Progress Measures for Grokking via Mechanistic Interpretability

Charlotte Peale & Sílvia Casacuberta

REFORM reading group

February 26, 2026

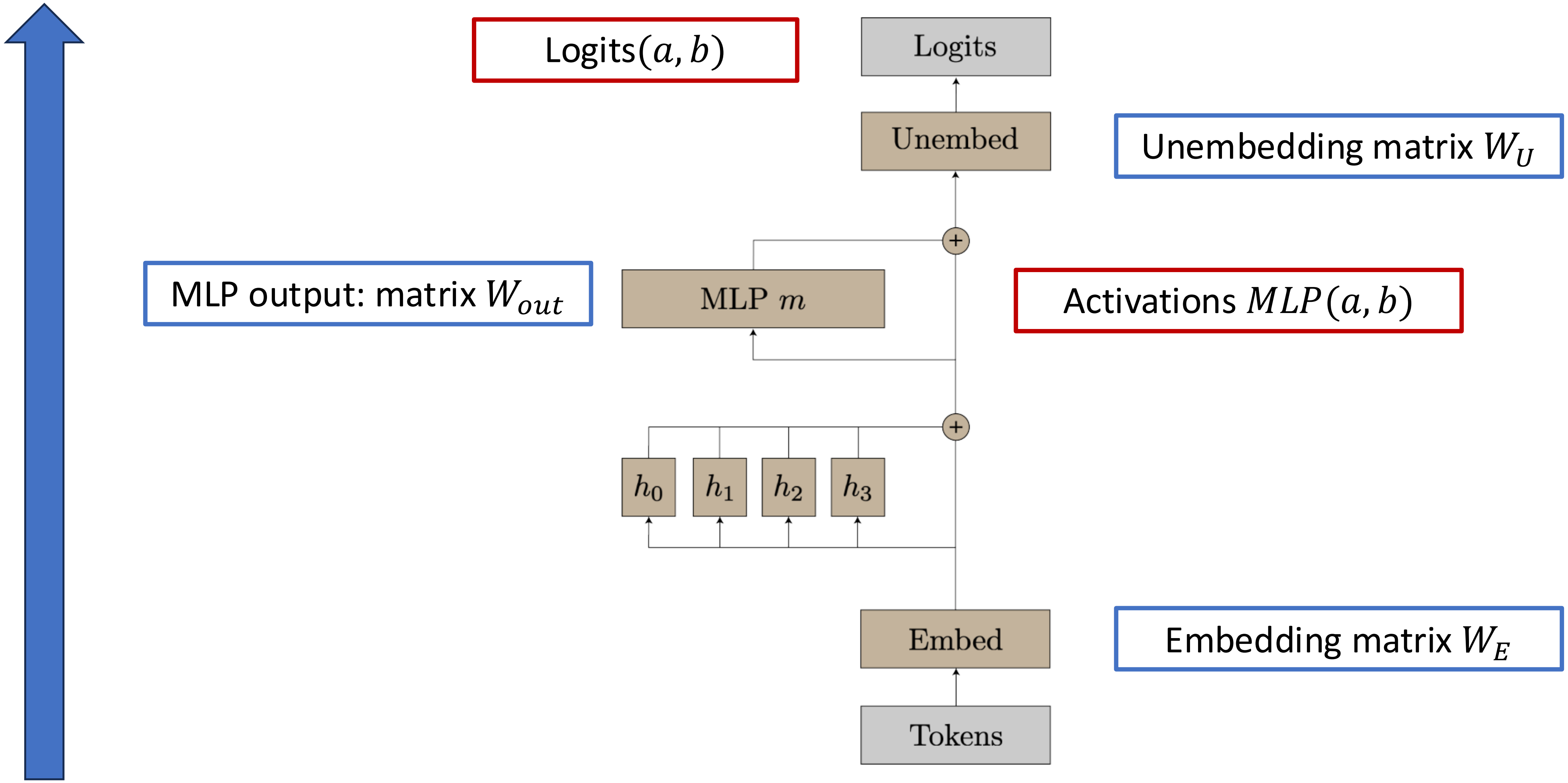
Emergence behavior

- NNs often exhibit emergent behavior, arising from scaling…
  - Amount of parameters
  - Training data
  - Training steps
- Question: How can we understand emergence?
  - Find <u>continuous</u> progress measures
- In this paper: use mechanistic interpretability to find progress measures
  - I.e., reverse-engineer learned behaviors into their individual components, by identifying the circuits within a model that implement a behavior
  - Use mech. interp. to discover progress measures empirically

