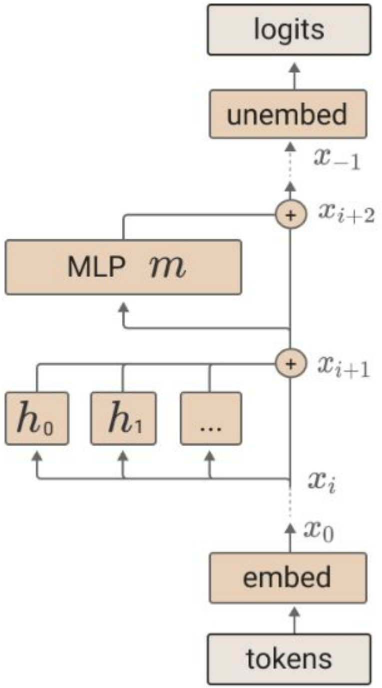
AI Alignment Cohort
Session 8

Introduction to Mechanistic Interpretabilit y

TransformerLens & Induction circuits

Transformer Architecture



The final logits are produced by applying the unembedding.

$$T(t) = W_U x_{-1}$$

An MLP layer, m, is run and added to the residual stream.

$$x_{i+2} = x_{i+1} + m(x_{i+1})$$

Each attention head, h, is run and added to the residual stream.

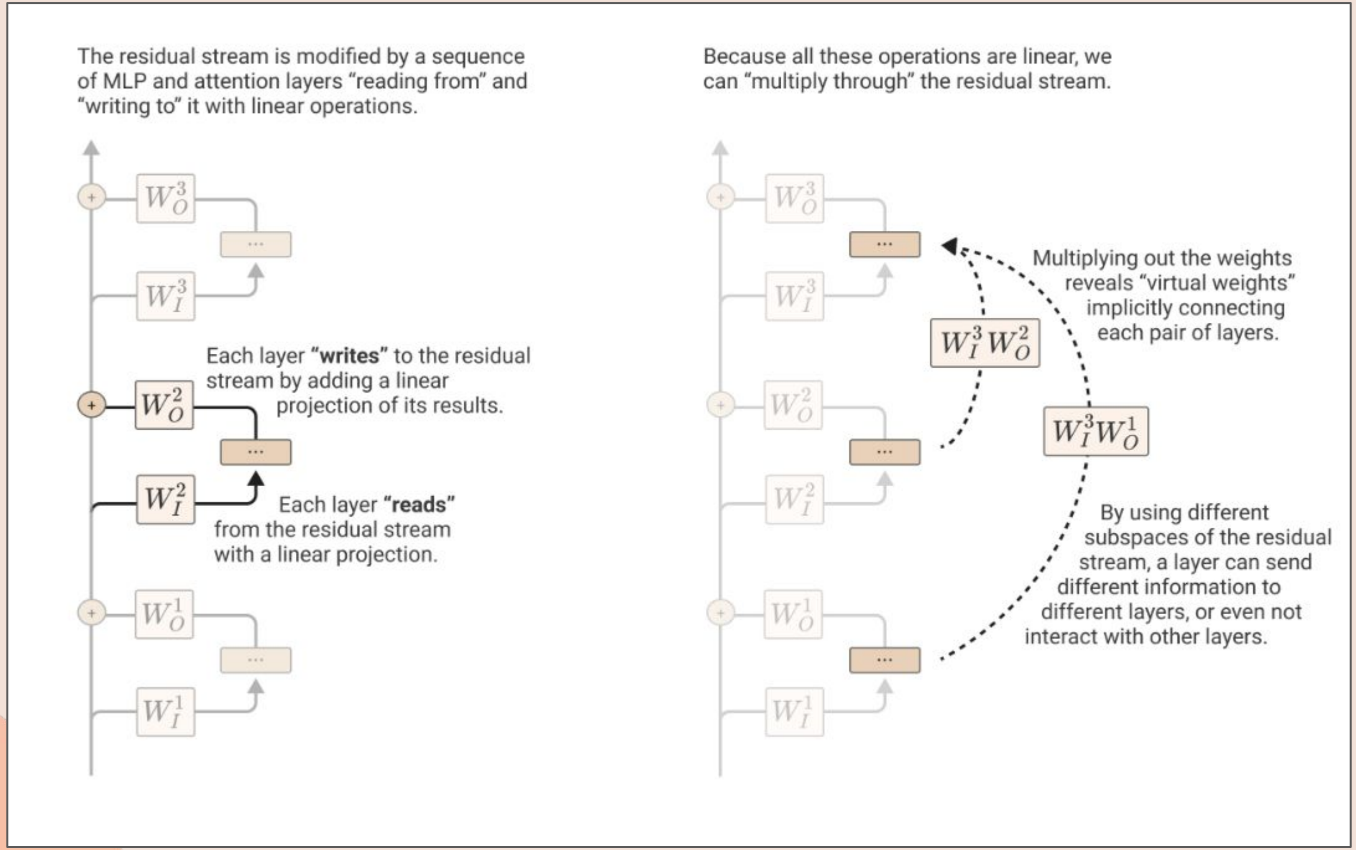
$$x_{i+1} = x_i + \sum_{h \in H_i} h(x_i)$$

