

# Interpretability using Dictionary Learning

Mohammad Taufeeque

Research Engineer, FAR AI  
Research Manager, Axiom Futures

Visit our website:



# Overview

1. What do Neural Networks learn?
2. Need for better tools for studying information flow in neural networks
3. Existing Approaches
  - a. Mechanistic Interpretability
  - b. Causal Abstraction
4. Dictionary Learning
  - a. Codebook Features
  - b. Sparse AutoEncoder (SAE)
5. Interpreting Features

# Notebooks

- [https://colab.research.google.com/github/taufeeque9/codebook-features/blob/main/tutorials/code\\_intervention.ipynb](https://colab.research.google.com/github/taufeeque9/codebook-features/blob/main/tutorials/code_intervention.ipynb)
- [https://colab.research.google.com/github/taufeeque9/codebook-features/blob/main/tutorials/tok\\_fsm.ipynb](https://colab.research.google.com/github/taufeeque9/codebook-features/blob/main/tutorials/tok_fsm.ipynb)

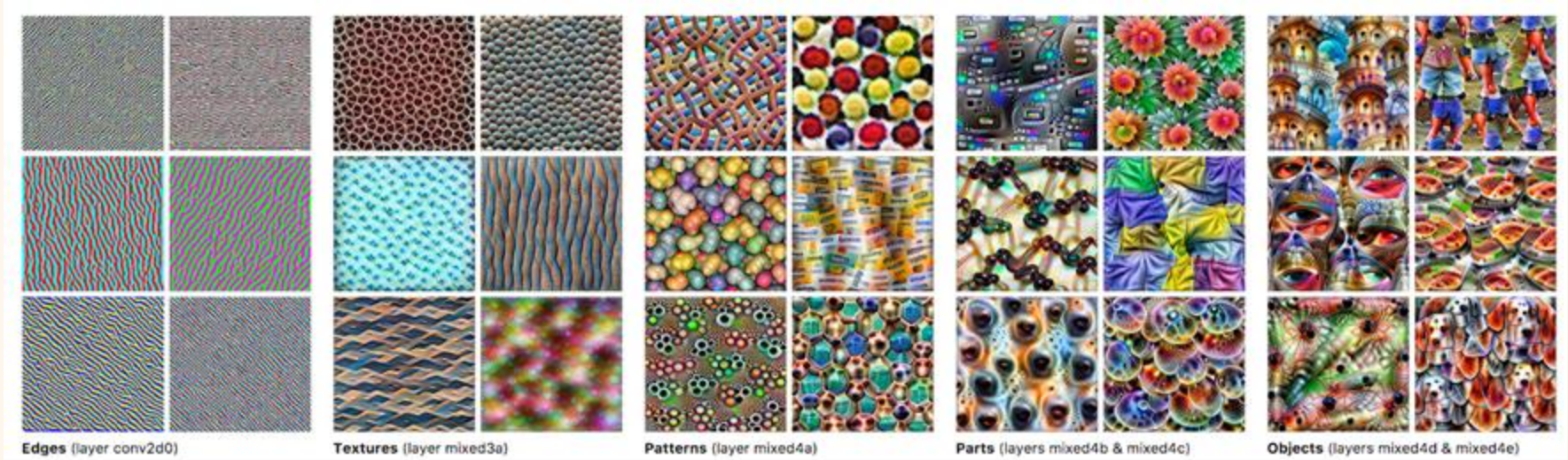
# What do Neural Networks learn?