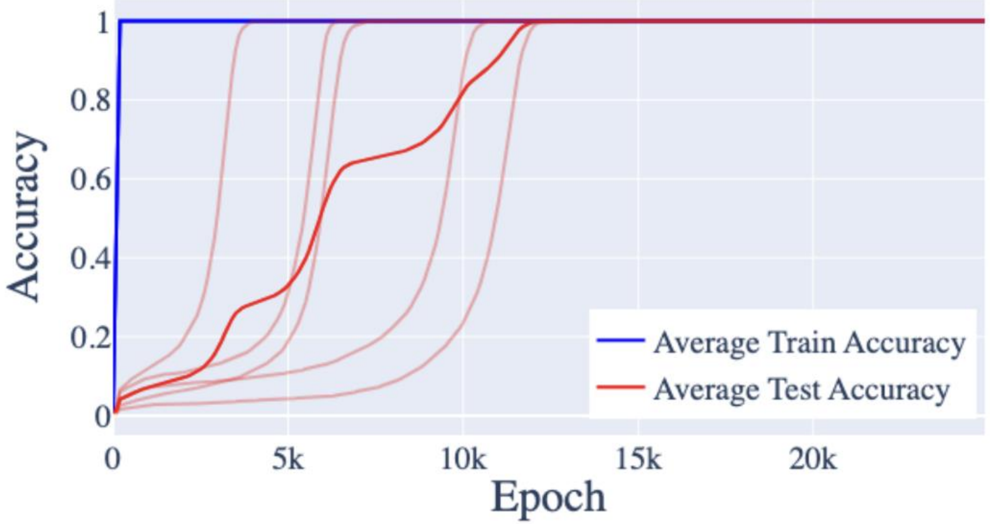
Abrupt emergence?

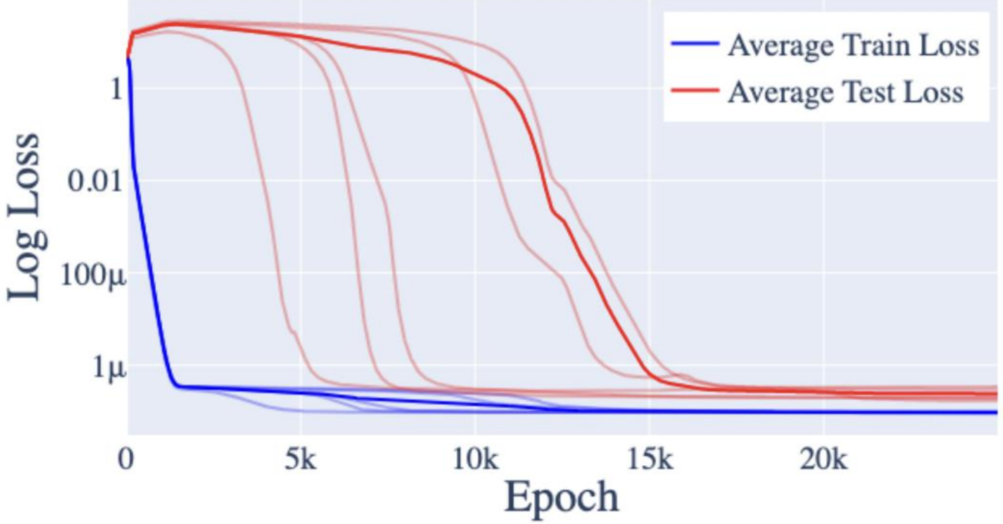
- Usually looks discontinuous
  - Abrupt phase transitions
  - Task ability
  - **Grokking**
    - Models first overfit, then abruptly transition to a generalizing solution after a large number of training steps
- Cross-entropy loss does not explain the phase changes
- Case study: modular addition
  - Input: $a, b \in \{0, \ldots, P-1\}$ for some prime P
  - Output: $a + b \; mod \; P$

Architecture: a one-layer transformer

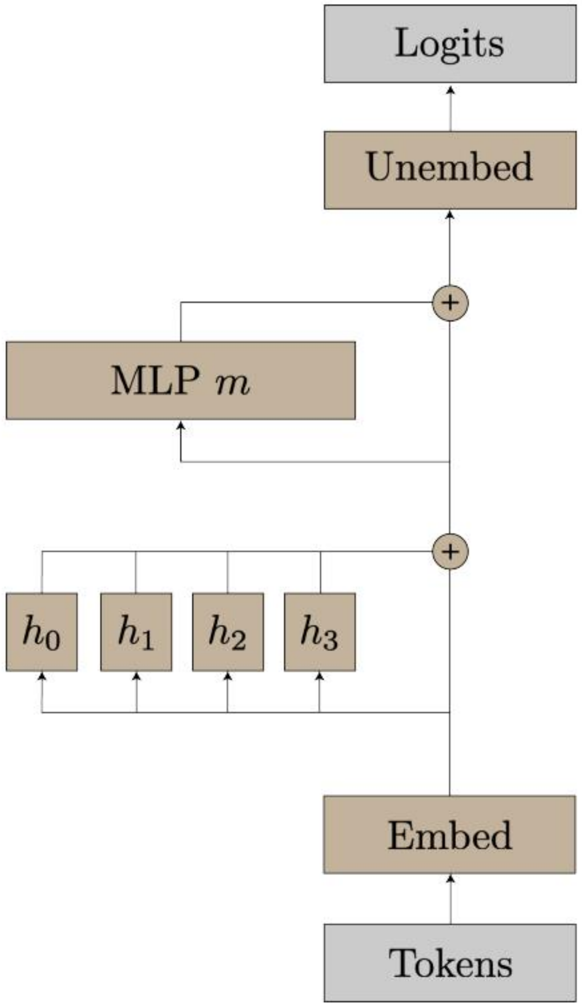


1-layer transformers on modular addition exhibit grokking





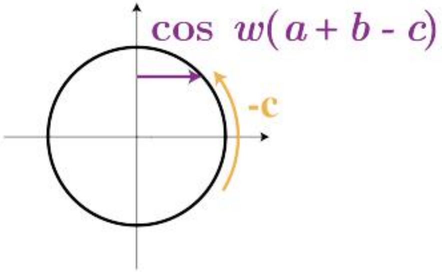
Use $P = 113$; grokking also occurs for other architectures and prime moduli, but does not occur without regularization



Reads off the logits for each $c \in \{0, 1, \dots, P-1\}$ by rotating by c to get $\cos(w(a+b-c))$ → maximized when $a + b = c \bmod P$!

