Introduction. This project partially reproduces Nanda *et al.* (2023). A one-layer transformer is trained on modular addition data, and *grokking* is observed.

Implementation Details. Following the original reference, a single-layer transformer is implemented, with token embeddings of dimension $d = 128$, learned positional embeddings, 4 attention heads, and an MLP with hidden dimension $n = 512$. No LayerNorm is used, and the unembedding matrix is not tied to the embedding matrix.

We use $P = 113$. Inputs are token sequences "$a\ b\ =$" with $a, b \in \{0, \ldots, P-1\}$ and "=" a special token. The model predicts $c = (a + b) \bmod P$ from the final position. The dataset consists of all P^2 pairs (a, b); the training split is a uniform random 30% subset, and test is the remaining 70%.

Full-batch training with the AdamW optimizer with fixed learning rate $\gamma = 0.001$ and weight decay $\lambda = 1$ is used. We perform 40,000 epochs of training. Test loss and accuracy are evaluated on all held-out pairs (a, b).

Let $W_E \in \mathrm{R}^{d \times P}$ denote the embedding matrix (ignoring the "=" row for simplicity), $W_{\text{out}} \in \mathrm{R}^{d \times n}$ the MLP output matrix, and $W_U \in \mathrm{R}^{P \times d}$ the unembedding matrix. Empirically, the skip connection around the MLP is small, so logits are approximated by

$$\text{Logits}(a, b) \approx W_U W_{\text{out}} \,\text{MLP}(a, b) = W_L \,\text{MLP}(a, b), \quad W_L := W_U W_{\text{out}} \in \mathrm{R}^{P \times n}.$$

Results. Grokking is observed for certain seeds, whereas for others no grokking occurs. Due to time constraints, two successful instances are recorded. The training and test loss and accuracy over 40,000 epochs are shown in Figure 1.

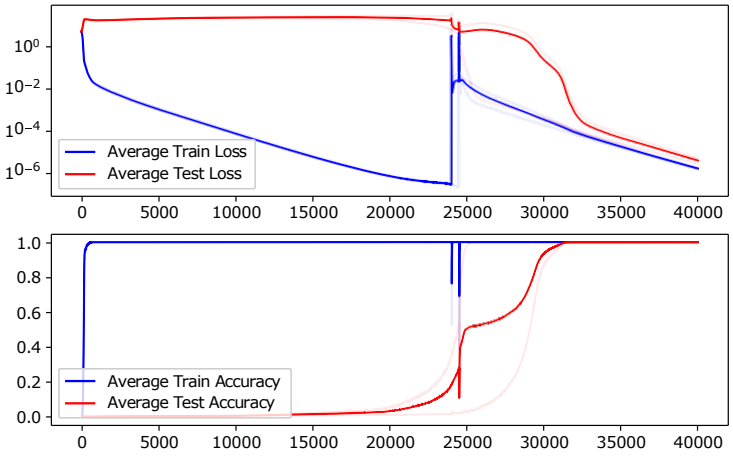

Figure 1: Training and test loss (*top*) and training and test accuracy (*bottom*) over 40,000 epochs. The model initially overfits, leading to increasing test loss, but later learns to generalize.

Unlike the reference, the training loss spikes when grokking first occurs. The reason for this phenomenon is unclear.

Following the reference, a Fourier analysis is performed on the embedding matrix and the neuron-logit map. We first compute the DFT along the token index dimension

$$\hat{W}_E = \text{FFT}(W_E),$$

then compute the ℓ_2 norm across the model dimension,

$$\text{norm}(k) := \|\hat{W}_E[k, :]\|_2.$$

Since the FFT results are symmetric with respect to $\lfloor (P-1)/2 \rfloor$, we analyze $k = 0, \ldots, 56$.