# **Need for better tools for studying information flow in neural networks**



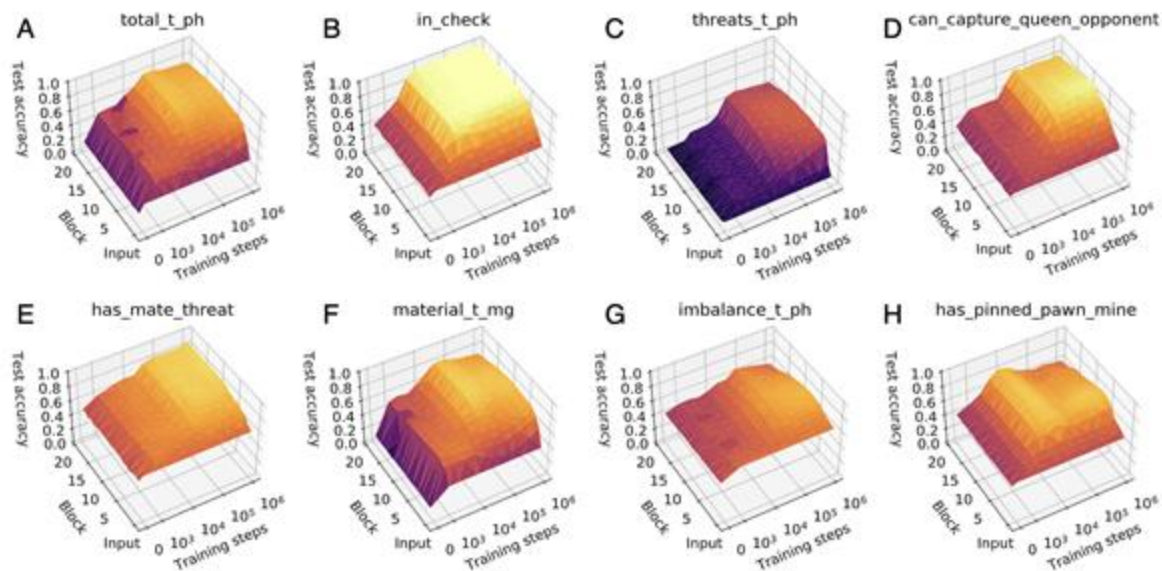
# Neural nets seem to learn hierarchies of features



Images (Olah+ 18)

# Neural nets seem to learn hierarchies of features

Chess  
(McGrath+21)



**Fig. 2.** What-when-where plots for a selection of Stockfish 8 and custom concepts. Following Fig. 1, we count a ResNet “block” as a layer. (A) Stockfish 8’s evaluation of total score. (B) Is the playing side in check? (C) Stockfish 8’s evaluation of threats. (D) Can the playing side capture the opponent’s queen? (E) Could the opposing side checkmate the playing side in one move? (F) Stockfish 8’s evaluation of “material score.” (G) Stockfish 8’s material score. Past  $10^5$  training steps this becomes less predictable from AlphaZero’s later layers. (H) Does the playing side have a pawn that is pinned to the king?

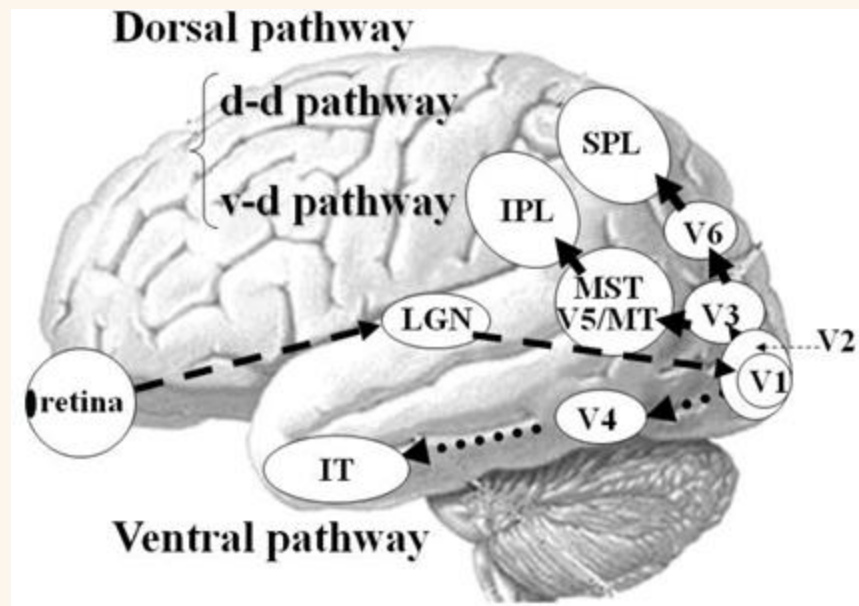
# Need better tools for studying information flow in neural networks