Computes logits using further trig identities:

$$\mathrm{Logit}(c) \propto \cos(w(a+b-c))$$
$$= \cos(w(a+b))\cos(wc) + \sin(w(a+b))\sin(wc)$$

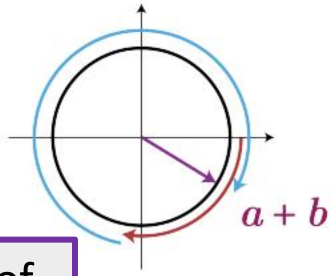


Calculates sine and cosine of $a+b$ using trig identities:

$$\sin(w(a+b)) = \sin(wa)\cos(wb) + \cos(wa)\sin(wb)$$
$$\cos(w(a+b)) = \cos(wa)\cos(wb) - \sin(wa)\sin(wb)$$

In the attention and MLP layers

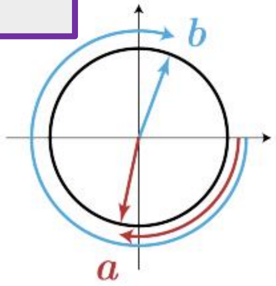
Representation of $a + b \bmod P$



Translates one-hot a, b to Fourier basis:

$$a \to \sin(wa), \cos(wa)$$
$$b \to \sin(wb), \cos(wb)$$

Project a, b using w_k

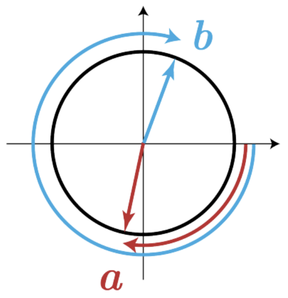


$w_k a, w_k b$ for various frequencies $w_k = \frac{2\,k\,\pi}{P}, k \in \mathbb{N}$

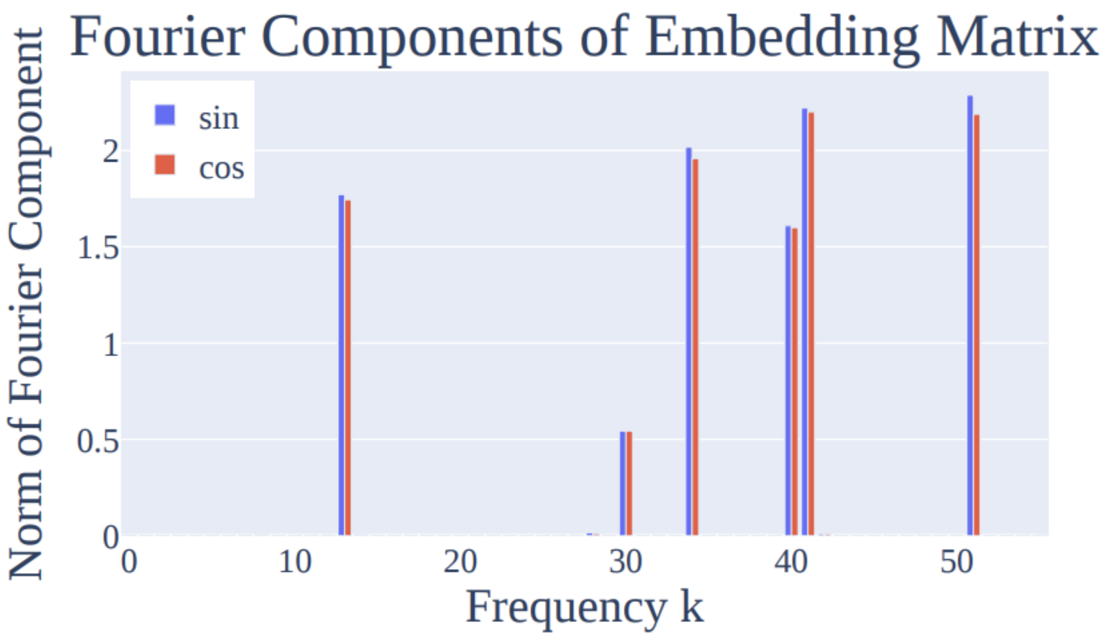
Evidence

Translates one-hot a, b to Fourier basis:
$a \to \sin(wa), \cos(wa)$
$b \to \sin(wb), \cos(wb)$

Discovering which w_k it uses



1. Periodicity in the embeddings.
   - Apply a Fourier transform along the input dimension of W_E
   - Compute the ℓ_2 norm along the other dimension

