Abrupt Learning in Transformers: A Case Study in Matrix Completion

Pulkit Gopalani[1]

Ekdeep Singh Lubana[2], **Wei Hu**[1]

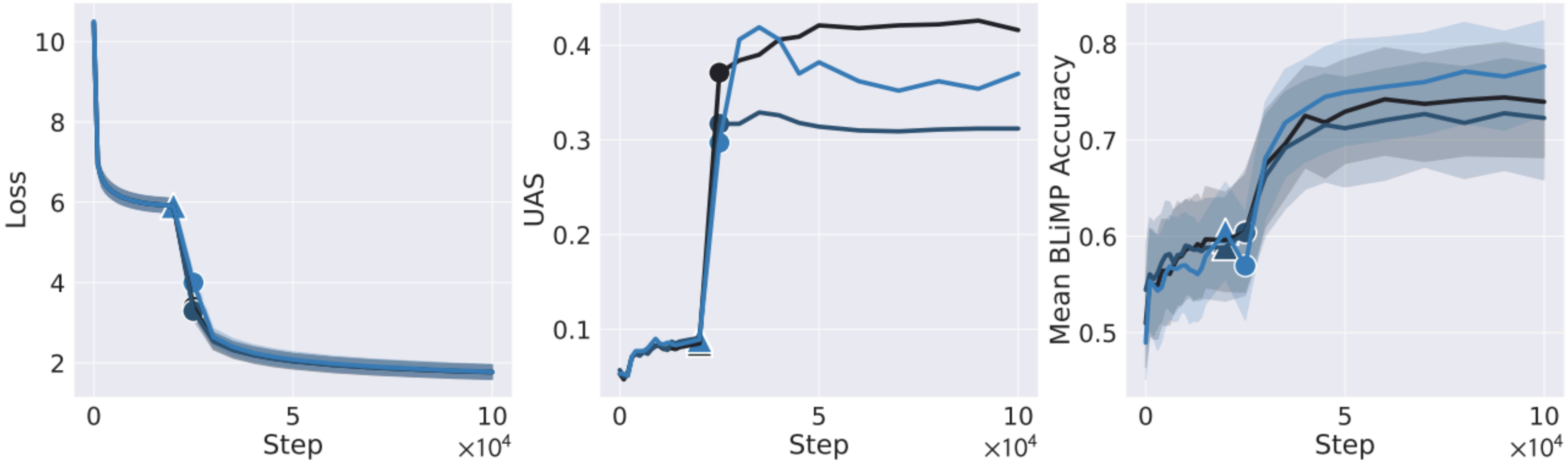
[1]University of Michigan, Ann Arbor
[2]Center for Brain Science, Harvard University

NeurIPS 2024

Introduction

- *Abrupt Learning*: Sudden drop in loss while training, with a jump in model performance

Question: Why do Transformers show abrupt learning while training?



[Chen et al. '24] Sudden Drops in the Loss: Syntax Acquisition, Phase Transitions, and Simplicity Bias in MLMs. ICLR 2024

