Can cluster models based on losses over time

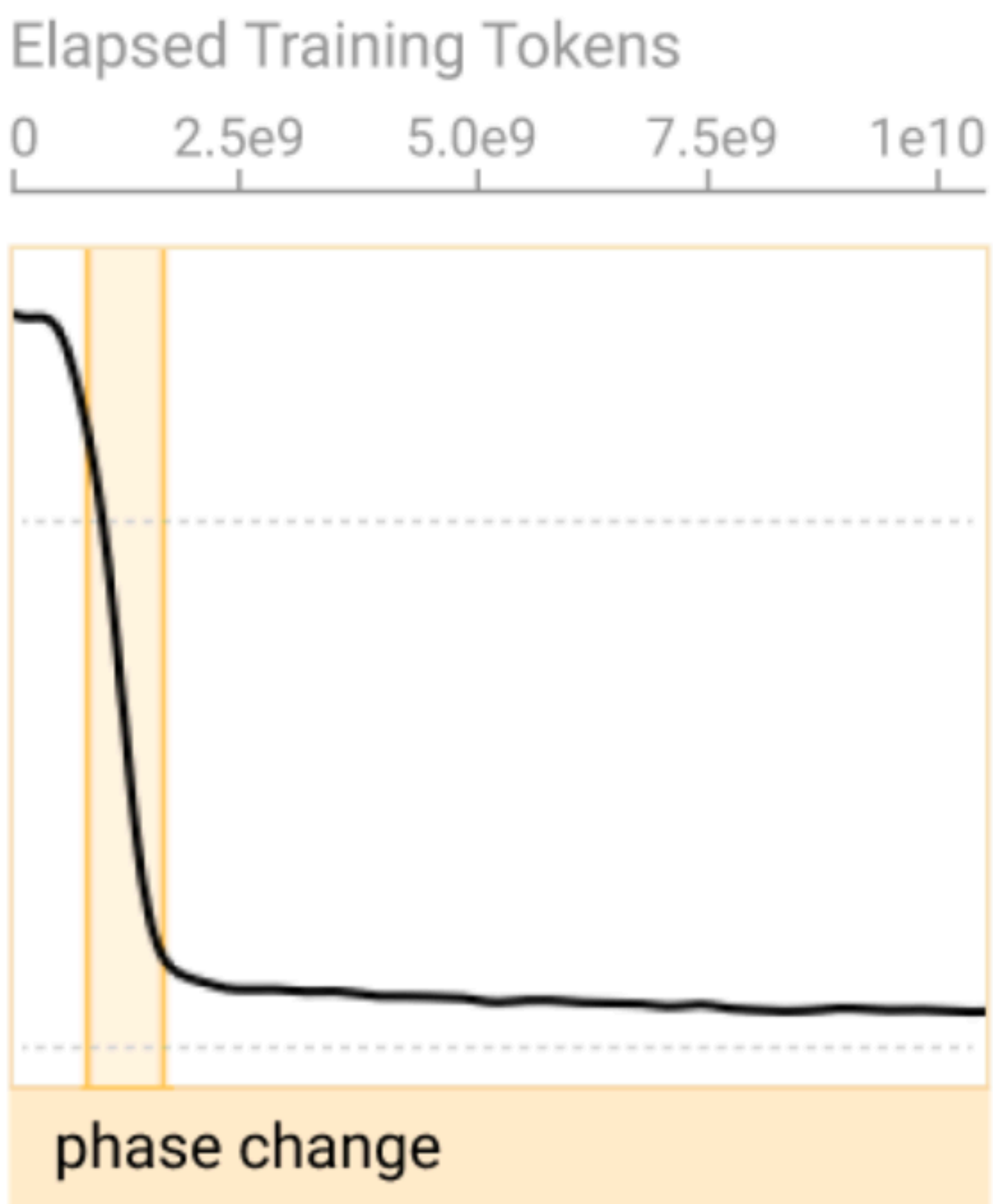
- Characterize performance by ICL score: loss(500th token) - loss(50th token) — average measure of how much better the model is doing later once it's seen more of the pattern