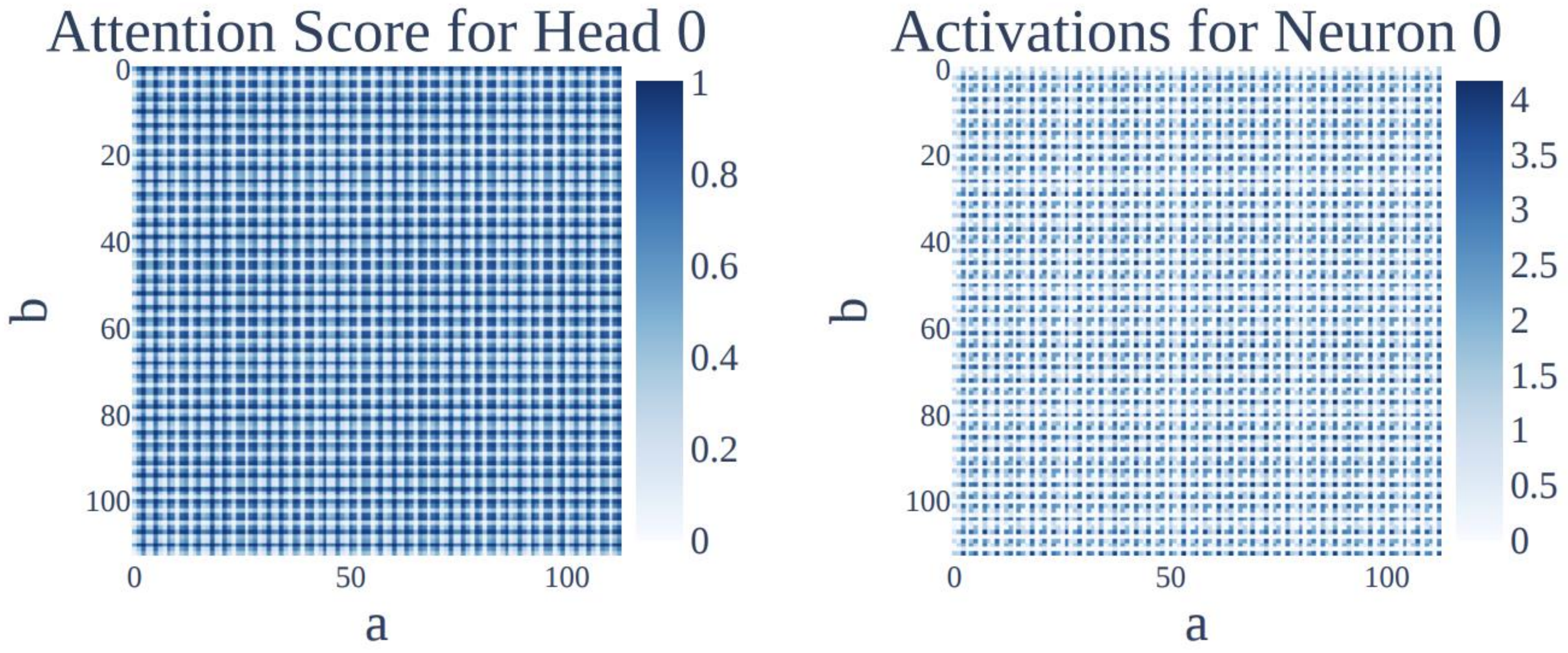


W_E is very sparse in the Fourier basis: only 6 frequencies have non-negligible norm → key frequencies