One residual block

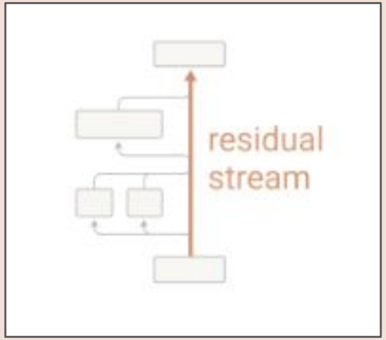
Token embedding.

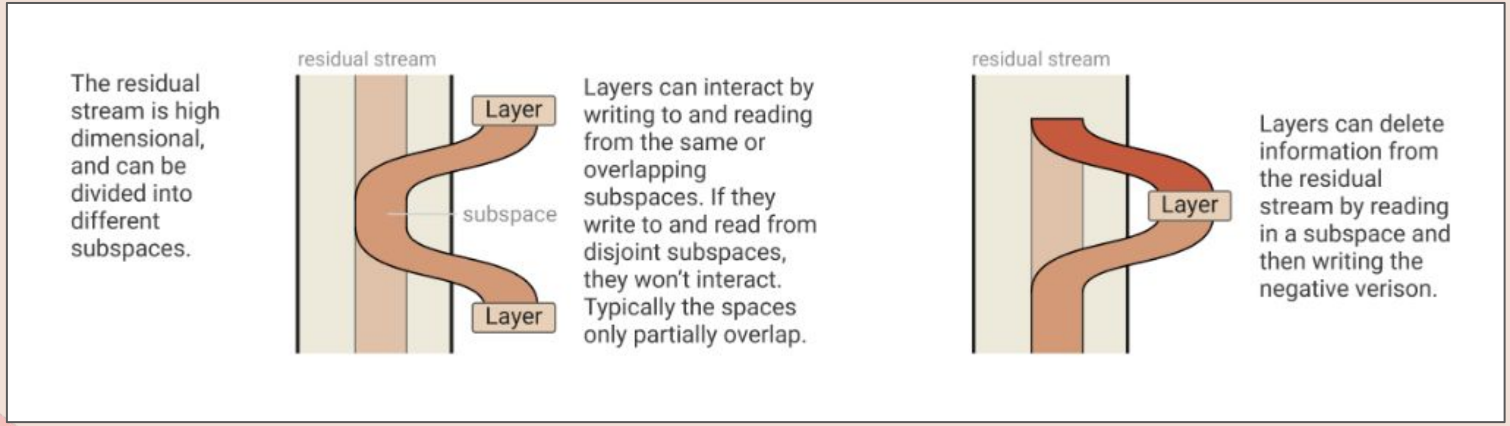
$$x_0 = W_E t$$

Residual Stream

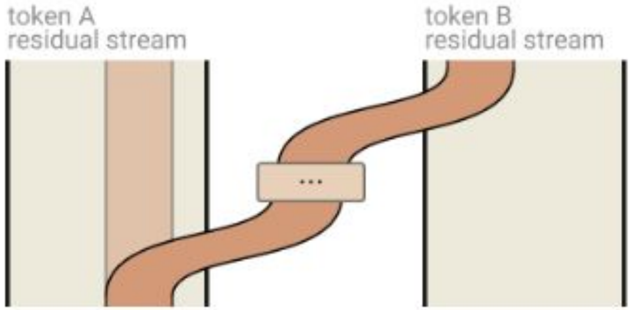


Residual Stream



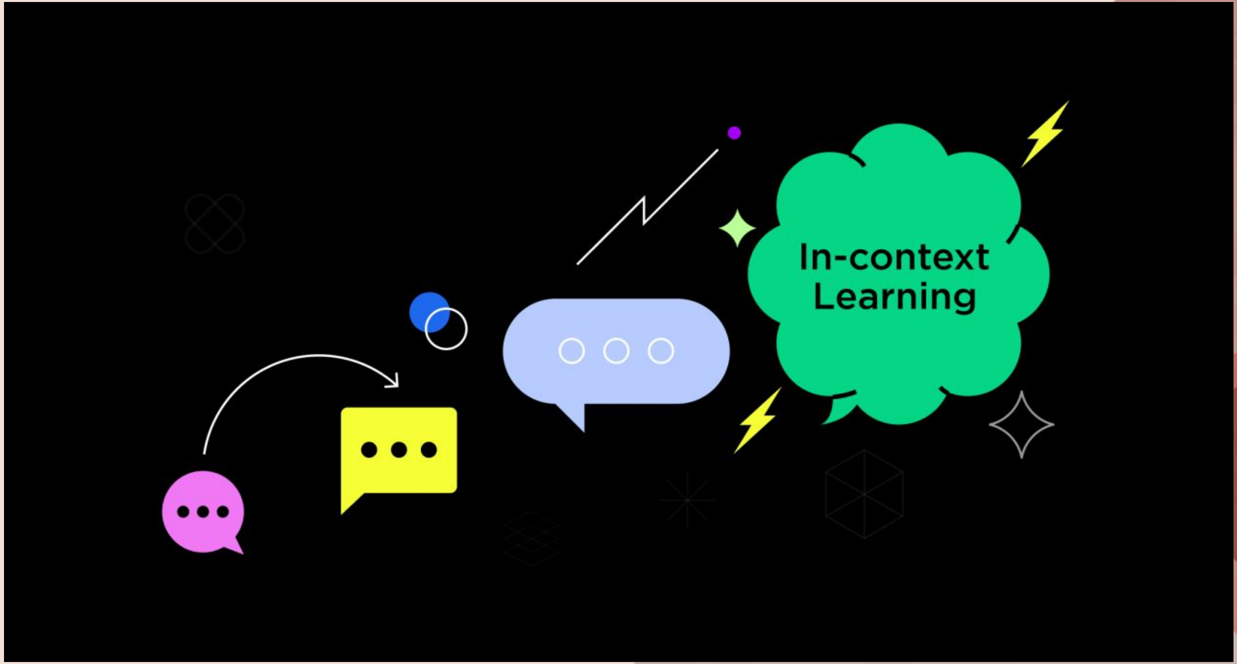


Attention Heads as Information Movement



Attention heads copy information from the residual stream of one token to the residual stream of another. They typically write to a different subspace than they read from.

