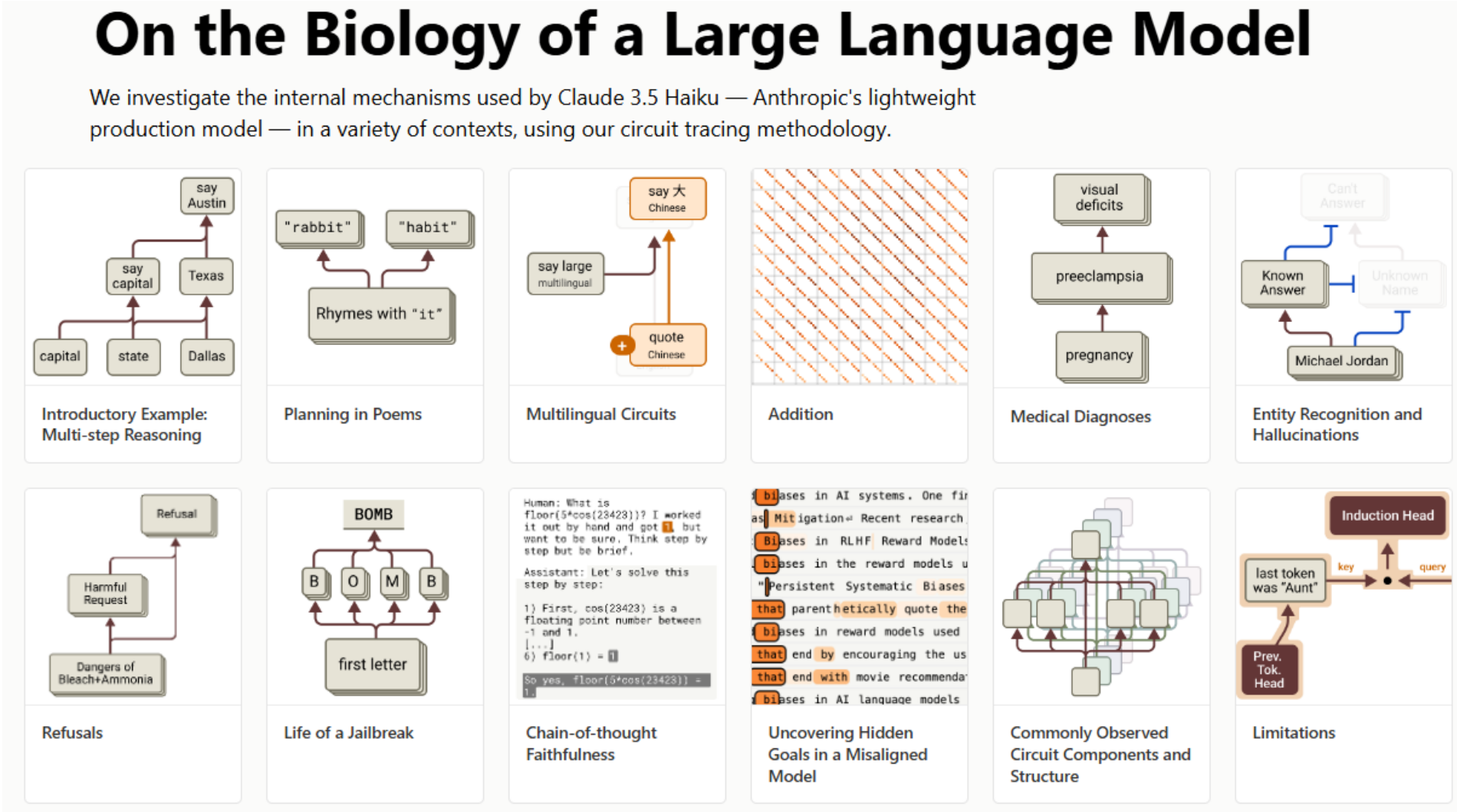
Machine Learning Department
School of Computer Science
Carnegie Mellon University

Reasoning Models
+
Mechanistic Interpretability

Pat Virtue & Matt Gormley
Lecture 26
Apr. 23, 2025

Mechanistic Interpretability



https://transformer-circuits.pub/2025/attribution-graphs/biology.html

Outline

Why is interpretability important?

Why is interpretability hard?