Evidence

2. Periodicity in the attention heads and MLP neuron activations.

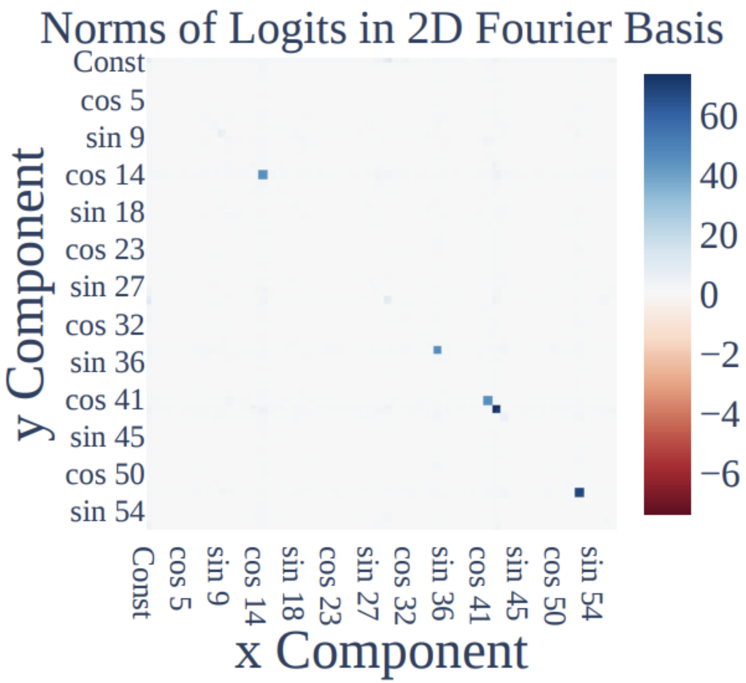
For $k = 35$ and $k = 42$



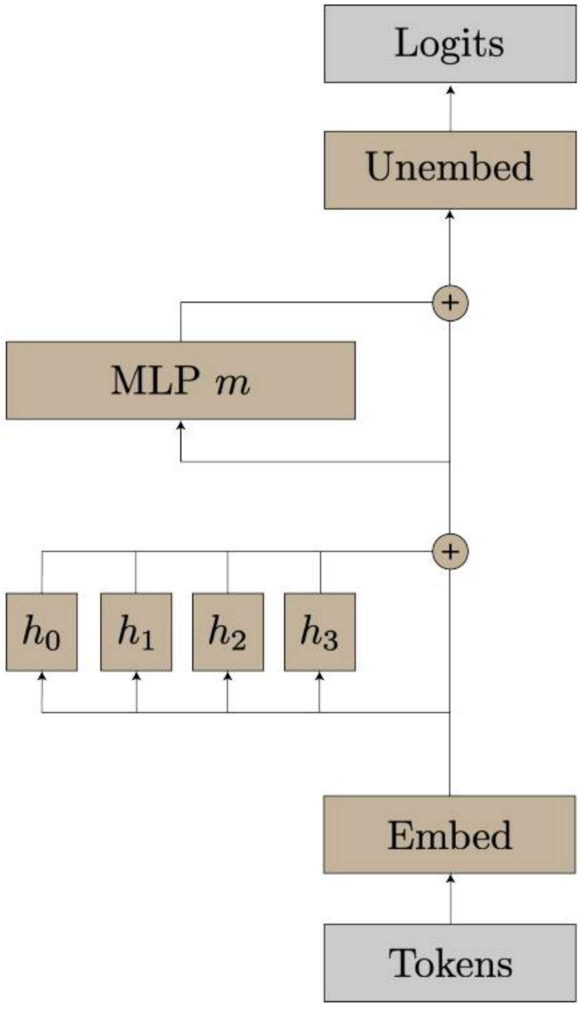
Evidence

3. Periodicity in logits.
   - 2D Fourier basis over the inputs, then take ℓ_2 norm over the output dim.
   - Only 20 significant components, corresponding to the products of sin and cos for the 5 key frequencies

4 possible sin/cos products → it correctly uses the same w_k

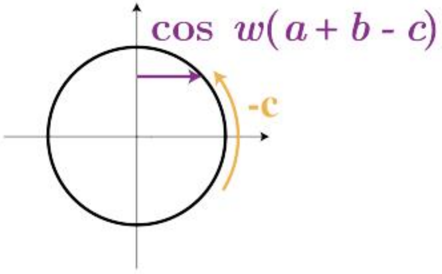


Reads off the logits for each $c \in \{0, 1, \ldots, P-1\}$ by rotating by c to get $\cos(w(a+b-c))$ → maximized when $a+b = c \bmod P$!



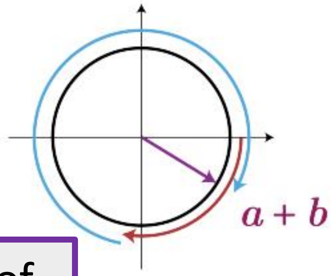
Computes logits using further trig identities:

$$\mathrm{Logit}(c) \propto \cos(w(a+b-c))$$
$$= \cos(w(a+b))\cos(wc) + \sin(w(a+b))\sin(wc)$$



Calculates sine and cosine of $a+b$ using trig identities:

$$\sin(w(a+b)) = \sin(wa)\cos(wb) + \cos(wa)\sin(wb)$$
$$\cos(w(a+b)) = \cos(wa)\cos(wb) - \sin(wa)\sin(wb)$$



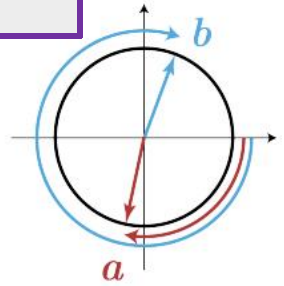
In the attention and MLP layers

Representation of $a+b \bmod P$

Translates one-hot a, b to Fourier basis:

$$a \to \sin(wa), \cos(wa)$$
$$b \to \sin(wb), \cos(wb)$$

Project a, b using w_k



$w_k a, w_k b$ for various frequencies

$$w_k = \frac{2k\pi}{P}, k \in \mathbb{N}$$

Trigonometric identity

- Matrix W_L: matrix mapping MLP activations to logits
  - It is approximately rank 10 (with the 5 key frequencies)

$$W_L = \sum_{k\in\{14,35,41,42,52\}} \cos(w_k)\, u_k^T + \sin(w_k)\, v_k^T$$

- The model implements the logits for a, b as:

$$\text{Logits}(a,b) = W_L \text{MLP}(a,b) \approx \sum_k \cos(w_k)\, u_k^T \text{MLP}(a,b) + \sin(w_k)\, v_k^T \text{MLP}(a,b)$$