Understand circuits in neural networks (also see Olah+ 2020)

"Go fishing" for scientific concepts

Be able to edit, do counterfactuals, have causal models, (Geiger+2022)

Rewire models to make them safer





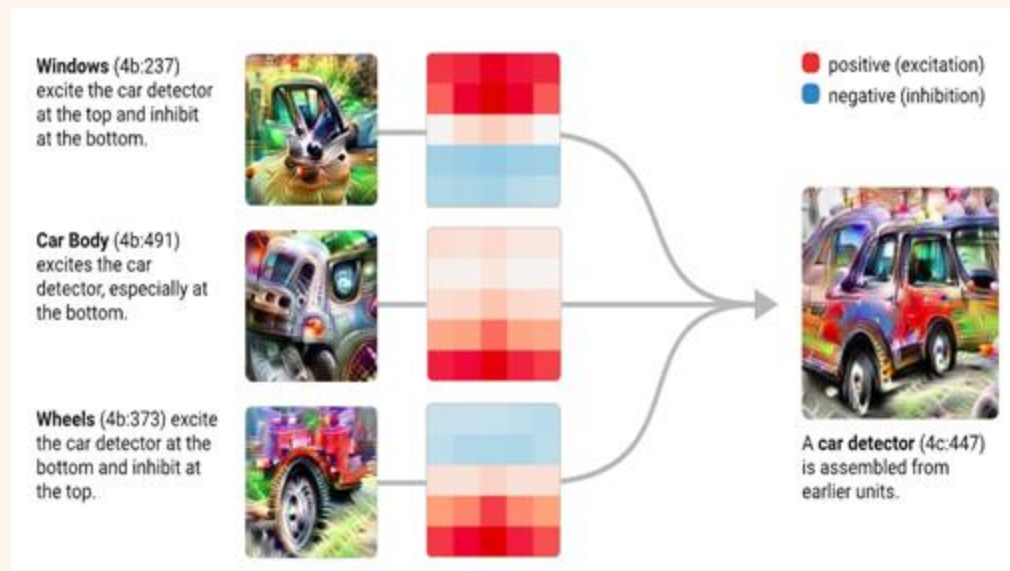
# Existing Approaches



# Mechanistic Interpretability

Reverse-engineering neural nets by studying weights, activations, and features

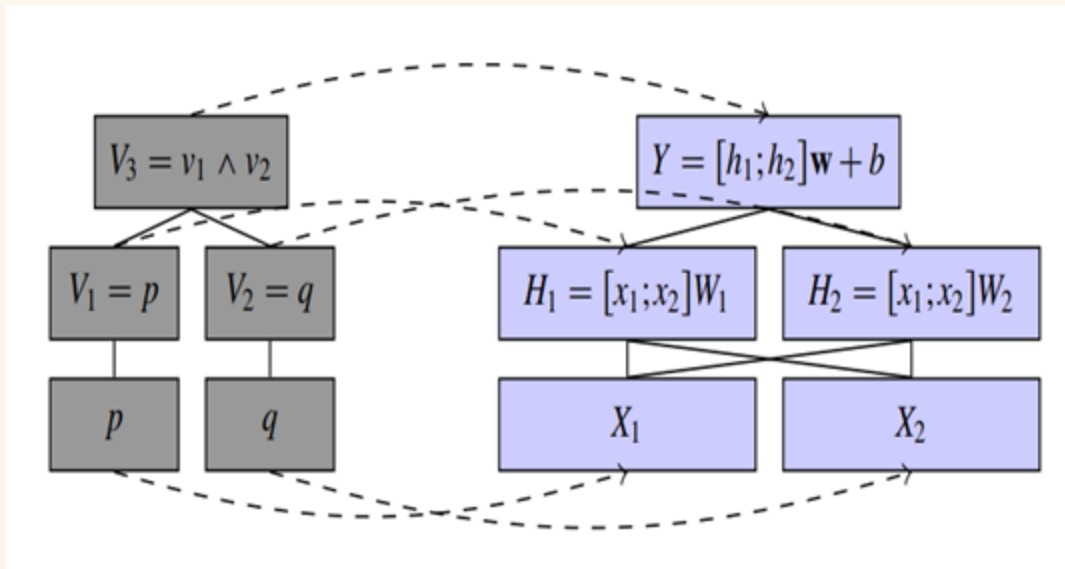
Superposition of features in neurons makes it difficult