Induction Heads

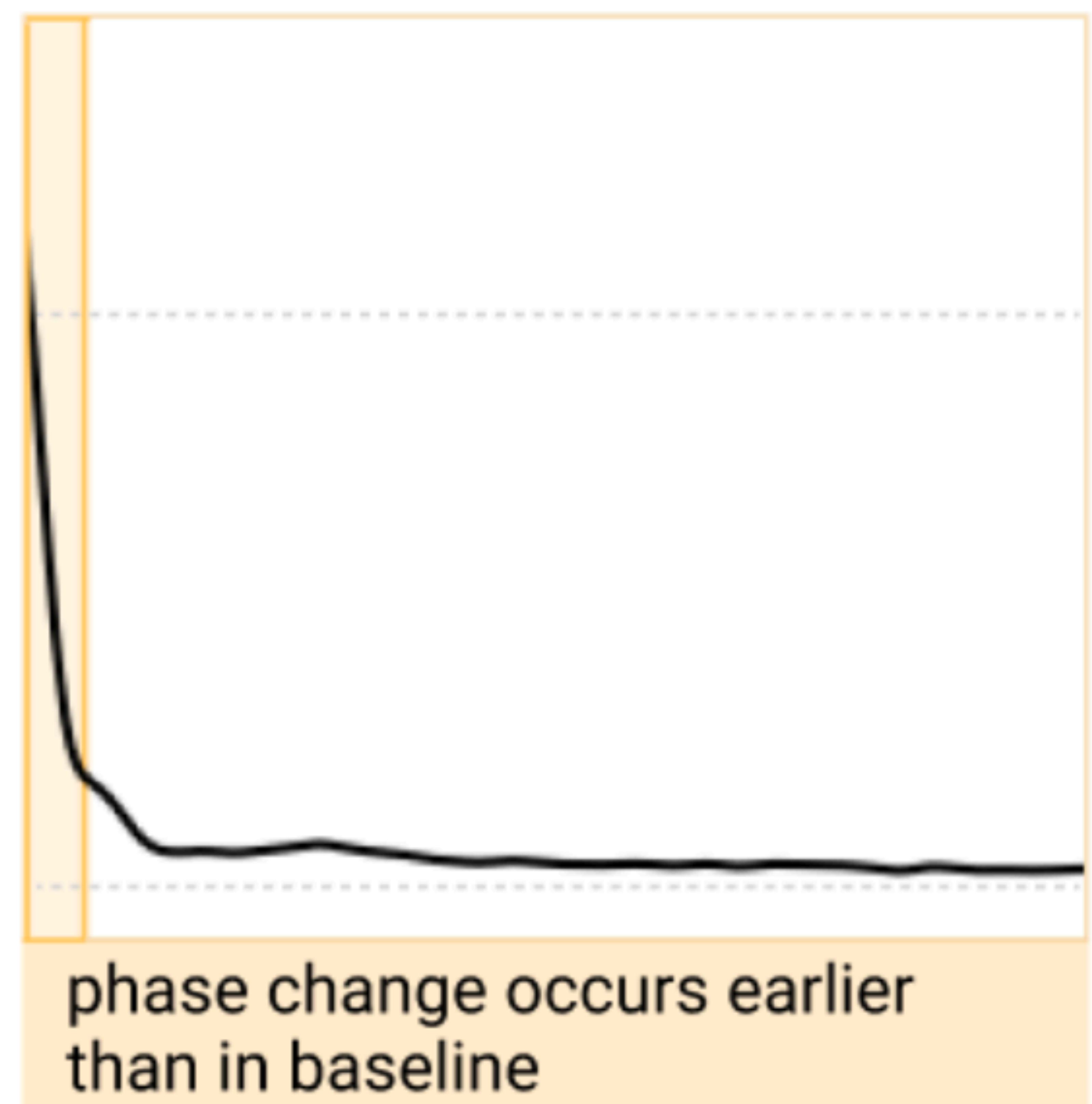
- Improvement in ICL (loss score) correlates with emergence of induction heads

Olsson et al. (2022)

Induction Heads



Change architecture to promote induction heads => phase change happens earlier



Olsson et al. (2022)

Induction Heads

- If you remove induction heads, behavior changes dramatically

Olsson et al. (2022)

Interpretability

- Lots of explanations for why ICL works — but these haven't led to many changes in how Transformers are built or scaled
- Several avenues of inquiry: theoretical results (capability of these models), mechanistic interpretability, fully empirical (more like that next time)