The Discrete-Log Clock: How a Transformer Learns Modular Multiplication

Danh Nguyen · Stanford University

Mechanistic Interpretability Workshop @ ICML 2026 · Seoul

The Problem

In (Nanda et al., 2023)[1], a specialized transformer is trained for the task of predicting the result of a + b mod 113 until grokked. The mechanism was and later dubbed the Clock mechanism by (Zhong et al., 2023)[2]. It uses sparse Fourier representations and trig identities to compose rotations on a circle.

<u>A model specialized for a*b mod p also exhibits grokking</u> when trained sufficiently long. However, the Clock mechanism cannot be applied directly as the DFT results in a dense spectrum. The interpretation breaks down.

Our work shows <u>that the transformer reduces multiplication to addition in the discrete-log space,</u> then uses Clock mechanism. We call this flow the Discrete-Log Clock.

The Identity

Analogous to log(a*b) = log(a) + log(b) in ℝ, there is an isomorphism from the multiplicative group (ℤ/pℤ)* to ℤ/(p-1) ℤ. Given g is a primitive root of prime p, the mapping is as follows. With p=113, g=3, we have this expression:

$$\log_g : (\mathbb{Z}/p\mathbb{Z})^* \xrightarrow{\cong} \mathbb{Z}/(p-1)\mathbb{Z} \implies a \cdot b \equiv 3^{(\alpha+\beta) \bmod 112} \pmod{113}$$

Method

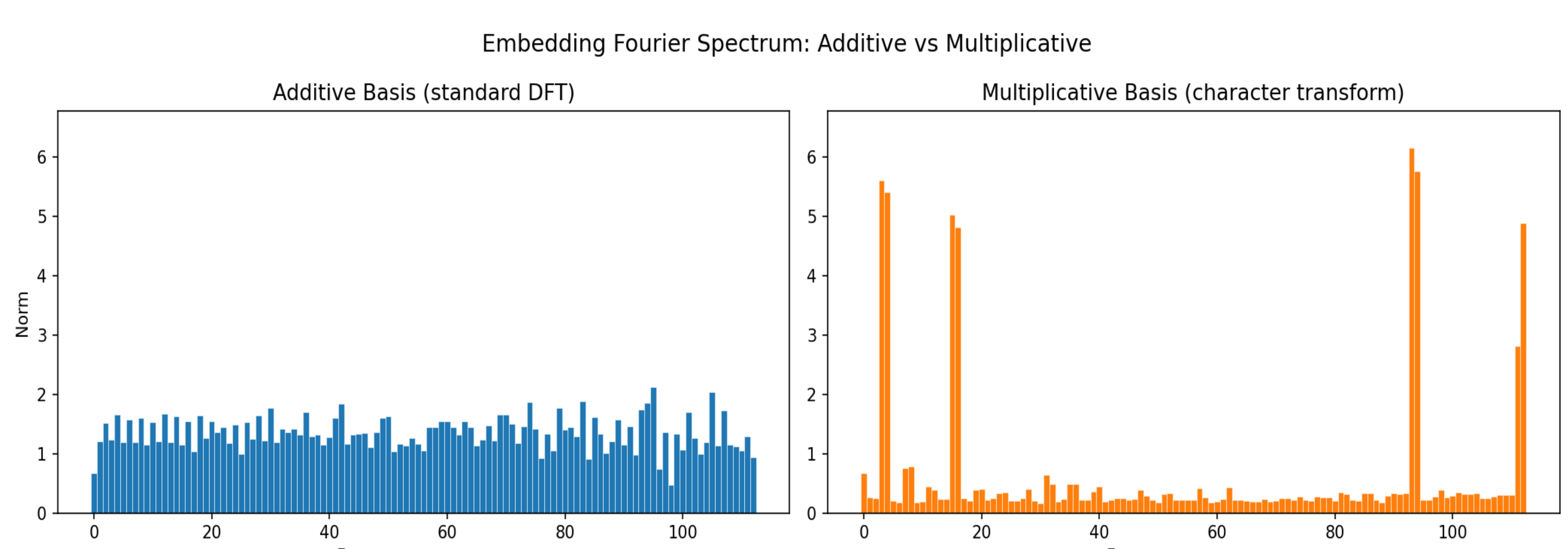
In the Clock mechanism, each entry of the embedding vector of token a is W_E[a] = [sin(ωka), cos(...] where frequency k is an integer ∈ {1,...,p-1}. This interpretation is coming from DFT the other axis of the W_E matrix (column-wise). Clock shows that only some k, called key frequencies, contains significantly more energy than the others. (no. of key << p). The spectrum is sparse.