# Causal Abstraction

Finds a high level causal model within a network

Limitations: Superposition, presupposition of model

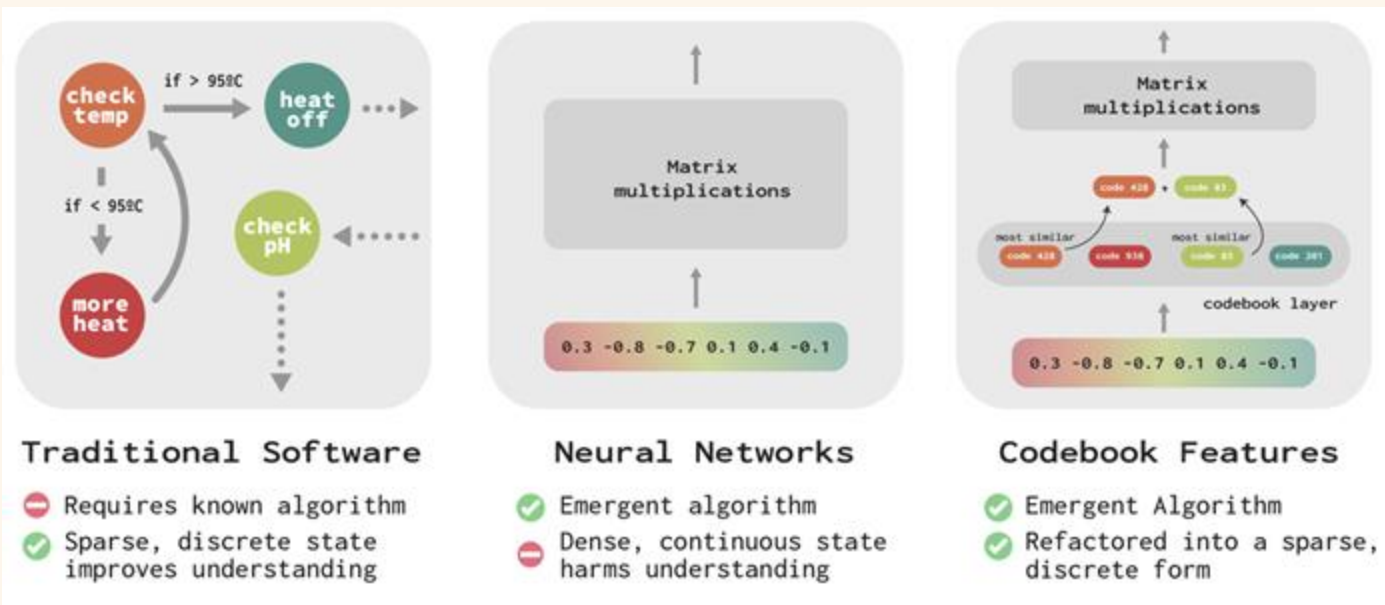


# Dictionary Learning

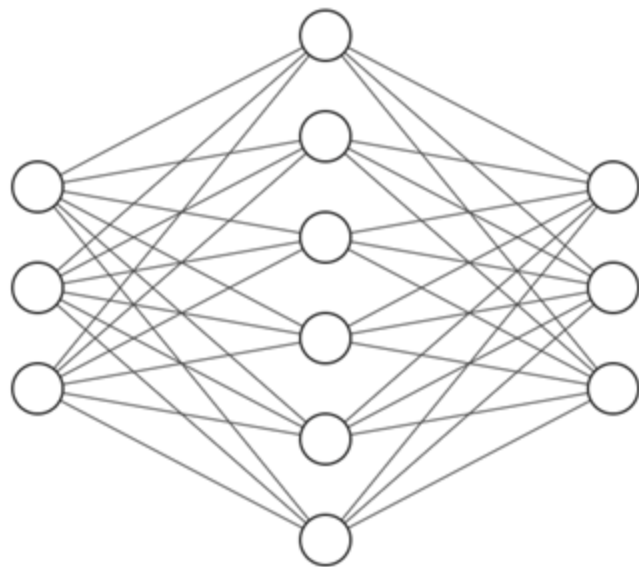


# Codebook Features

- Interpreting activations and weights of NNs is extremely difficult
- Can we constraint activations within a NN to a discretized set of features?

