AI/ML Model Interpretability

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

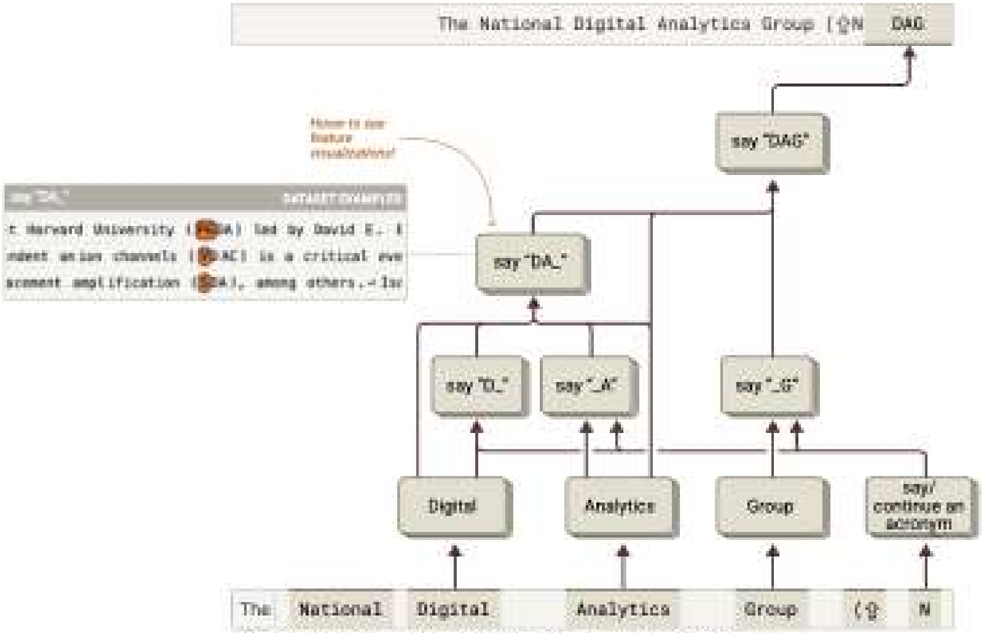
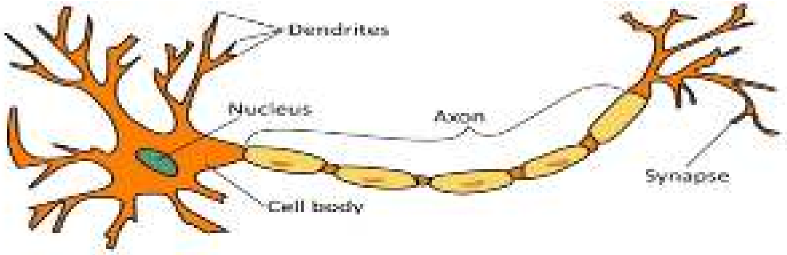

The National Digital Analytics Group (⇧N
DAG
say "DAG"
Hover to see feature visualizations!
say "DA_"
DATASET EXAMPLES
t Harvard University (DDA) led by David E. I
ndent anion channels (DAC) is a critical eve
ncement amplification (DDA), among others.- Isc
say "D_"
say "_A"
say "_G"
Digital
Analytics
Group
say/ continue an acronym
The
National
Digital
Analytics
Group
(⇧
N

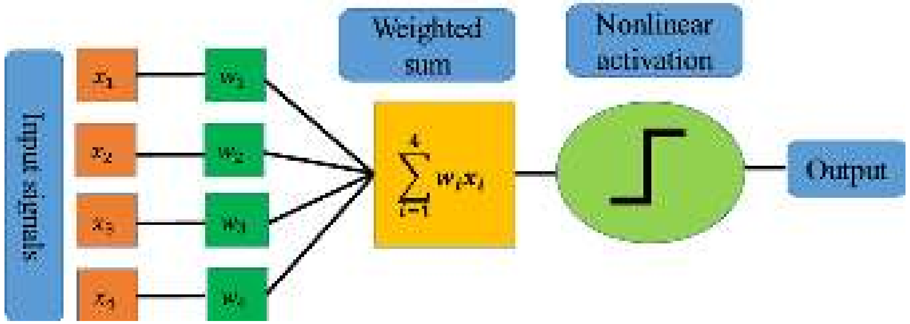
Figure 8: A simplified diagram of the attribution graph for 18L completing a fictional acronym.