In-Context Learning

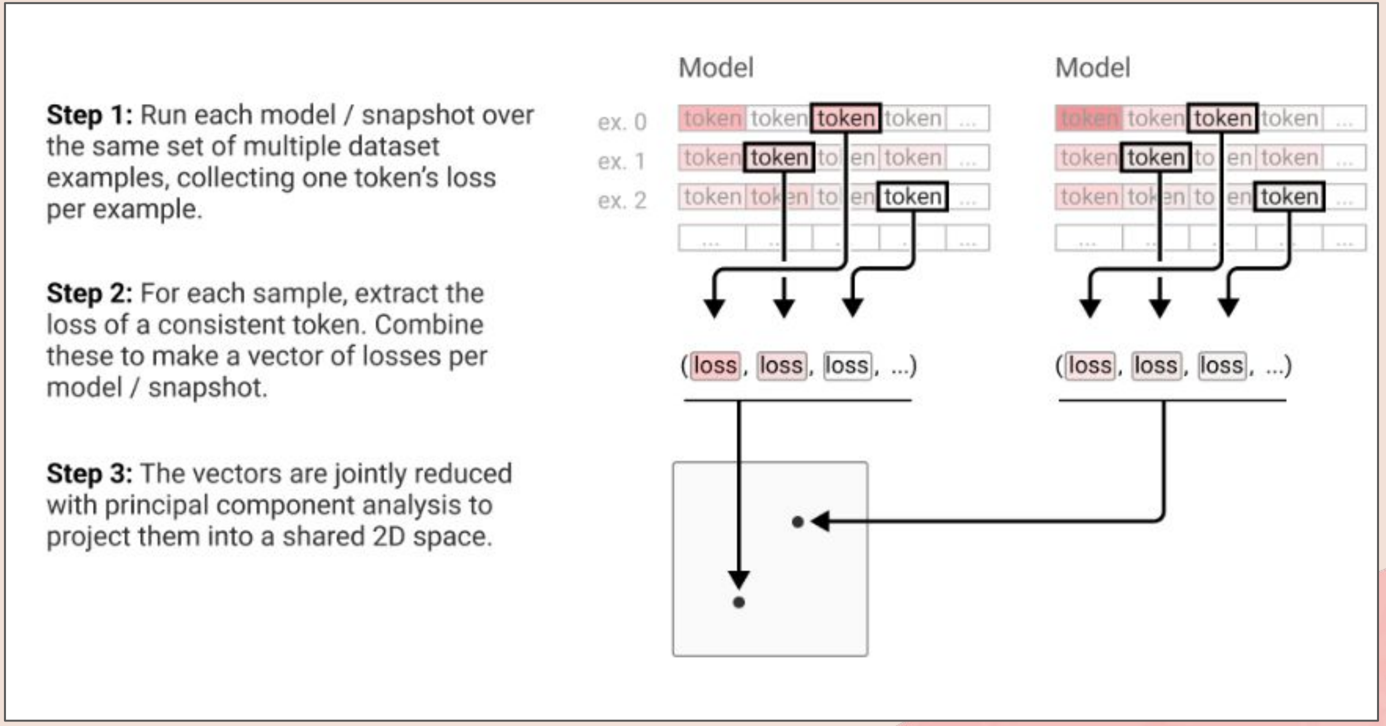


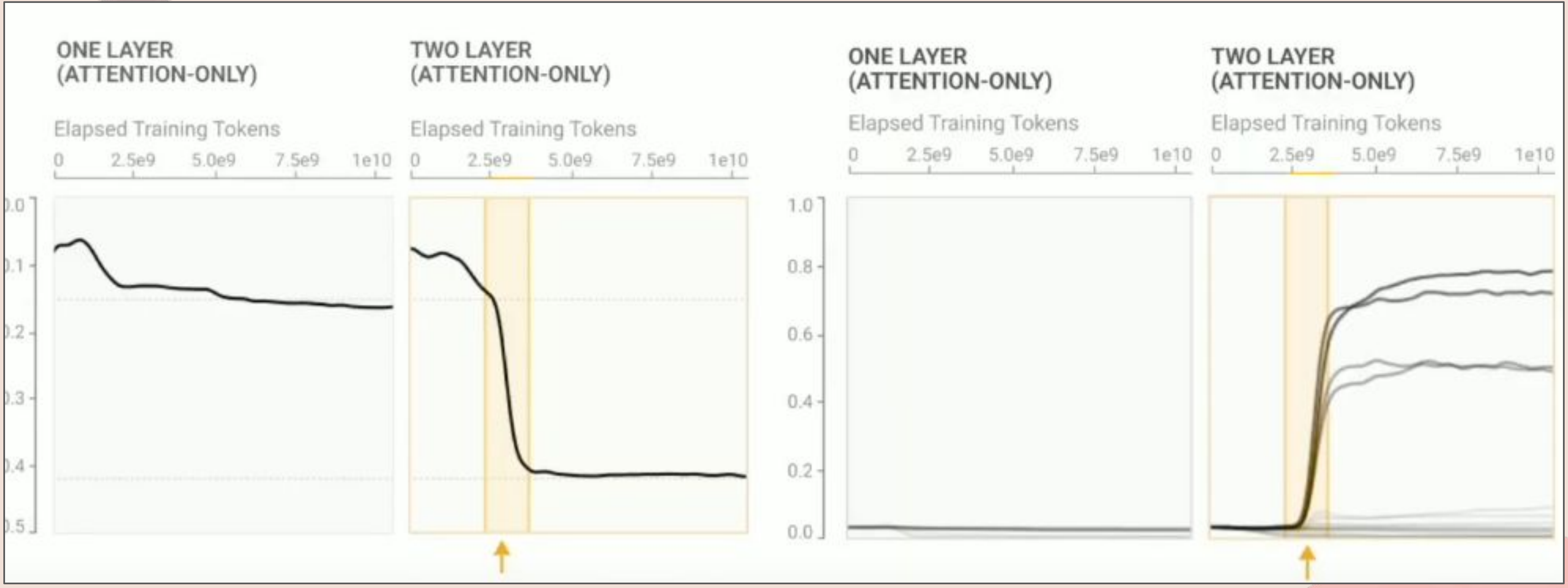
Induction Heads

Induction heads are implemented by a circuit consisting of a pair of attention heads in different layers that work together to copy or complete patterns