Understanding Transformers Using Math

Model: Practically useful, easy to analyze

- Masked Language Model (MLM) - BERT

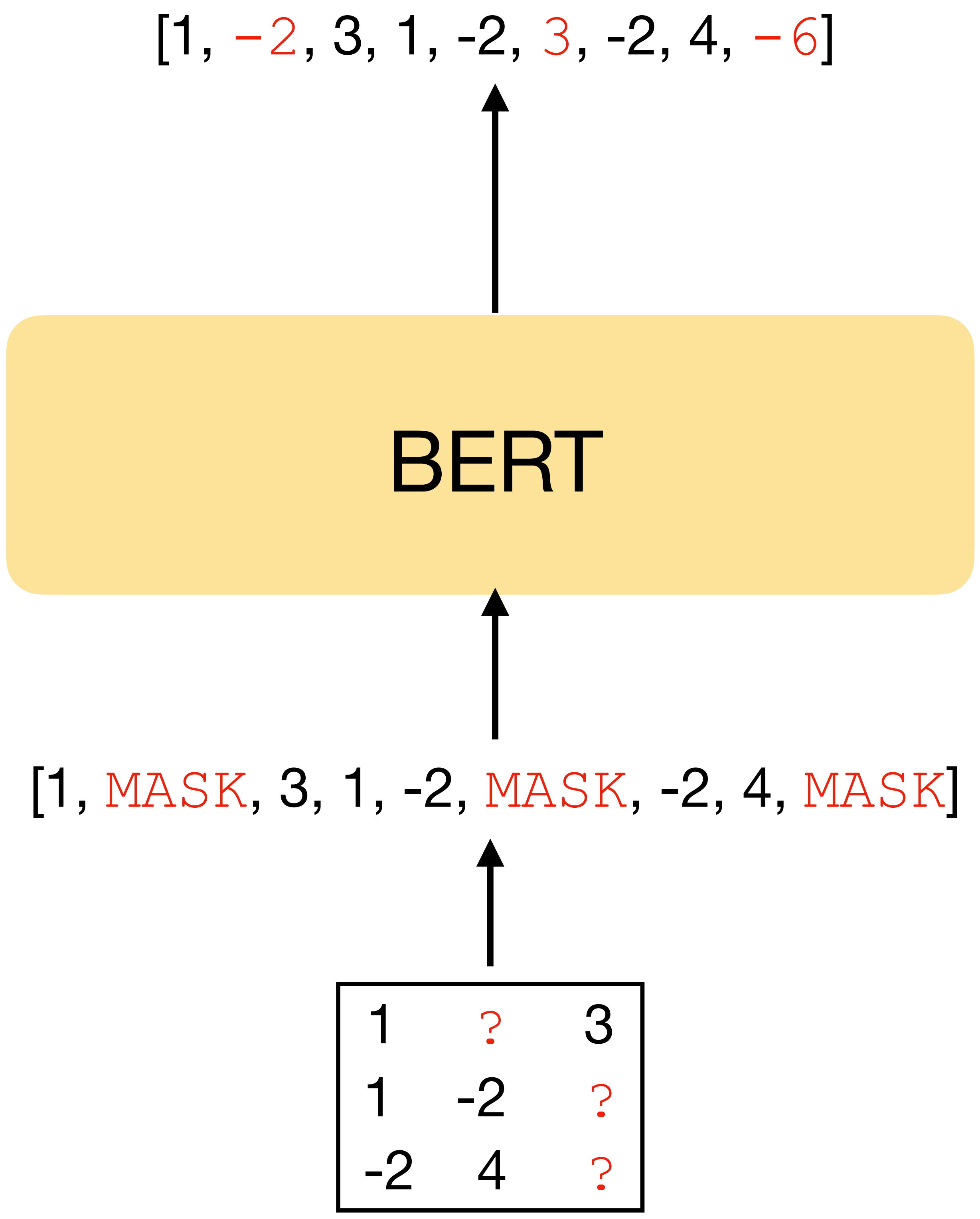
Data: Simple + controllable task, mathematical formulation

- Low Rank Matrix Completion (LRMC)

LRMC ↔ MLM

- LRMC is analogous to MLM

- Input matrix as a sequence; mask elements like words in MLM



1	?	3
1	-2	?
-2	4	?