Multiples of $\cos(w_k(a+b))$ and $\sin(w_k(a+b))$

Trigonometric identity

Similarly, logits are well approximated by a weighted sum of $\cos(w(a+b-c))$'s

W_L Component	Fourier components of $u_k^T \mathrm{MLP}(a,b)$ or $v_k^T \mathrm{MLP}(a,b)$	FVE
$\cos(w_{14}c)$	$44.6\cos(w_{14}a)\cos(w_{14}b) - 43.6\sin(w_{14}a)\sin(w_{14}b) \approx 44.1\cos(w_{14}(a+b))$	93.2%
$\sin(w_{14}c)$	$44.1\sin(w_{14}a)\cos(w_{14}b) + 44.1\cos(w_{14}a)\sin(w_{14}b) \approx 44.1\sin(w_{14}(a+b))$	93.5%
$\cos(w_{35}c)$	$40.7\cos(w_{35}a)\cos(w_{35}b) - 43.6\sin(w_{35}a)\sin(w_{35}b) \approx 42.2\cos(w_{35}(a+b))$	96.8%
$\sin(w_{35}c)$	$41.8\sin(w_{35}a)\cos(w_{35}b) + 41.8\cos(w_{35}a)\sin(w_{35}b) \approx 41.8\sin(w_{35}(a+b))$	96.5%
$\cos(w_{41}c)$	$44.8\cos(w_{41}a)\cos(w_{41}b) - 44.8\sin(w_{41}a)\sin(w_{41}b) \approx 44.8\cos(w_{41}(a+b))$	97.0%
$\sin(w_{41}c)$	$44.5\sin(w_{41}a)\cos(w_{41}b) + 44.5\cos(w_{41}a)\sin(w_{41}b) \approx 44.5\sin(w_{41}(a+b))$	97.0%
$\cos(w_{42}c)$	$64.6\cos(w_{42}a)\cos(w_{42}b) - 68.5\sin(w_{42}a)\sin(w_{42}b) \approx 66.6\cos(w_{42}(a+b))$	96.4%
$\sin(w_{42}c)$	$67.8\sin(w_{42}a)\cos(w_{42}b) + 67.8\cos(w_{42}a)\sin(w_{42}b) \approx 67.8\sin(w_{42}(a+b))$	96.4%
$\cos(w_{52}c)$	$60.5\cos(w_{52}a)\cos(w_{52}b) - 65.5\sin(w_{52}a)\sin(w_{52}b) \approx 63.0\cos(w_{52}(a+b))$	97.4%
$\sin(w_{52}c)$	$64.5\sin(w_{52}a)\cos(w_{52}b) + 64.5\cos(w_{52}a)\sin(w_{52}b) \approx 64.5\sin(w_{52}(a+b))$	98.2%

Table 1: For each of the directions u_k or v_k (corresponding to the $\cos(w_k)$ and $\sin(w_k)$ components respectively) in the unembedding matrix, we take the dot product of the MLP activations with that direction, then perform a Fourier transform (middle column; only two largest coefficients shown). We then compute the fraction of variance explained (FVE) if we replace the projection with a single term proportional to $\cos(w_k(a+b))$ or $\sin(w_k(a+b))$, and find that it is consistently close to 1.