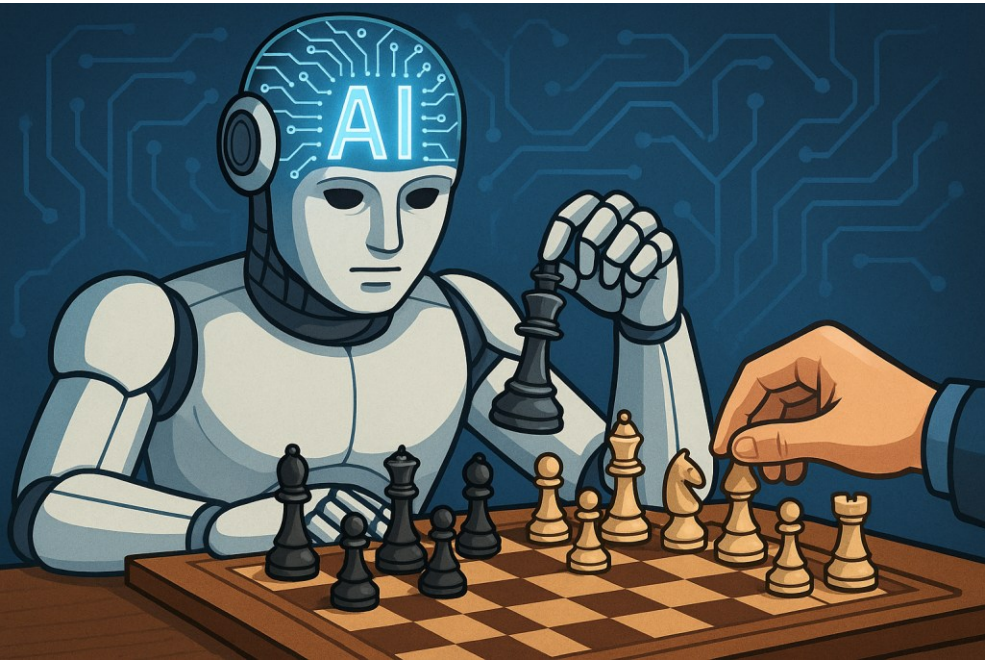
- Superposition

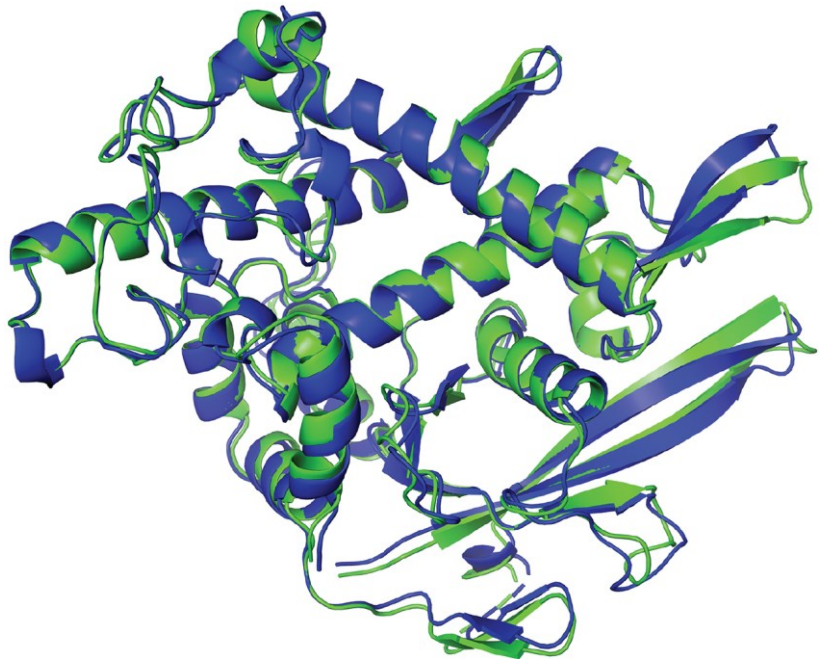
Mechanistic Interpretability Techniques

- Replacement models
- Sparse Autoencoders
- Circuits
- Cross layer transcoders
- Attribution Graph

Why is Interpretability Important

- Safety (Corrective Action)
- Safety (Preventative Action)
- Preventing AI Apocalypse
- Learning from AI





Why is Interpretability Hard?

- Ideally their would be a sparse set of neurons active for a given human-interpretable feature

Why is Interpretability Hard?