Experimental Setup

- 4-layer BERT model; 8 Attention heads in each layer

- Input data sampled as

$$X = UV^\top, U, V \in \mathbb{R}^{7\times 2}$$

$$U_{ij}, V_{ij} \sim \mathrm{Unif}[-1,1]$$

i.e., X is 7x7, rank-2 matrix

- Online training on mean-squared-error (MSE) loss on all entries

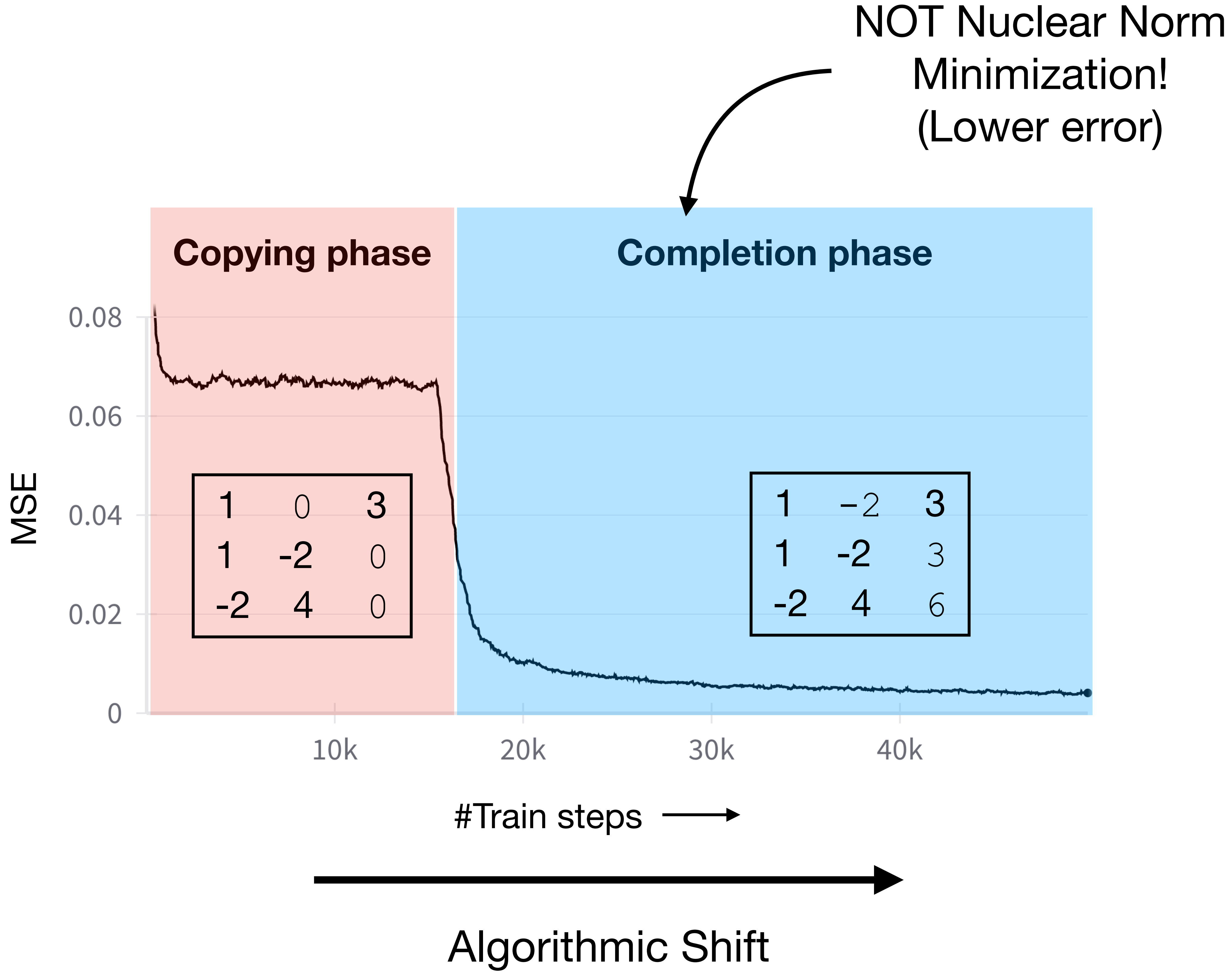
$$L = \frac{1}{n^2}\sum_{i,j=1}^{n}(\hat{X}_{ij} - X_{ij})^2$$

