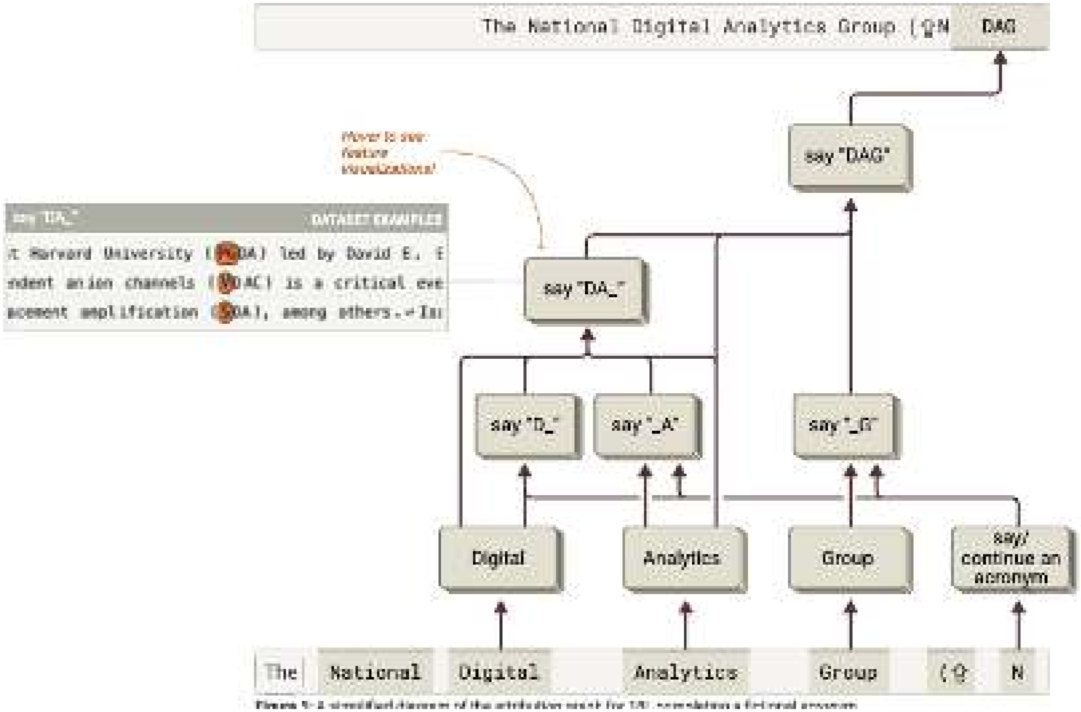
AI/ML Model Interpretability: Methods and Attribution Graphs

Understanding AI Model Internals

- **Goal**: Make AI models transparent and interpretable through systematic analysis
- **Challenge**: Understanding AI models is similar to biological research - complex systems requiring sophisticated tools



- **Outcomes**:
  - Make predictions about unexpected AI outputs
  - "Microscopes" for AI model internals https://transformer-circuits.pub/

The Landscape of Interpretability

- **Observation**: Model behaviors to understand underlying mechanisms
- **Sparse Autoencoders (SAEs)**: Identifying concepts (features) in the model