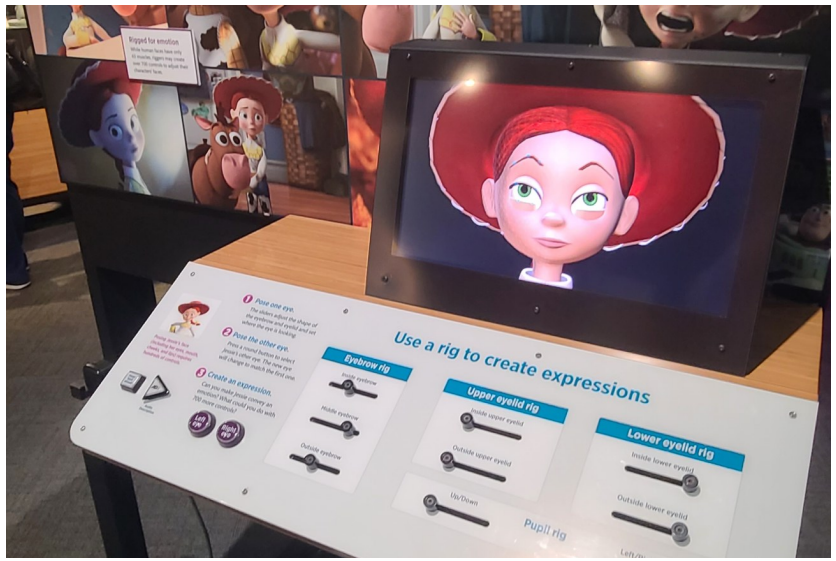
- Superposition is a major problem
  - "Features" rarely activate in a single location in the network
  - Activations for a human-interpretable feature is almost always distributed across many, many location within the networs, e.g. across heads, across MLP neurons, across layers

Replacement Models

- Idea: For specific blocks within a network, train a "replacement block" to mimic the functionality (input → ouput) of the original block...but importantly:
  - Make the replacement block more interpretable
- Techniques:
  - Sparse autoencoders
  - Circuits
  - Cross layer transcoders

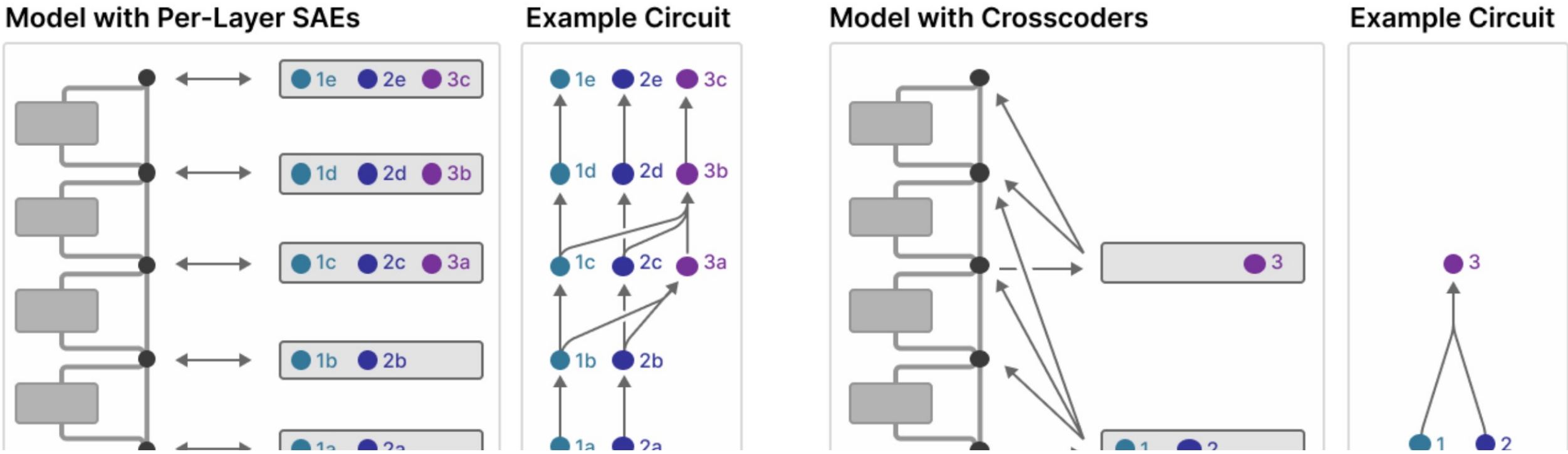
Sparse Autoencoders

- Replace MLP layer in transformer block with an autoencoder version with:
  - more neurons (features) in the hidden layer
  - an added regularization to encourage sparsity in the activations, e.g. L1 regularization



Circuits and Cross-layer Transcoders

- Cross-layer: allow replacement blocks to directly access all earlier replacement blocks



https://transformer-circuits.pub/2024/crosscoders/index.html

Outline

Why is interpretability important?