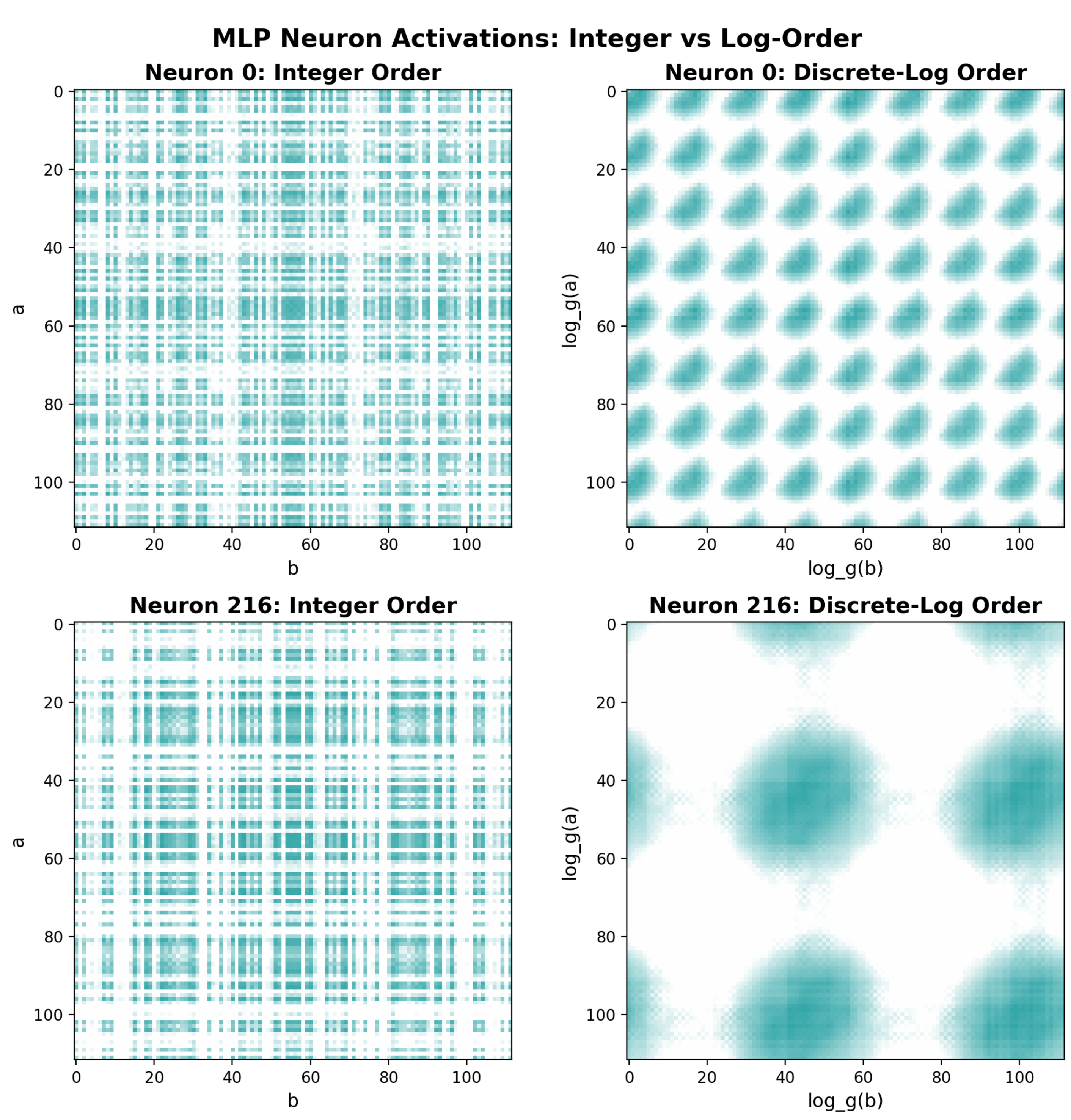
When one tries to apply the same technique directly on the multiplication model, they will find the spectrum is dense, with no key frequency. However, with the identity we have shown, instead of embedding a as W_E[a], we embed it as W_E[log_g(a)].