Results

- BERT can be trained to solve LRMC to low error

- Training BERT shows abrupt learning

- Changes in model components and mechanism after sudden drop



$$\begin{matrix} 1 & 0 & 3 \\ 1 & -2 & 0 \\ -2 & 4 & 0 \end{matrix}$$

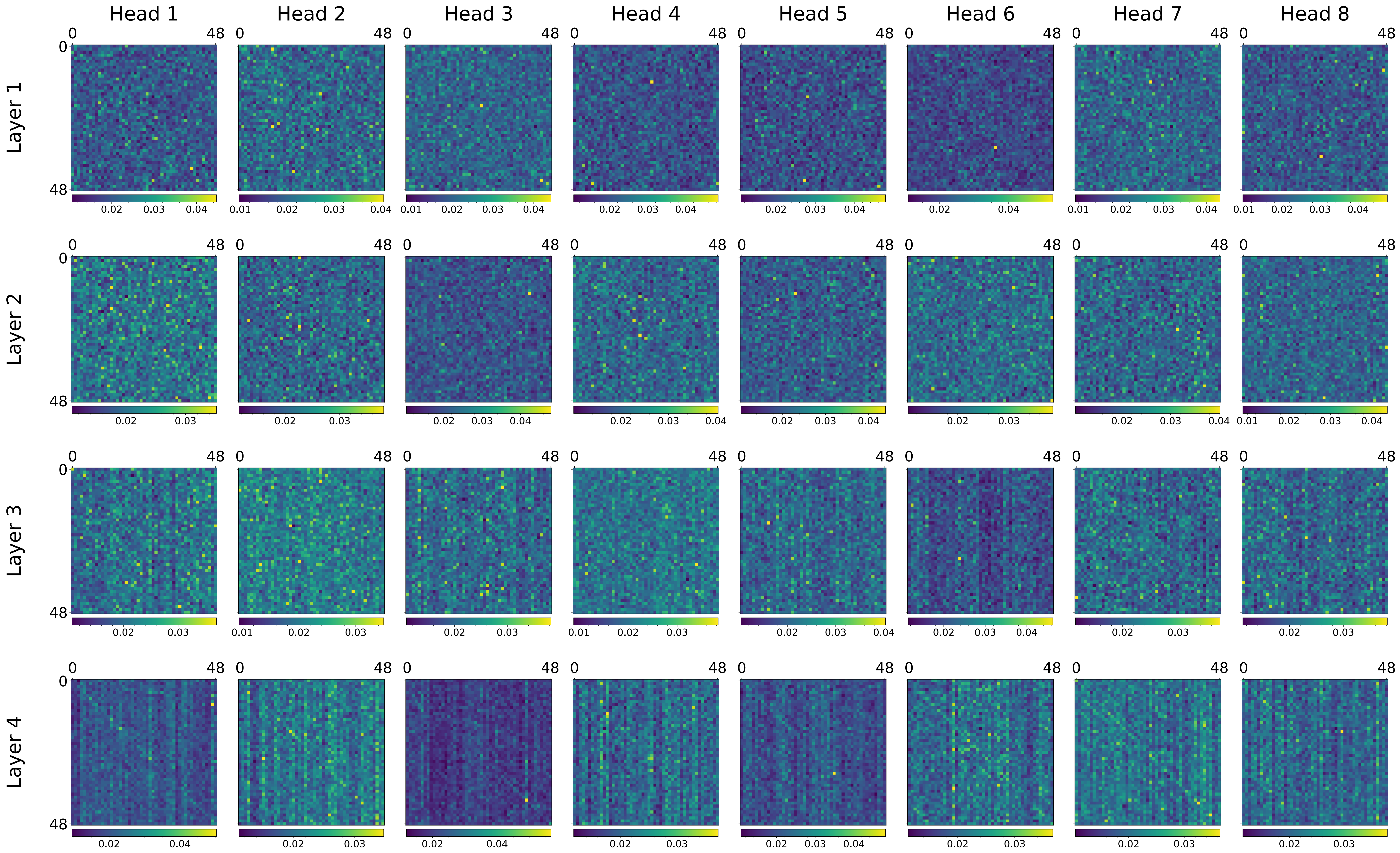
$$\begin{matrix} 1 & -2 & 3 \\ 1 & -2 & 3 \\ -2 & 4 & 6 \end{matrix}$$

NOT Nuclear Norm Minimization! (Lower error)