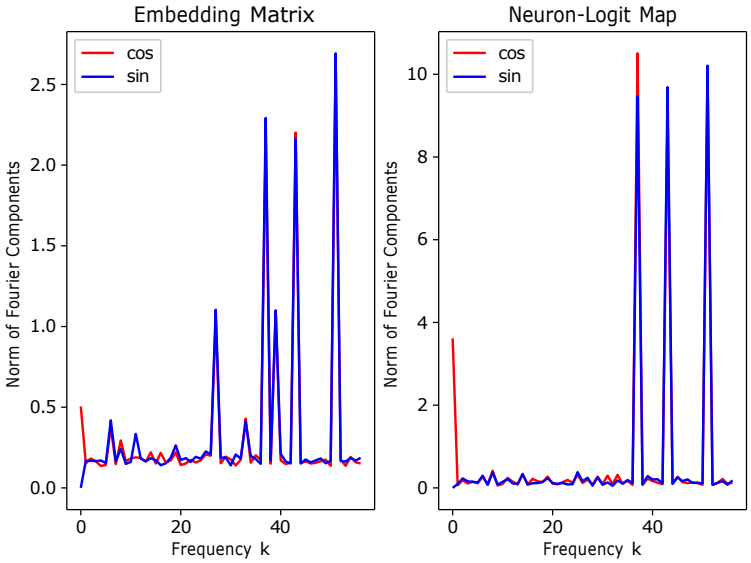

Figure 2: (*Left*) Norms of the Fourier components of the embedding matrix W_E. Six peaks are observed. (*Right*) Norms of the Fourier components of the neuron-logit map W_L. Of these, three "key frequencies" remain, corresponding to $k \in \{37, 43, 51\}$.

To show that these three key frequencies dominate W_L, we examine its low-rank structure using a Scree plot and explained energy.

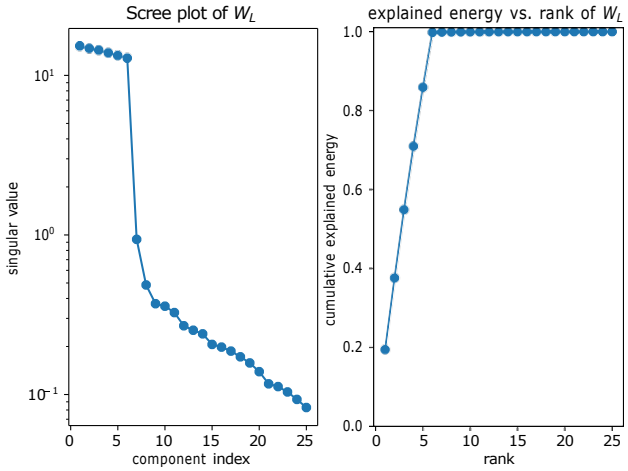

Figure 3: Scree plot (*left*) and explained energy (*right*) of W_L. Six singular values explain most of the energy, indicating approximate rank 6.

We now identify the different phases of grokking using the progress measures proposed in the reference.

We define the *restricted* and *excluded* loss, corresponding to the loss when the neuron-logit map W_L consists of only the key frequencies or all other frequencies, respectively. Following the reference, the restricted loss is evaluated on the full dataset, while the excluded loss is evaluated on the training dataset. Fourier sparsity of W_E and W_L is measured via the Gini coefficient, and the total sum of squared weights is also tracked. The results are shown in Figure 4.