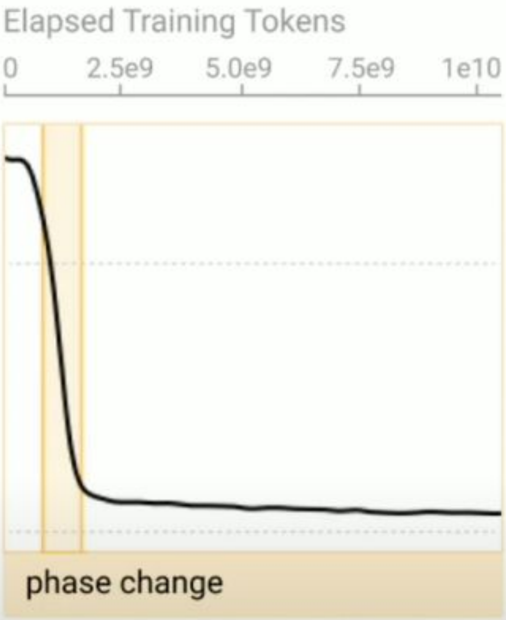
ICL Score

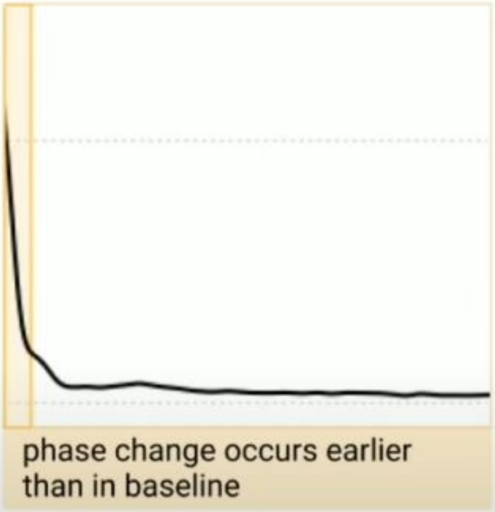
In-context learning score: the loss of the 500th token in the context minus the average loss of the 50th token in the context, averaged over dataset examples