Algorithmic Shift

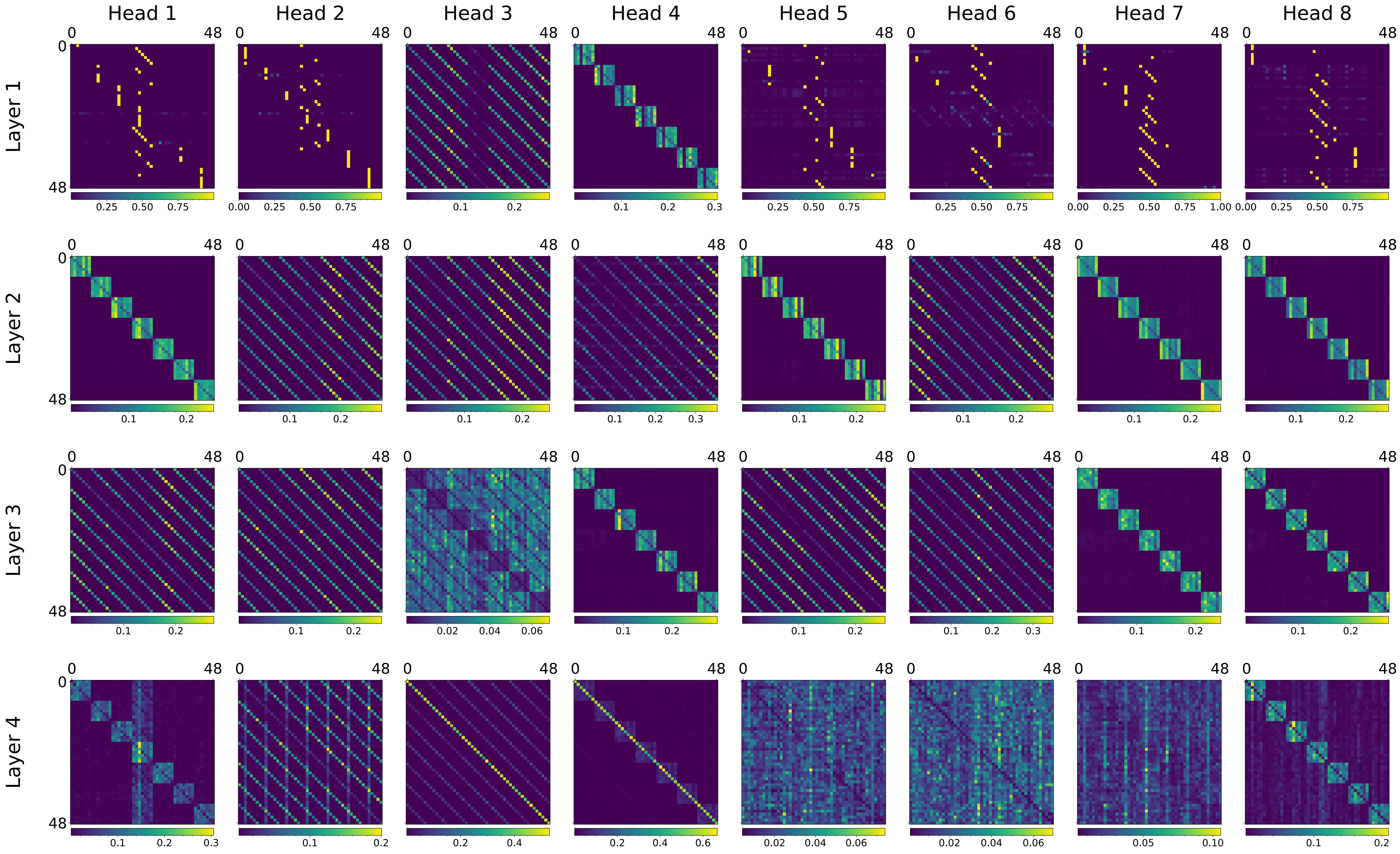
Attention Heads - Before the Sudden Drop

No clear interpretation of how various heads combine different elements in the input