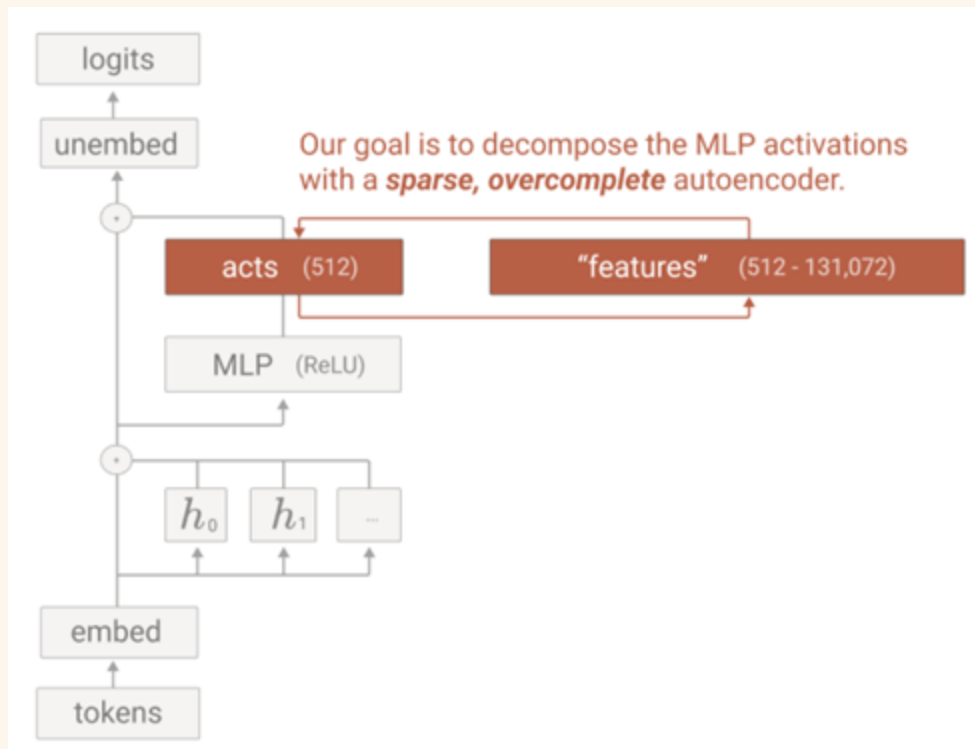


# SAE



Input Layer  $\in \mathbb{R}^3$     Hidden Layer  $\in \mathbb{R}^6$     Output Layer  $\in \mathbb{R}^3$