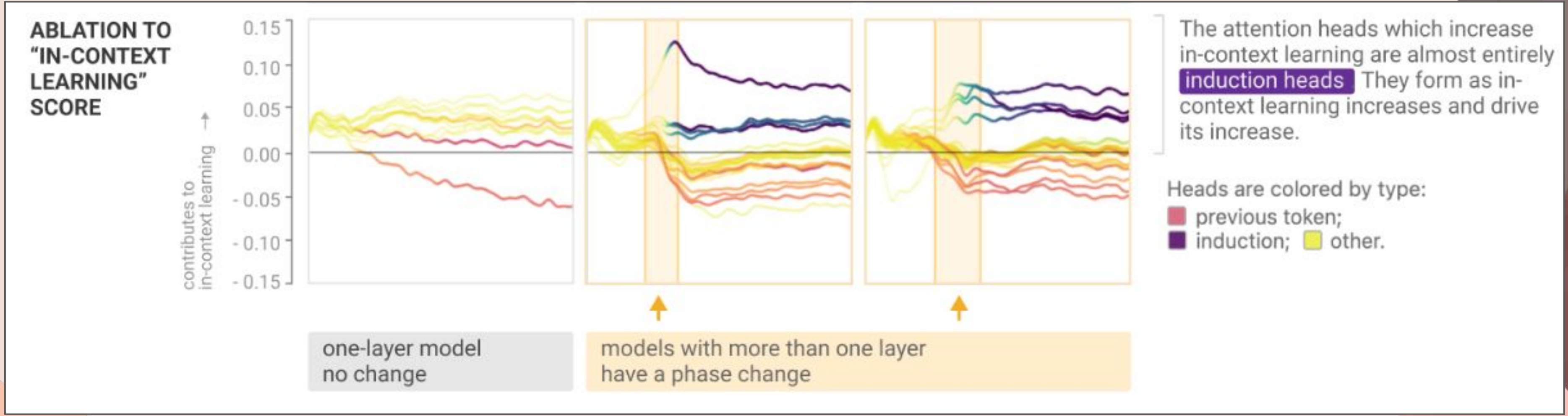



ONE LAYER (ATTENTION-ONLY)
TWO LAYER (ATTENTION-ONLY)
ONE LAYER (ATTENTION-ONLY)
TWO LAYER (ATTENTION-ONLY)
Elapsed Training Tokens
0
2.5e9
5.0e9
7.5e9
1e10
0.1
0.2
0.3
0.4
0.5
1.0
0.8
0.6
0.4
0.2
0.0