Final step

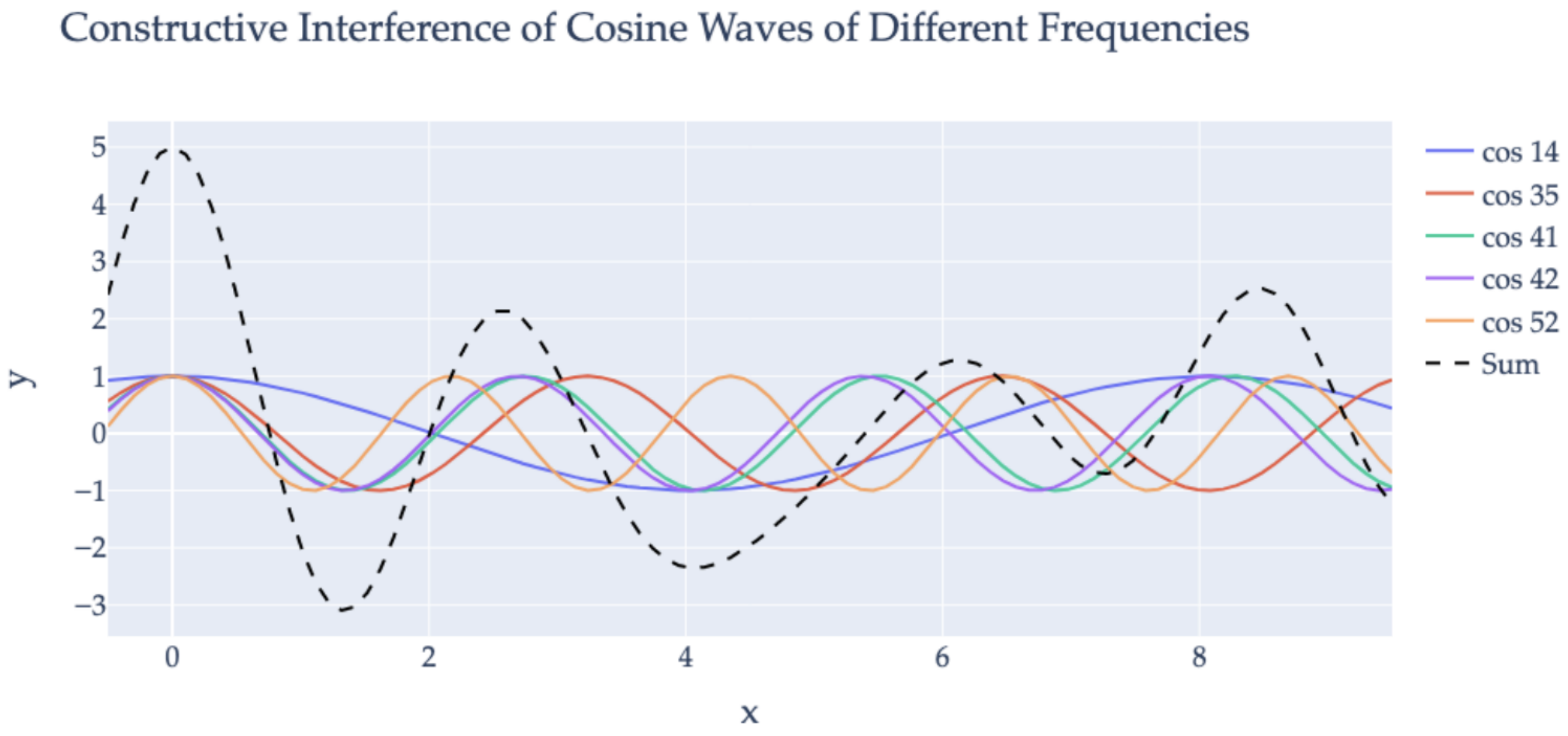
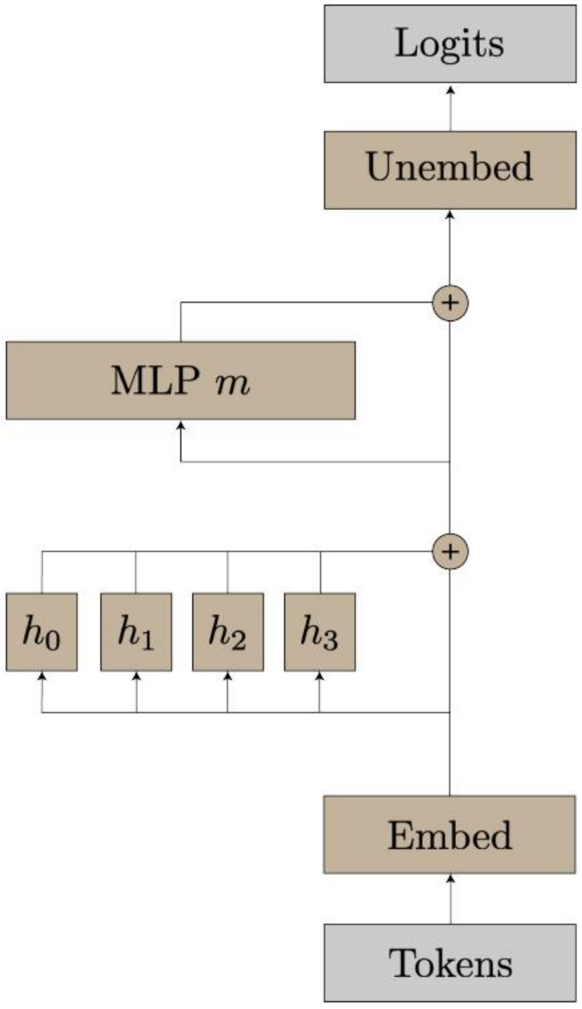
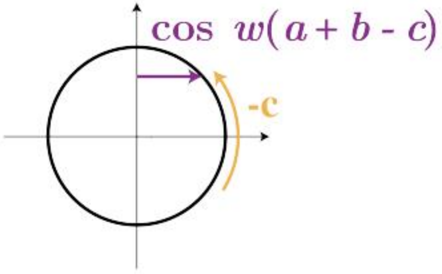


Figure 8: As discussed in Appendix B, while for every $k \in [0, \ldots P-1]$, $\cos\left(\frac{2k\pi}{P}x\right)$ achieves its maximum value (1) at $x = 0 \mod 113$, it still has additional peaks at different values that are close to the maximum value. However, by adding together cosine waves of the 5 keyfrequencies, the model constructs a periodic function where the value at $x = 0 \mod 113$ is significantly larger than its value anywhere else.

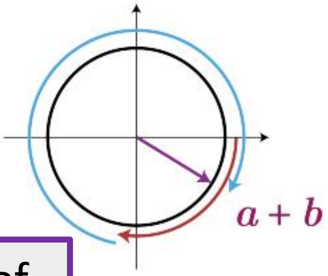


Reads off the logits for each $c \in \{0, 1, \ldots, P-1\}$ by rotating by c to get $\cos(w(a+b-c))$ → maximized when $a + b = c \bmod P$!



Computes logits using further trig identities:

$$\mathrm{Logit}(c) \propto \cos(w(a+b-c))$$
$$= \cos(w(a+b))\cos(wc) + \sin(w(a+b))\sin(wc)$$

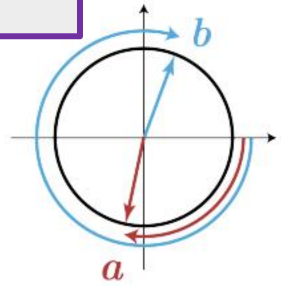


Calculates sine and cosine of $a + b$ using trig identities:

$$\sin(w(a+b)) = \sin(wa)\cos(wb) + \cos(wa)\sin(wb)$$
$$\cos(w(a+b)) = \cos(wa)\cos(wb) - \sin(wa)\sin(wb)$$