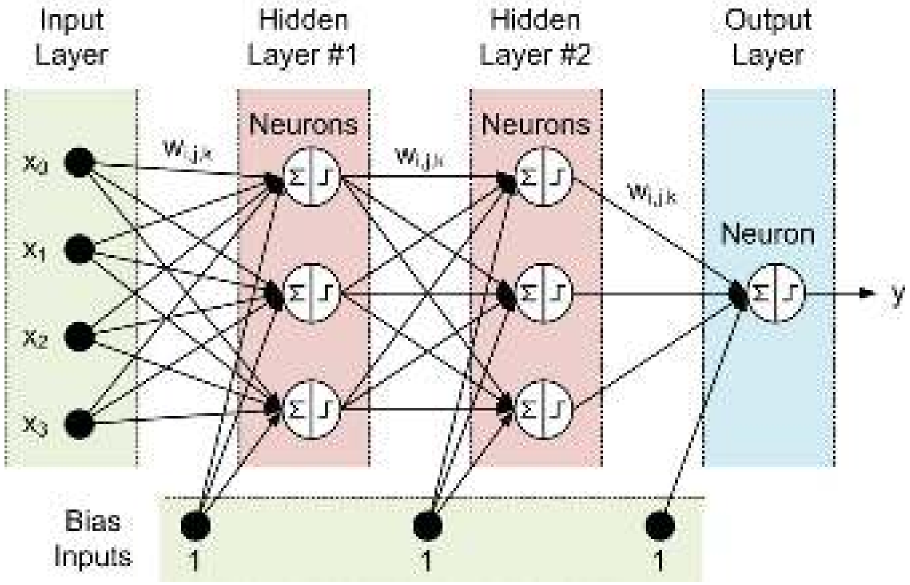
- Predictions about unexpected AI outputs
- "Microscopes" for AI model internals

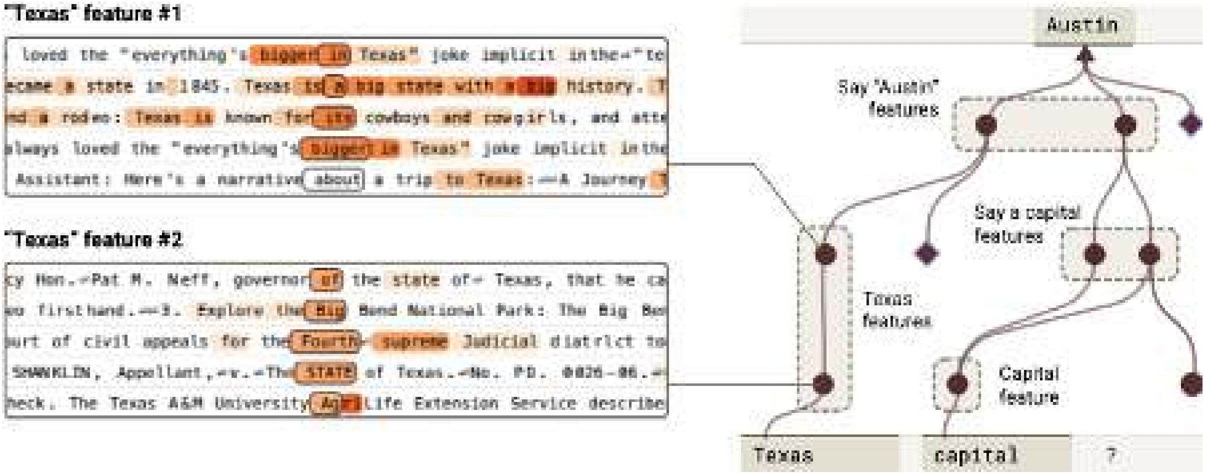
Lindsey, et al., "On the Biology of a Large Language Model", Transformer Circuits, 2025.

One slide overview (MLP=ANN=FF)