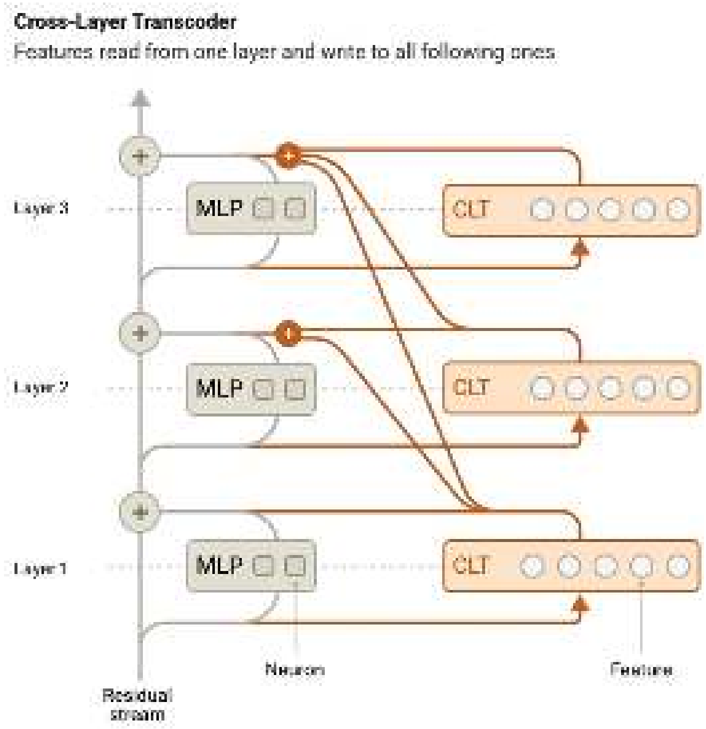
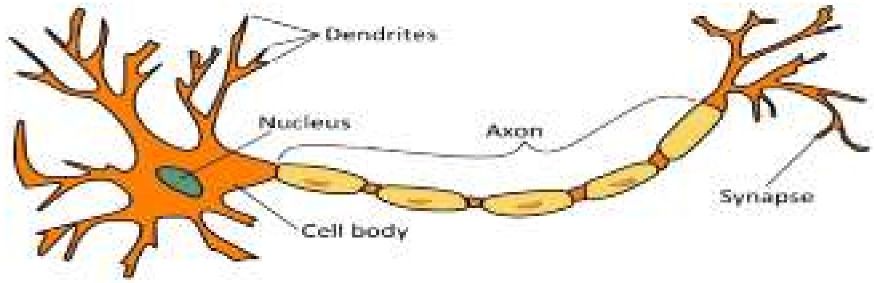
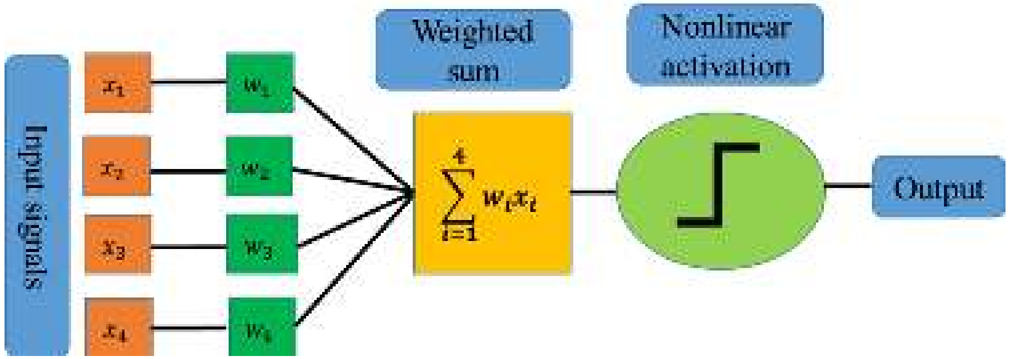


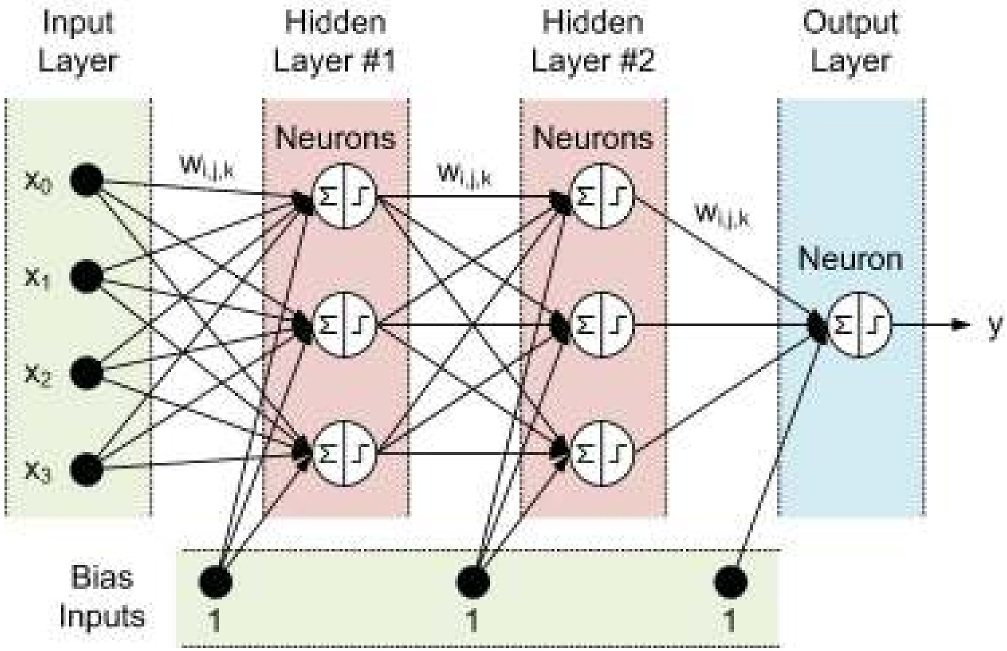
Figure 1: The cross-layer transcoder (CLT) forms the core architecture of our replacement model.