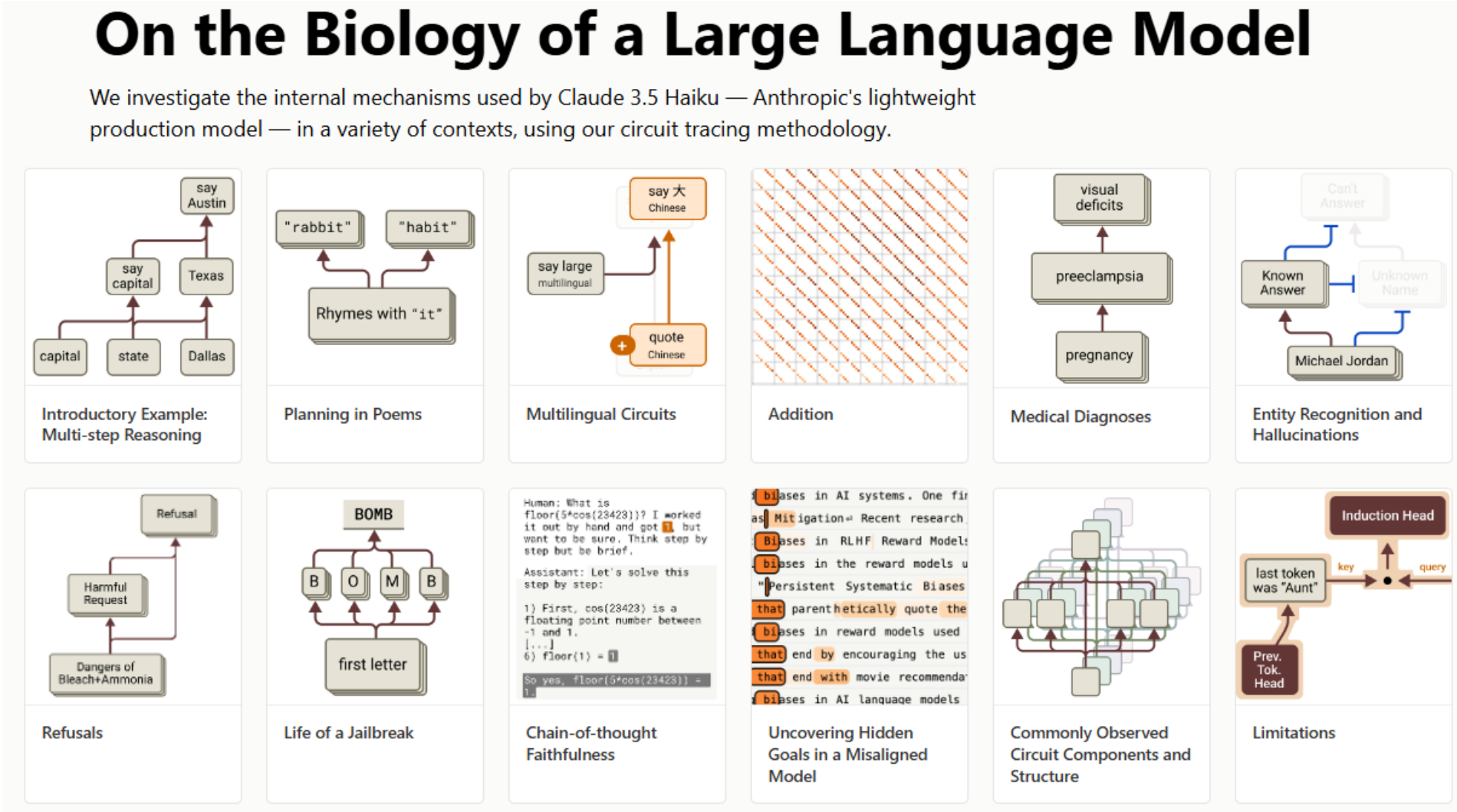
Why is interpretability hard?

- Superposition

Mechanistic Interpretability Techniques

- Replacement models
- Sparse Autoencoders
- Circuits
- Cross layer transcoders
- Attribution Graph

Mechanistic Interpretability



https://transformer-circuits.pub/2025/attribution-graphs/biology.html