"Texas" feature #1

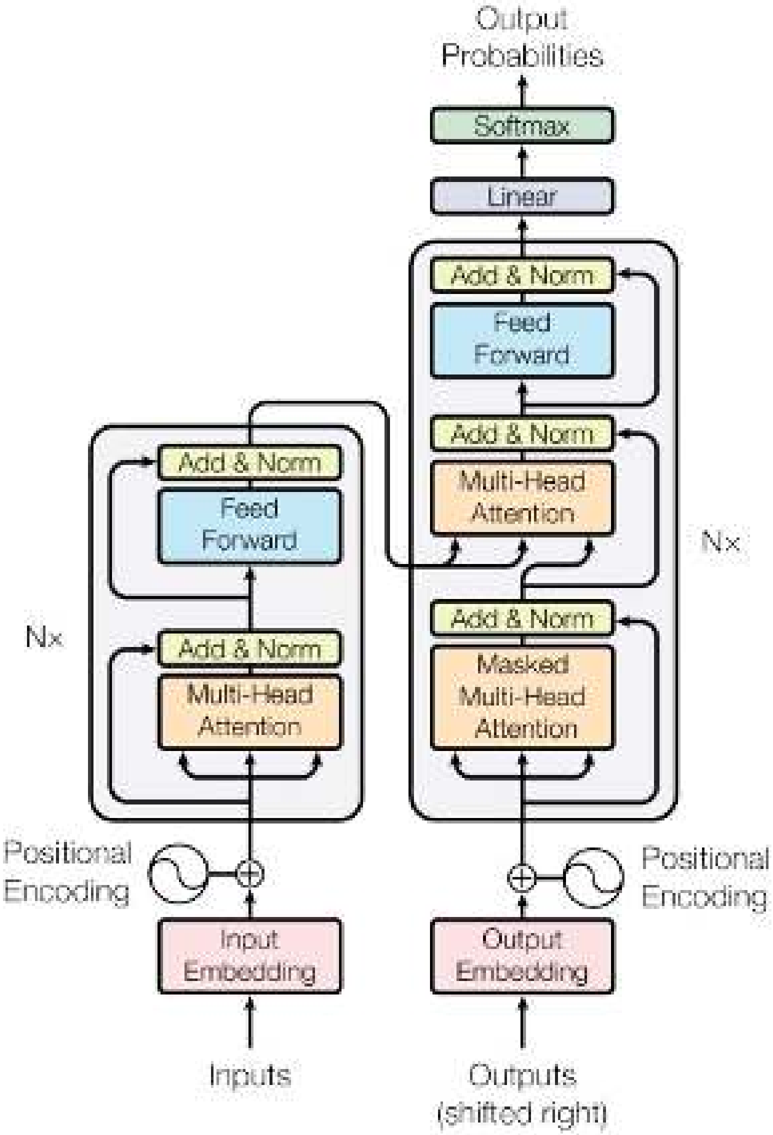
loved the "everything's bigger in Texas" joke implicit in the↵"te
ecame a state in 1845. Texas is a big state with a big history. T
nd a rodeo: Texas is known for its cowboys and cowgirls, and atte
always loved the "everything's bigger in Texas" joke implicit in the
Assistant: Here's a narrative about a trip to Texas:↵↵A Journey T

"Texas" feature #2

cy Hon.↵Pat M. Neff, governor of the state of↵ Texas, that he ca
so firsthand.↵↵3. Explore the Big Bend National Park: The Big Ben
ourt of civil appeals for the Fourth↵ supreme Judicial district to
SHANKLIN, Appellant,↵v.↵The STATE of Texas.↵No. PD. 0826-06.↵
heck. The Texas A&M University AgriLife Extension Service describe

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

Transformer Architecture of LLMs

- 

  Attention layer followed by **FF Layer**

- Attention Heads: long-range connections within text

- Feed Forward: same as in previous slide
- N x times : multiple heads run in parallel
- If you trained a neural network, similar plus attention

Vaswani A et al. Attention is all you need. arXiv:1706.03762. 2017;30

The Landscape of Interpretability

- Neurons are "Poly-semantic" (superposition phenomenon)
- Sparse Autoencoders: Identifying knowledge (features) stored in the neurons