Setup

- 1-layer transformer, d_model=128, 4 heads
- MLP: 512 neurons, ReLU, no LayerNorm
- Task: a·b mod 113, all a,b ∈ {1,...,112}
- 30% train, AdamW (wd=1.0), 40K epochs
- Groks at epoch ~10K → 100% test accuracy

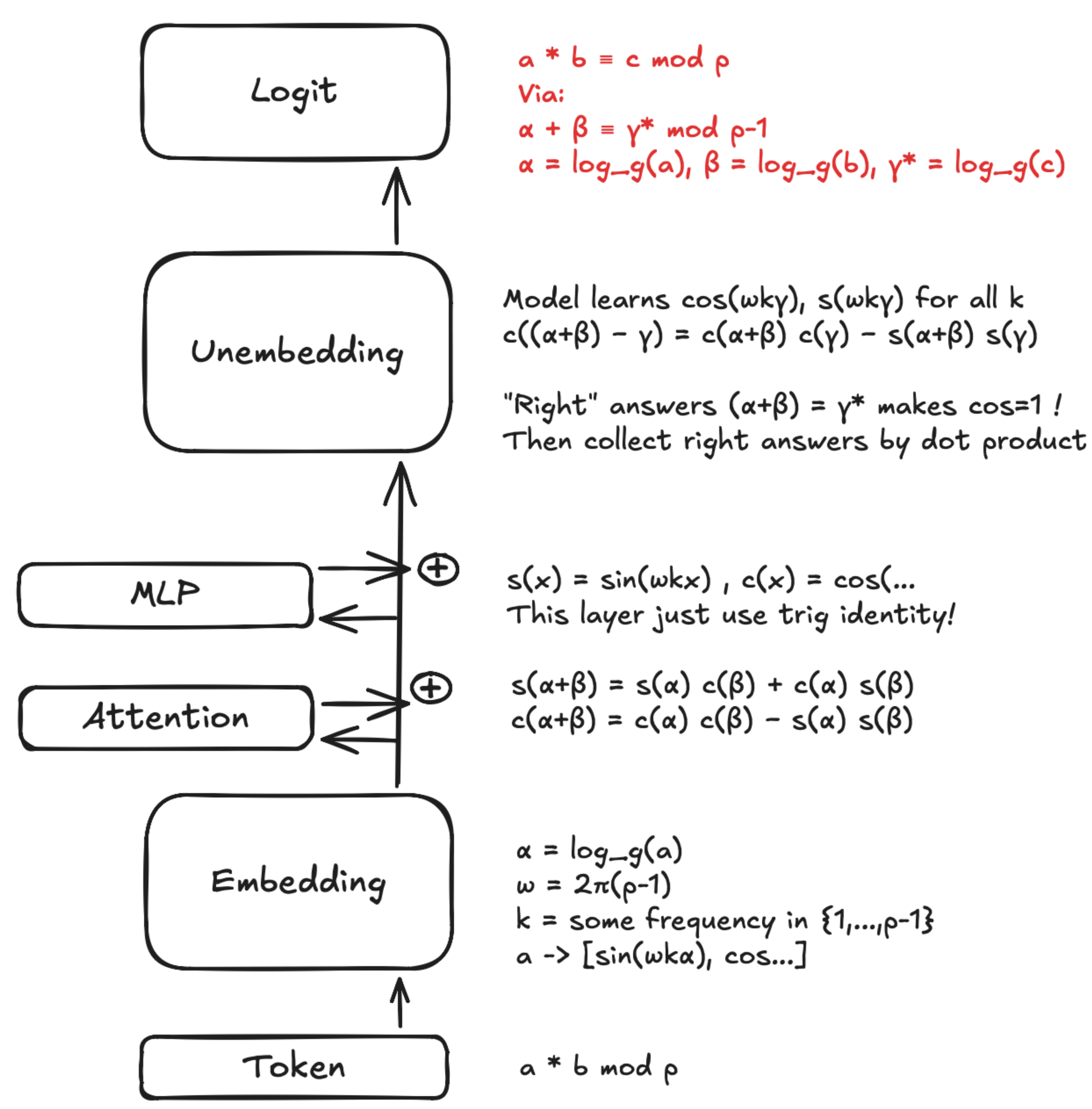
Results





After the transformation, the Fourier spectrum becomes distinctively sparse. The following plots explicitly have the same y-axis scaling for comparison. Quantitatively, the sparsity metric - Gini coefficient goes from 0.071 (left) to 0.579 (right).

Interpretation: the reverse-engineered flow



Conclusion & Future Work

We have shown that a transformer trained on modular multiplication learns the Discrete Log Clock by reducing multiplication to addition in discrete log space.

We have also achieved similar grokking results on modular exponentiation (a^b mod p). Preliminary results suggest that the model implements multiple number theoretic concepts to achieve generalization. We analyze the exponentiation models using the techniques developed from this paper: Using the right transformations before the Fourier spectral analysis.

1 N. Nanda, L. Chan, T. Lieberum, J. Smith, and J. Steinhardt, "Progress Measures for Grokking via Mechanistic Interpretability", ICLR, 2023.
2 Z. Zhong, Z. Liu, M. Tegmark, and J. Andreas, "The Clock and the Pizza: Two Stories in Mechanistic Explanation of Neural Networks," NeurIPS, 2023.