In the attention and MLP layers

Representation of $a + b \bmod P$



Translates one-hot a, b to Fourier basis:

$$a \to \sin(wa), \cos(wa)$$
$$b \to \sin(wb), \cos(wb)$$

Project a, b using w_k

$w_k a, w_k b$ for various frequencies

$$w_k = \frac{2\,k\,\pi}{P}, k \in \mathbb{N}$$

Progress measures

- Metrics that can be computed during training that track the progress, *including during phase transitions*
- Restricted loss: measure how well intermediate versions of the model can do with only the 5 key frequencies
  - Measure the loss of the ablated network
- Excluded loss: remove *only* the key frequencies from the logits but keep the rest
  - Measure this on the training data
  - Memorizing solution should be spread out in the Fourier domain → ablating a bit doesn't hurt much, but it should hurt the generalizing solution

1. Memorization

- Decline of both excluded and train loss; unused w_k frequencies

2. Circuit formation

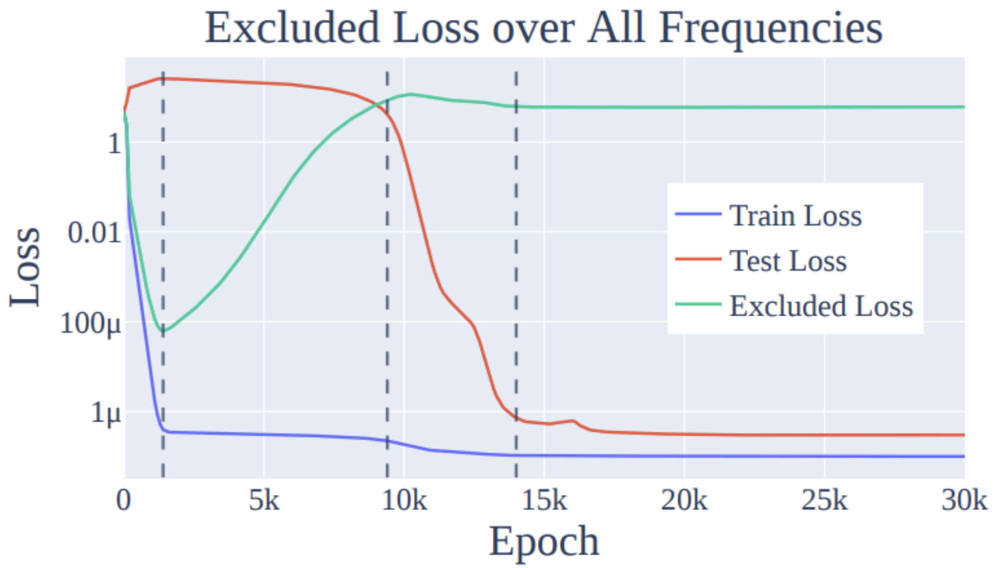
- Excluded loss rises, restricted loss starts to fall
- Smooth transition from memorizing to Fourier mult. algorithm
- Occurs *well before* the grokking occurs!

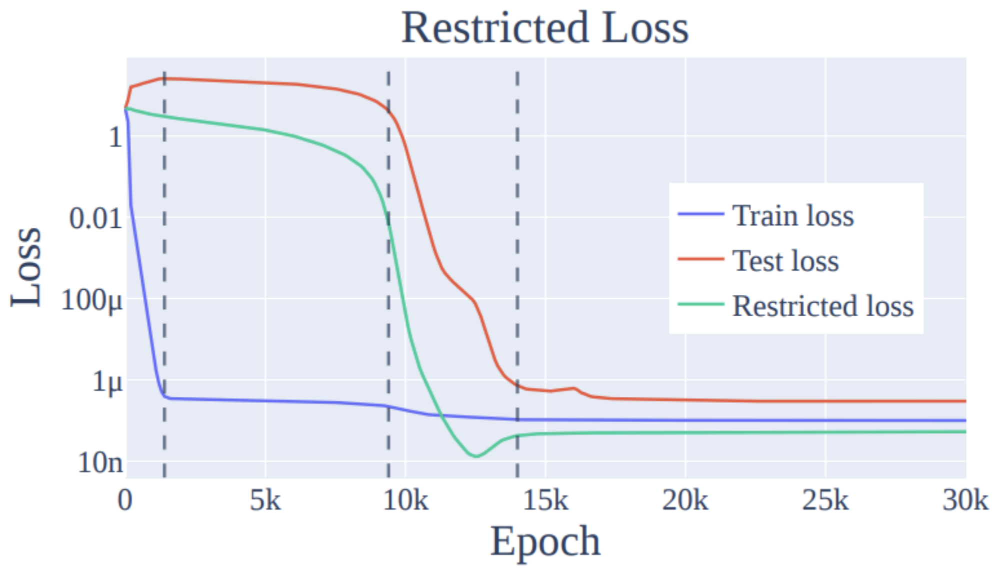
3. Cleanup

- Excluded loss plateaus, restricted loss continues to drop, test loss suddenly drops
- Completed Fourier circuit

Restricted: **only** the key frequencies

Excluded: **without** the key frequencies





Grokking, rather than being a sudden shift, arises from the gradual amplification of structured mechanisms encoded in the weights, followed by the later removal of memorizing components