# SAE

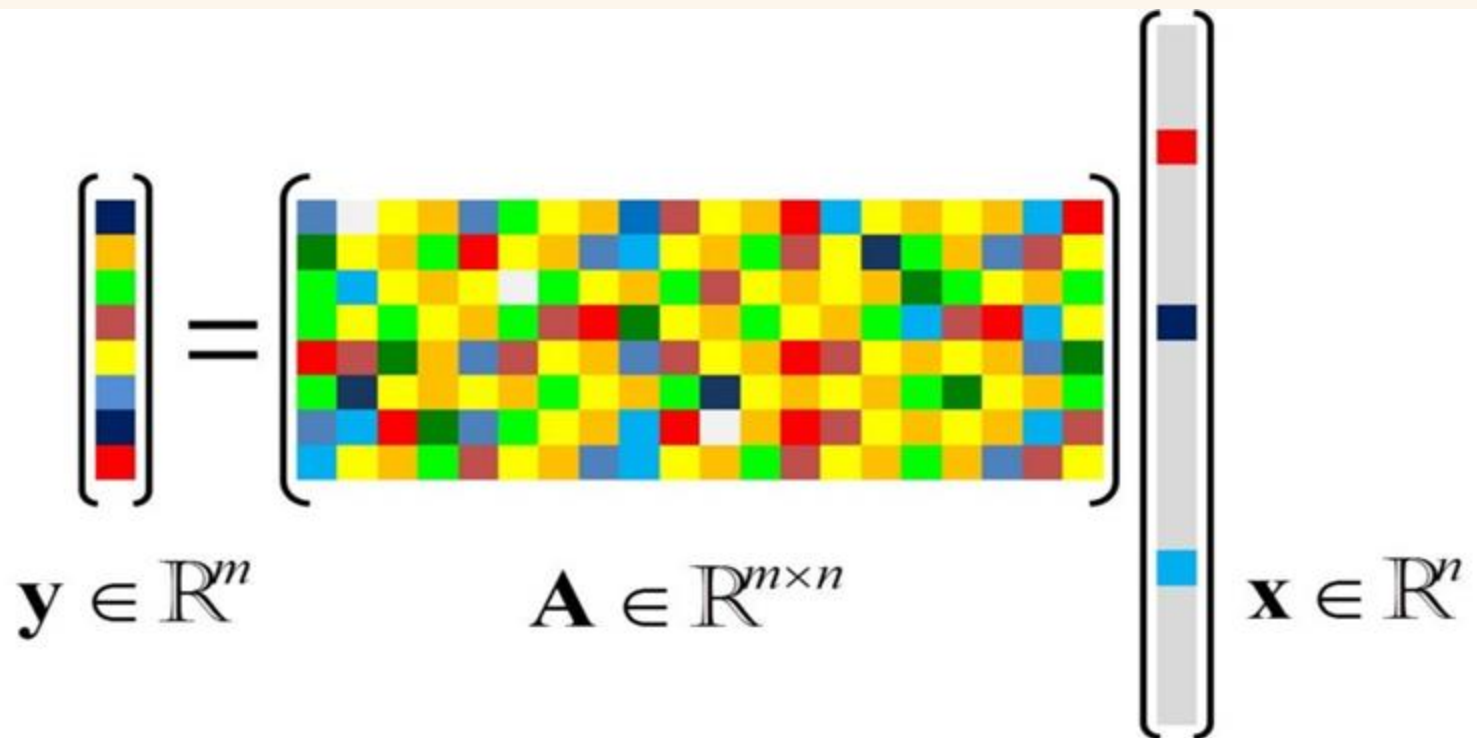


# Motivation

- Codebook/SAE can ideally separate features into different vectors
- Codebook/SAE try to find the causal structure in an unsupervised way unlike causal abstraction methods



# Sparse Coding



The diagram illustrates the sparse coding equation  $y = Ax$ . On the left, a vertical vector  $y \in \mathbb{R}^m$  is shown with 6 colored blocks (dark blue, orange, green, brown, yellow, dark blue) and a red block at the bottom. In the center is a matrix  $A \in \mathbb{R}^{m \times n}$  with a grid of colored squares (yellow, green, blue, red, brown, dark blue, light blue, white) and an equals sign to its left. On the right, a vertical vector  $x \in \mathbb{R}^n$  is shown with a long grey bar and a few colored blocks (red, dark blue, light blue) indicating non-zero entries.

$$\mathbf{y} \in \mathbb{R}^m = \mathbf{A} \in \mathbb{R}^{m \times n} \mathbf{x} \in \mathbb{R}^n$$

# Codebook

$$\begin{bmatrix} \text{colored vector} \end{bmatrix} = \begin{bmatrix} \text{codebook matrix} \end{bmatrix} \begin{bmatrix} \text{binary vector} \end{bmatrix}$$

$\mathbf{y} \in \mathbb{R}^m$        $\mathbf{A} \in \mathbb{R}^{m \times n}$        $\mathbf{X} \in \{0, 1\}^n, \|\mathbf{X}\|_0 = k$