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

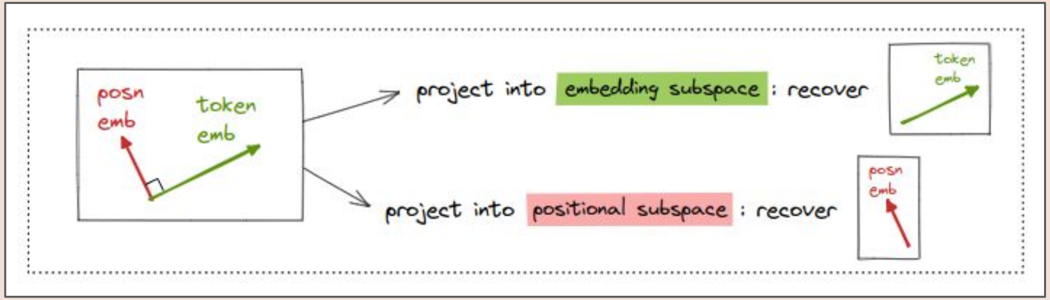


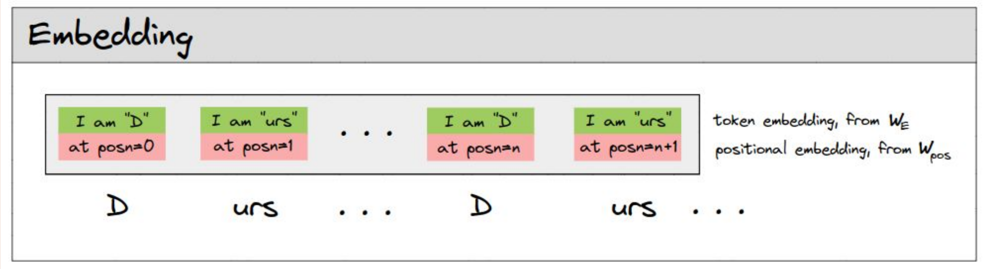




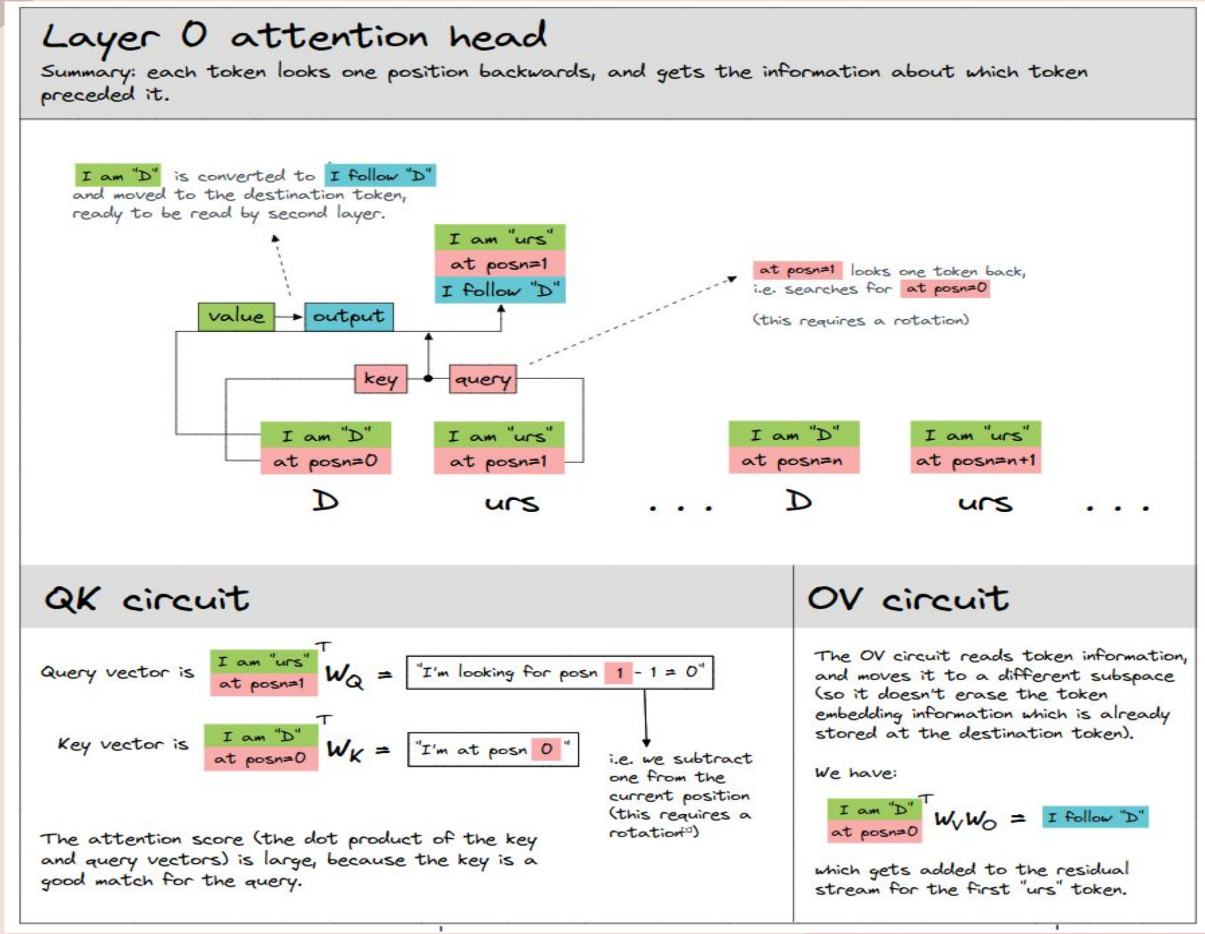
ABLATION TO "IN-CONTEXT LEARNING" SCORE
0.15
0.10
0.05
0.00
- 0.05
- 0.10
- 0.15
contributes to in-context learning
one-layer model no change
models with more than one layer have a phase change
The attention heads which increase in-context learning are almost entirely induction heads They form as in-context learning increases and drive its increase.
Heads are colored by type:
previous token;
induction;
other.

Subspaces in the Residual Stream & Embedding





Layer 0 Attention Head



Layer 0 Attention Head Mathematically

Mathematically, the attention scores are:

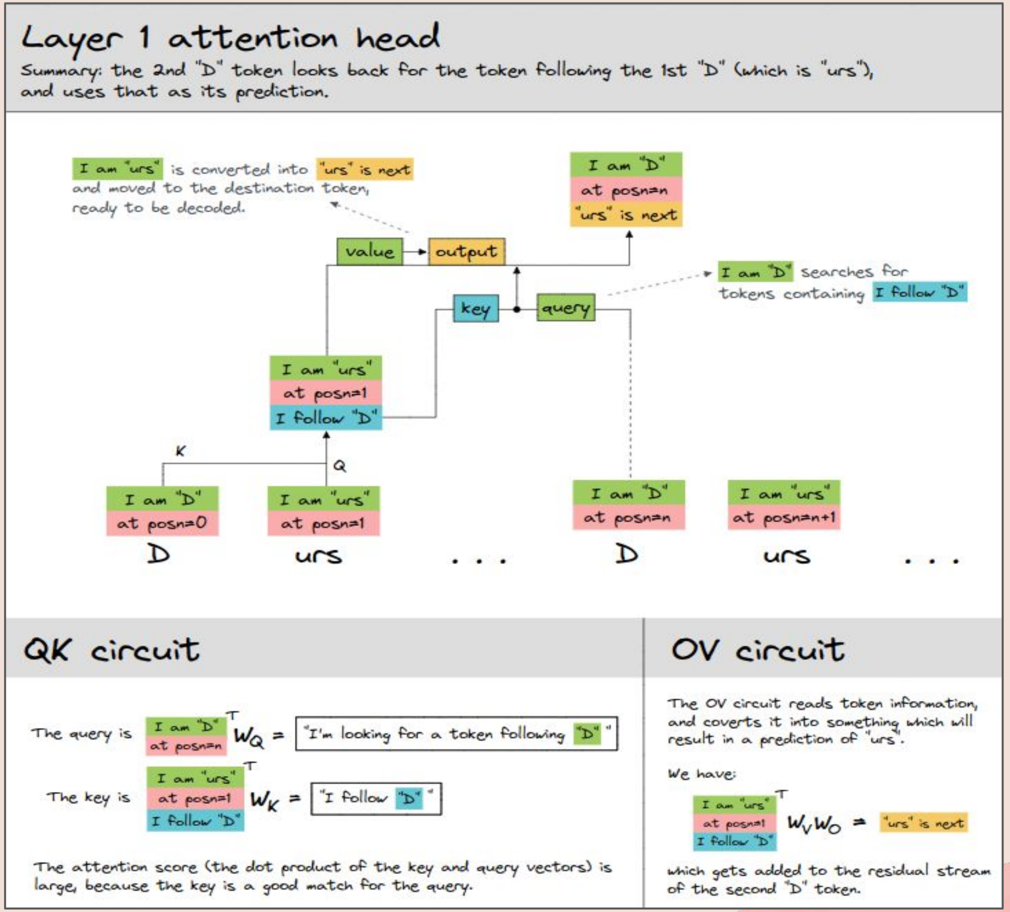
$$\boxed{\begin{array}{c}\text{I am "X"}\\ \text{at posn}=i\end{array}}^T W_Q W_K^T \boxed{\begin{array}{c}\text{I am "Y"}\\ \text{at posn}=j\end{array}} \approx \boxed{\text{at posn}=i}^T W_Q W_K^T \boxed{\text{at posn}=j} = \begin{cases}\text{large} & \text{if } j = i - 1\\ \text{small} & \text{if else}\end{cases}$$

How can we prove this?

Define the "full QK circuit" $W_{pos} W_{QK} W_{pos}^T$

The (i, j) th element of this matrix is $\boxed{\text{at posn}=i}^T W_Q W_K^T \boxed{\text{at posn}=j}$, so we can verify this.

Layer 1 Attention Head



Layer 1 Attention Head Mathematically

Mathematically, the attention scores are:

$$\begin{bmatrix}\text{I am "X"}\\ \text{at posn i}\\ \text{I follow "P"}\end{bmatrix}^T W_Q W_K^T \begin{bmatrix}\text{I am "Y"}\\ \text{at posn j}\\ \text{I follow "Q"}\end{bmatrix} \approx \boxed{\text{I am "X"}}^T W_Q W_K^T \boxed{\text{I follow "Q"}}$$

$$= \underbrace{\left(\boxed{\text{I am "X"}}^T W_Q\right)}_{\text{"I am looking for something which follows "X""}} \cdot \underbrace{\left(\boxed{\text{I follow "Q"}}^T W_K\right)}_{\text{"I follow "Q""}} = \begin{cases}\text{large} & \text{if "X" = "Q"}\\ \text{small} & \text{if else}\end{cases}$$

How can we prove this?

Define the "full K-composition circuit" $W_E W_{QK}^1 (W_{OV}^0)^T W_E^T$

Superscript 0 means the head in layer 0, and 1 to means the head in layer 1.

If X and Q are tokens, then the (X, Q) th element of this matrix is:

$$\boxed{\text{"X"}}^T W_E W_{QK}^1 (W_{OV}^0)^T W_E^T \boxed{\text{"Q"}} = \boxed{\text{I am "X"}}^T W_{QK}^1 (W_{OV}^0)^T \boxed{\text{I am "Q"}}$$

$$= \boxed{\text{I am "X"}}^T W_{QK}^1 \boxed{\text{I follow "Q"}}$$

So we just need to verify that this is $\begin{cases}\text{large} & \text{if "X" = "Q"}\\ \text{small} & \text{if else}\end{cases}$

quAIdditch

In the context of transformer models, which specific circuit is responsible for writing information into the residual stream that helps subsequent layers make predictions about token sequences?

OV (Output-Value) circuit