Attention Heads - After the Sudden Drop

Different attention heads attend to different interpretable parts of the input



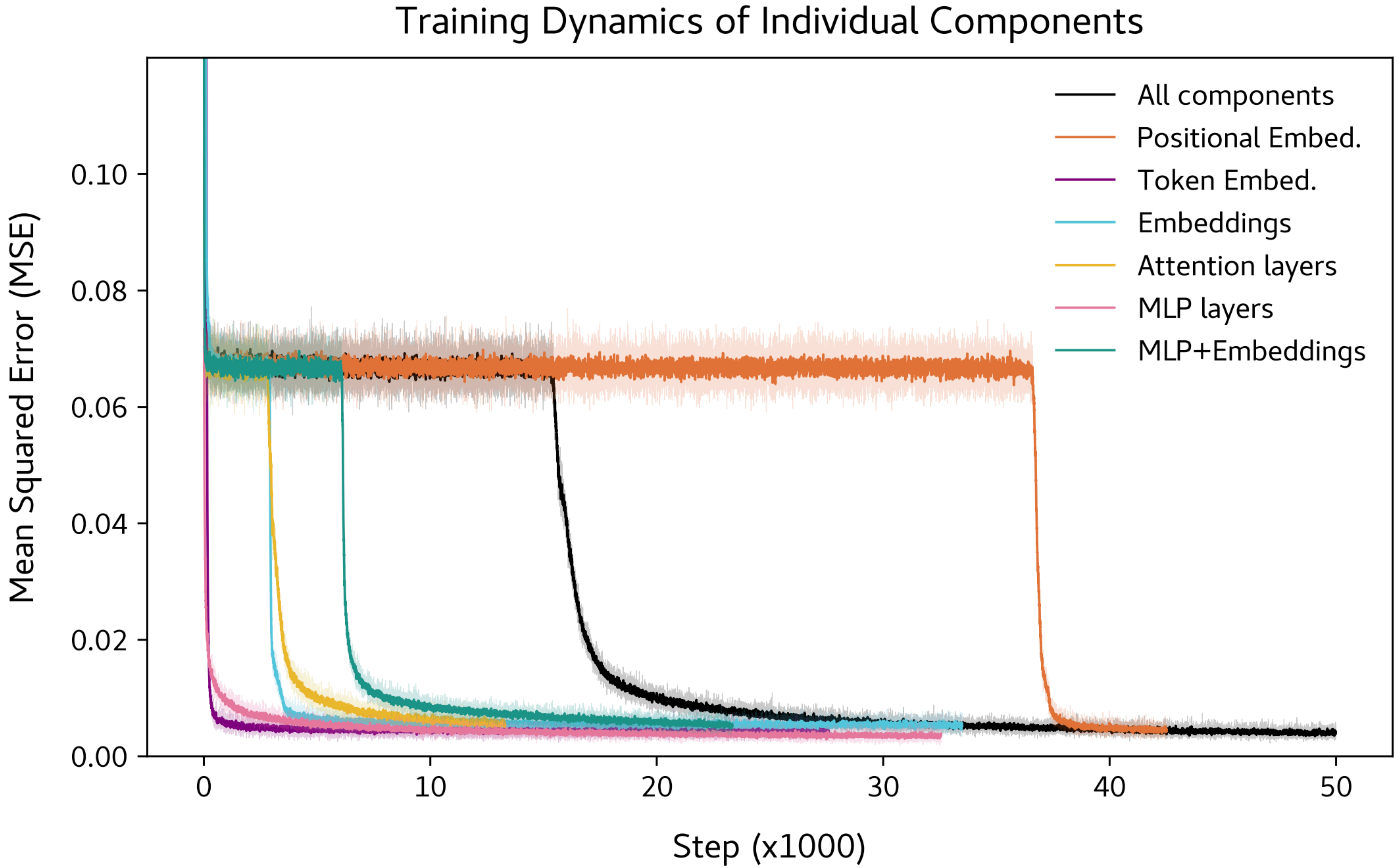
Sudden Drop in Loss

- Different components have qualitatively different computational roles

- Can we understand training dynamics of the full model through dynamics of parts of the model?

- Train each ‘component’ individually, fix the others to value @ $t = 50K$

- Component: Positional / Token Embedding, Attention layers, MLP layers


Training Dynamics of Individual Components
Mean Squared Error (MSE)
0.10
0.08
0.06
0.04
0.02
0.00
0
10
20
30
40
50
Step (x1000)
All components
Positional Embed.
Token Embed.
Embeddings
Attention layers
MLP layers
MLP+Embeddings

Hypothesis

- Based on our results, we hypothesize,

Learning required structure from data through components like Attention layers, embeddings is what leads to sudden drop in loss observed in training Transformers.

Thank You!