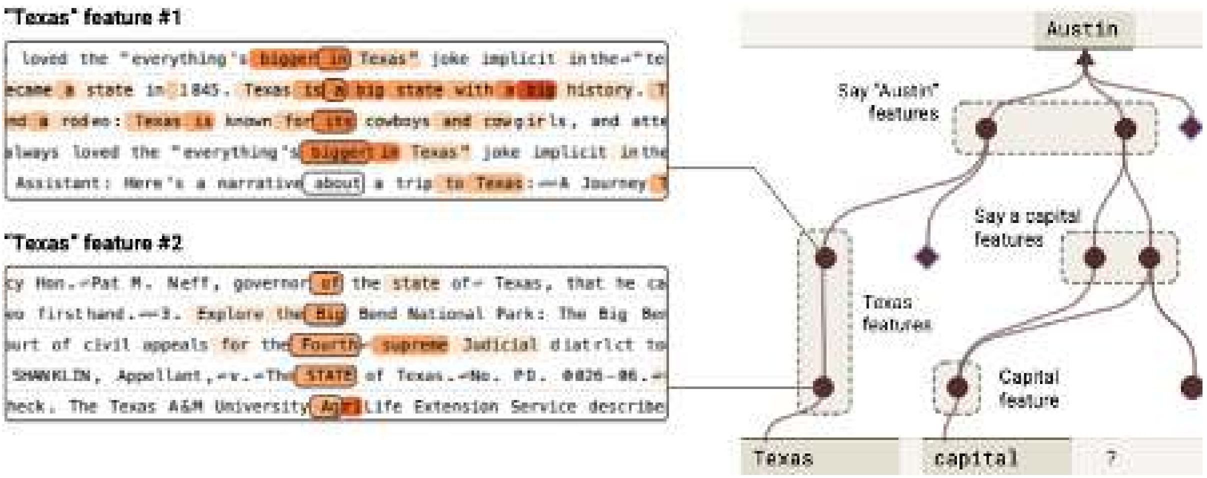
- **Linear Probes**: Internal linear representations of specific concepts
- **Intervention Experiments**: Steering, neural activation patching, and ablations

One slide overview


Dendrites
Nucleus
Axon
Synapse
Cell body


Input signals
x_1
x_2
x_3
x_4
w_1
w_2
w_3
w_4
Weighted sum
$\sum_{i=1}^{4} w_i x_i$
Nonlinear activation
Output


Input Layer
Hidden Layer #1
Hidden Layer #2
Output Layer
Neurons
Neurons
Neuron
x_0
x_1
x_2
x_3
$w_{i,j,k}$
$w_{i,j,k}$
$w_{i,j,k}$
y
Bias Inputs
1
1
1



Ameisen, et al., "Circuit Tracing: Revealing Computational Graphs in Language Models", Transformer Circuits, 2025.

Attribution Graphs for Studying Model Biology

- Compute interactions between features active on specific prompts
- Create interactive graphs showing feature-feature interactions
- Identify important interaction chains influencing model output
- Per-prompt analysis revealing computational pathways

Transformer Architecture: Circuit-Based View

- Each residual block contains:
  - **Attention Layer** followed by **MLP Layer**
  - Both layers "read" from and "write" to the residual stream
- **Attention Heads**: Independent operations outputting results added to residual stream
- **Linear Projections**: Read input from residual stream, write results back via addition

The transformer architecture can be viewed as a series of circuit components that process information through a central residual stream, enabling mathematical analysis of information flow.

The Residual Stream: Mathematical Properties

- Foundation for circuit-based interpretability methods
- **Additive Structure**: Each layer adds its output to the stream
- **End-to-End Functions**: Attention-only models can be written as sum of interpretable functions mapping tokens to logit changes
- Each layer **adds** its results into the residual stream
- Attention heads can be understood as **independent operations**

Elhage, et al., "A Mathematical Framework for Transformer Circuits", Transformer Circuits Thread, 2021.

The linear, additive structure of the residual stream is unique among neural architectures and provides a mathematical foundation for understanding transformer computations. This mathematical framework has enabled significant discoveries in mechanistic interpretability and provides tools for understanding transformer behavior.