Cross-Layer Transcoder

Features read from one layer and write to all following ones

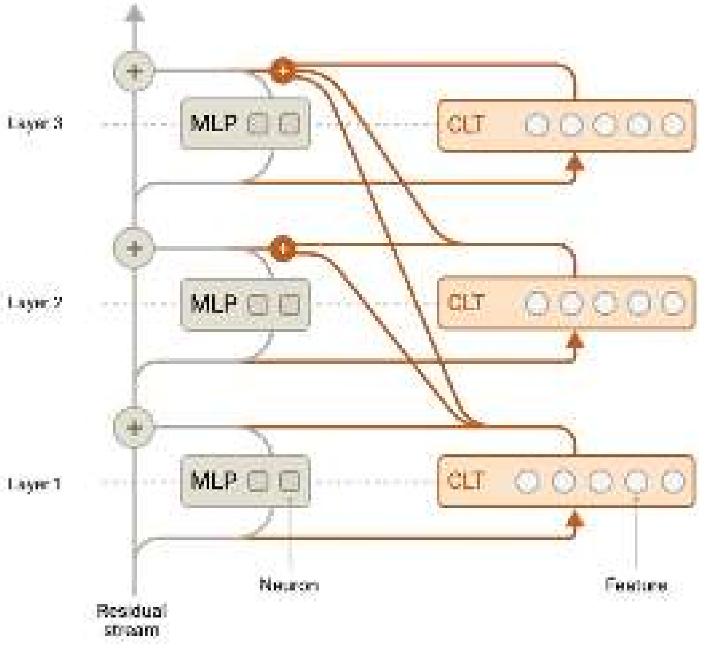


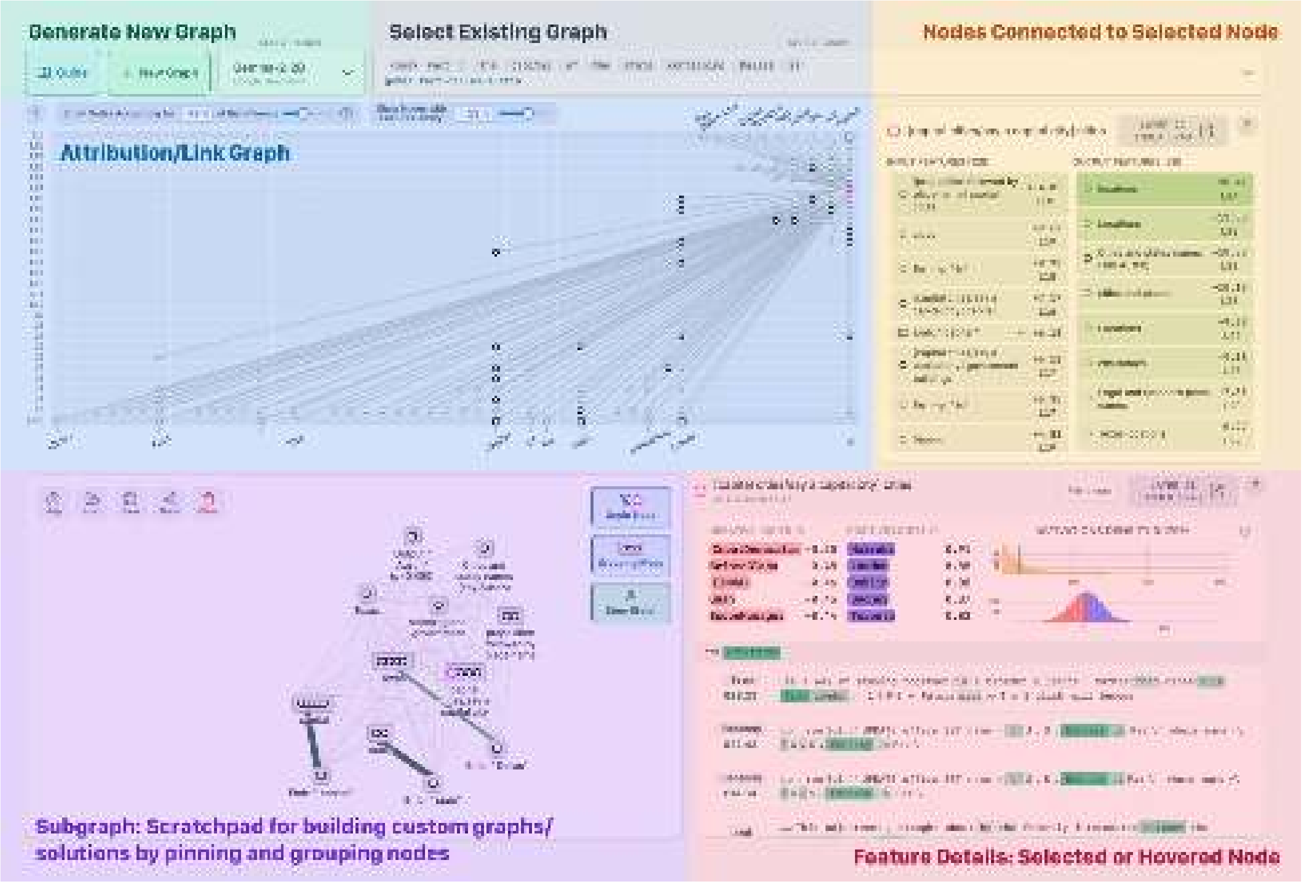
Figure 1: The cross-layer transcoder (CLT) forms the core architecture of our replacement model.

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

https://transformer-circuits.pub/

Attribution Graphs

- Interactions between features (text from input) activating neurons
- Graphs showing feature-feature interactions on specific prompts


Generate New Graph
Select Existing Graph
Nodes Connected to Selected Node
Attribution/Link Graph
Subgraph: Scratchpad for building custom graphs/ solutions by pinning and grouping nodes
Feature Details: Selected or Hovered Node

- Interaction chains influencing model output
- Prompt analysis revealing circuits

https://www.neuronpedia.org/

Thanks !

- https://github.com/KKrampis/presentations
- Transformers the tech behind LLMs https://tinyurl.com/3b1b-Transformers
- https://tinyurl.com/alignment-nanda-papers
- https://tinyurl.com/nanda-become-interp-researcher
- https://www.neelnanda.io/mechanistic-interpretability/quickstart