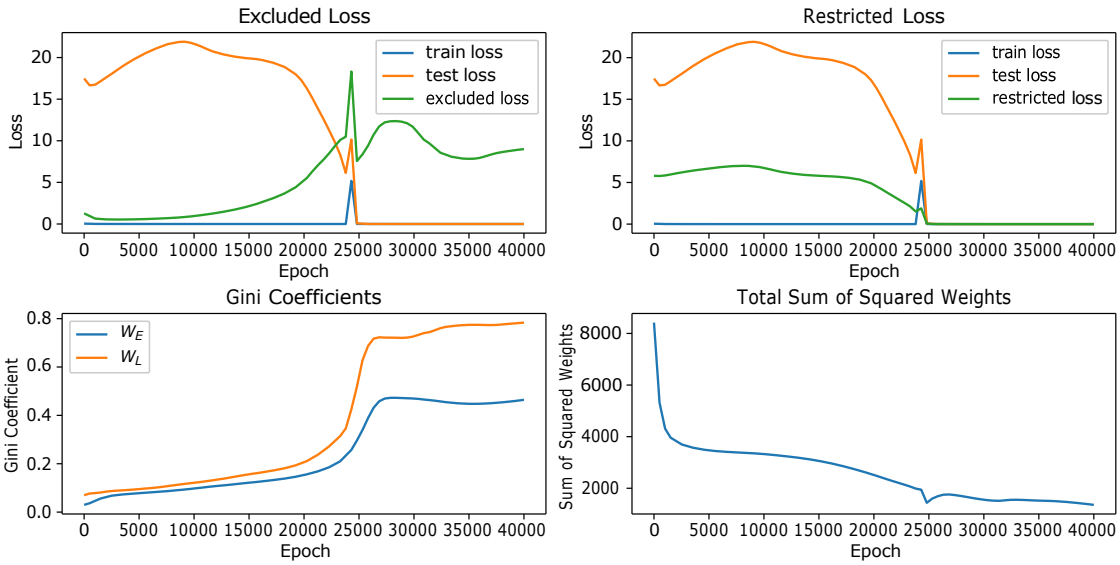

Figure 4: Progress measures during training. (*Top left*) Excluded loss increases during circuit formation, indicating suppression of memorization-based components. (*Top right*) Restricted loss decreases before test loss improves, showing that the Fourier circuit forms prior to generalization. (*Bottom*) At the grokking transition, Fourier sparsity increases sharply and the total squared weight decreases slightly, corresponding to the cleanup phase.

Strong weight decay $\lambda = 1.0$ is essential for grokking. When weight decay is removed, the model continues to fit the training data but fails to reorganize its internal representation. As a result, neither Fourier sparsification nor test generalization occurs, as shown in Figure 5.

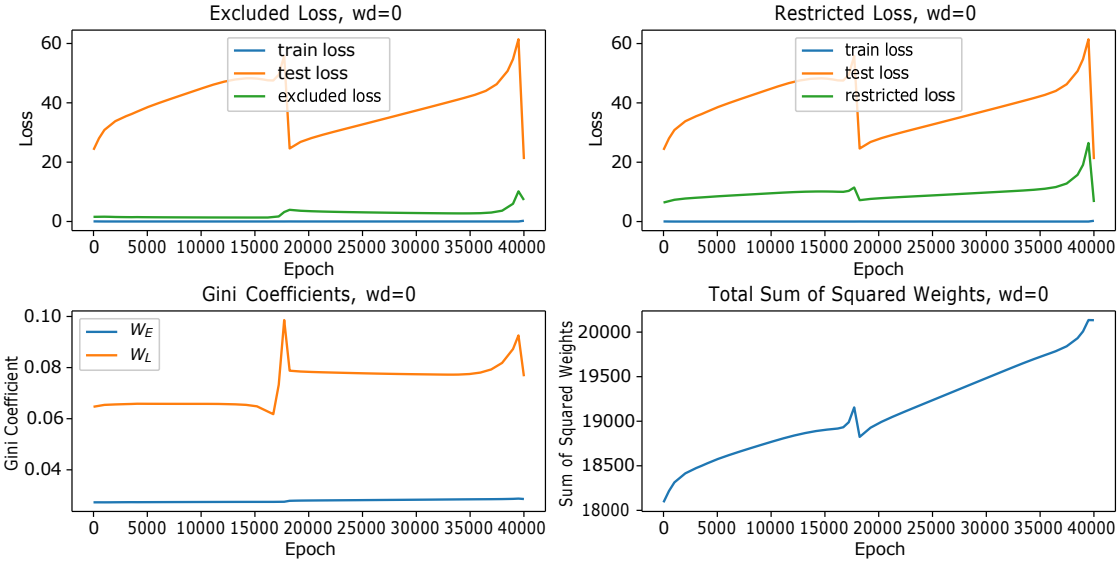

Figure 5: Same plots as Figure 4, but with zero weight decay. Grokking does not occur.

Discussion. The Fourier peaks of the embedding matrix indicate which group characters are available in the internal representation space. In contrast, peaks in the neuron-logit map indicate which group characters directly determine the logits. A peak at frequency k in W_L means that the output logits depend on the character $x_k(a + b)$.

The approximate rank-6 structure of W_L supports this interpretation. Each Fourier frequency corresponds to two real modes (sine and cosine), so three key frequencies produce six dominant singular values.

The reference identifies three phases of grokking. While these phases are less sharply separated here due to implementation and seed differences, similar trends are observed:

- *Memorization:* training loss decreases as the model fits the training data.
- *Circuit formation:* training loss reaches zero while excluded loss rises gradually, indicating slow formation of the Fourier circuit.
- *Cleanup:* restricted and test loss drop to zero while Fourier sparsity increases sharply, indicating a transition to a sparse Fourier solution.

Unlike the reference, the total squared weight does not drop as sharply; the cause is unclear.

Without weight decay, the model has no incentive to increase excluded loss or decrease the total squared weight. Consequently, no generalization occurs, highlighting the role of regularization in grokking.

References. Nanda, N., Chan, L., Lieberum, T., Smith, J., Steinhardt, J. (2023). *Progress Measures for Grokking via Mechanistic Interpretability.* ICLR 2023.