**Questions till now?**



# Method: Codebook

1. Initialize codebook at every layer of the network with random vectors
2. Define the loss function:

$$\mathbb{L} = \mathbb{L}_{base} + \sum_{i=1}^N \|C_{out}^i - sg[C_{in}^i]\|_2^2$$

1. Train the codebooks and the model parameters using gradient descent and straight-through estimator

# Interpreting the Learned Features



# Codebook Models are Capable Language Models

(a) TinyStories 1-Layer Model

Language Model	Loss	Acc
*Pretrained	1.82	56.22
Finetuned	1.57	59.27
†Attn, $k = 8$	1.66	57.91
MLP, $k = 100$	1.57	59.47

(b) WikiText-103 410M 24-Layer Model

Language Model	Loss	Acc
*Finetuned (Wiki)	2.41	50.52
Finetuned 160M (Wiki)	2.72	46.75
†Attn, $k = 8$	2.74	46.68
Attn, $k = 64$	2.55	48.44
MLP, $k = 100$	3.03	42.47
MLP, grouped $16 \times (k = 64)$	2.57	48.46

# Feature Activations

(a) WikiText-103

Code	Interpretation	Example Activations
7.12.7884	Months (after preposition)	at Toulon in <b>August</b> The ship began trials [...] and spent three weeks in <b>September</b> attached to 14 : 30 on 7 <b>December</b> . The division had the [...] a major attack until 8 <b>December</b> on <b>August</b> 31, a Utah [...] On <b>September</b> 1, 1987
4.15.6101	Evaluative words	Initially , the New Zealand attack progressed <b>well</b> Superman from the main timeline is <b>successfully</b> teleported into only HWMs evaluated as " <b>excellent</b> " are used by NHC
1.9.295	Names starting with 'B'	In one account from the Bah <b>amas</b> , a mating pair ascended while John and Roy <b>Boulting</b> noted that [...] <b>Bocks</b> car, sometimes called <b>Bock</b> 's Car, is the name of the United States Army Air Forces B- <b>29</b> bomber
4.14.4742	Years in 2000s	As of <b>2011</b> , the International Shark Attack File lists In <b>2014</b> , a study at the University of Amsterdam with Fabian Cancellara kicked off his <b>2010</b> campaign with an overall victory at the Tour of
9.3.3727	Square Units	Atlanta encompasses 134.0 <b>square miles</b> (347.1km <sup>2</sup> ) it covered more than 55 square metres (590 <b>sq</b> ft) 6 percent or 101,593 <b>square</b> kilometres (39,225 <b>sq</b> mi) of [...]



# Steering using Topic Codes

Topic	Codes	Original generations	Steered generations
Video game	18	<b>The war was fought</b> on two fronts. The war was initiated in 1914 between Austria-Hungary and Serbia, when the Entente Powers signed a treaty of friendship between the two countries. In October 1914, Tschichky was sent to defend the German Empire'	<b>The war was fought</b> on both sides, and was only the second game to deal with one-on-one battles, following SimCity 2D Blade II. The game was released to critical acclaim, with praise particularly directed to the new console
Football	18	<b>The war was fought</b> on two fronts. The war was initiated in 1914 between Austria-Hungary and Serbia, when the Entente Powers signed a treaty of friendship between the two countries. In October 1914, Tschichky was sent to defend the German Empire'	<b>The war was fought</b> in its first forty years. In the summer of 1946, the Cardinals of the All-America Football Conference (AAFC) were rapidly becoming the favorites for NFL Hall-of-Fame coach Jim Mora, who had
Movie	12	<b>The novel was published in</b> November 2009 by MacChinnacle, a London publishing house. The book's publishers, Syco, published the book in the United Kingdom and the United States on 1 November 2009. The book received generally positive reviews from critics, who praised the	<b>The novel was published in</b> the United States and Canada. The film was directed by Joe Hahn and stars Steven Spielberg as Lucas, Neil Patrick Harris, and Jude Lawder as Lucas's best friend, Jonathan Miller. The plot follows a character (Lucas

# Thank you!

Visit our website:



Axiom Futures

Incubated by



Impact  
Academy