

UNIVERSITY OF
MISSOURI-ST. LOUIS

An ongoing approach to interpret LLMs

Rumandeep Singh | Department Of Computer Science

08/23/2025



Fact: the capital of
the state containing
Dallas is

Fact: the capital of
the state containing
Dallas is

**State containing
Dallas is Texas**



**Capital of Texas
is Austin**

Q:

Do LLM's actually perform these two steps internally?

Or

Do they use some “shortcut”?

Modern LLM perform this multi-step reasoning, which coexists alongside the “shortcut” reasoning.

How can we verify this multi-step reasoning?

Mechanistic Interpretability

This technique aims to reverse-engineer these Neural networks.

We Look inside Neural Networks using

1. Features
2. Circuits

Looking Inside Neural Networks

We study the interactions between features to trace its intermediate steps it took to produce a responses.

Visualized using **attribution graphs**, a graphical representation of the computational steps the model uses to determine its output for a particular input

- Nodes represent features
- Edges represent the causal interactions between them.

Looking Inside Neural Networks

We use Neuronpedia.org which is an open platform for ML.

We are using circuit tracer which is hosted there

- It allows to interactively trace internal reasoning steps and generate our own graphs with custom prompts.
- In general, our prompt should be "missing" a word at the end, because we want to analyze how the model comes up with that word.

Looking Inside Neural Networks

We use Neuronpedia.org which is an open platform for ML.

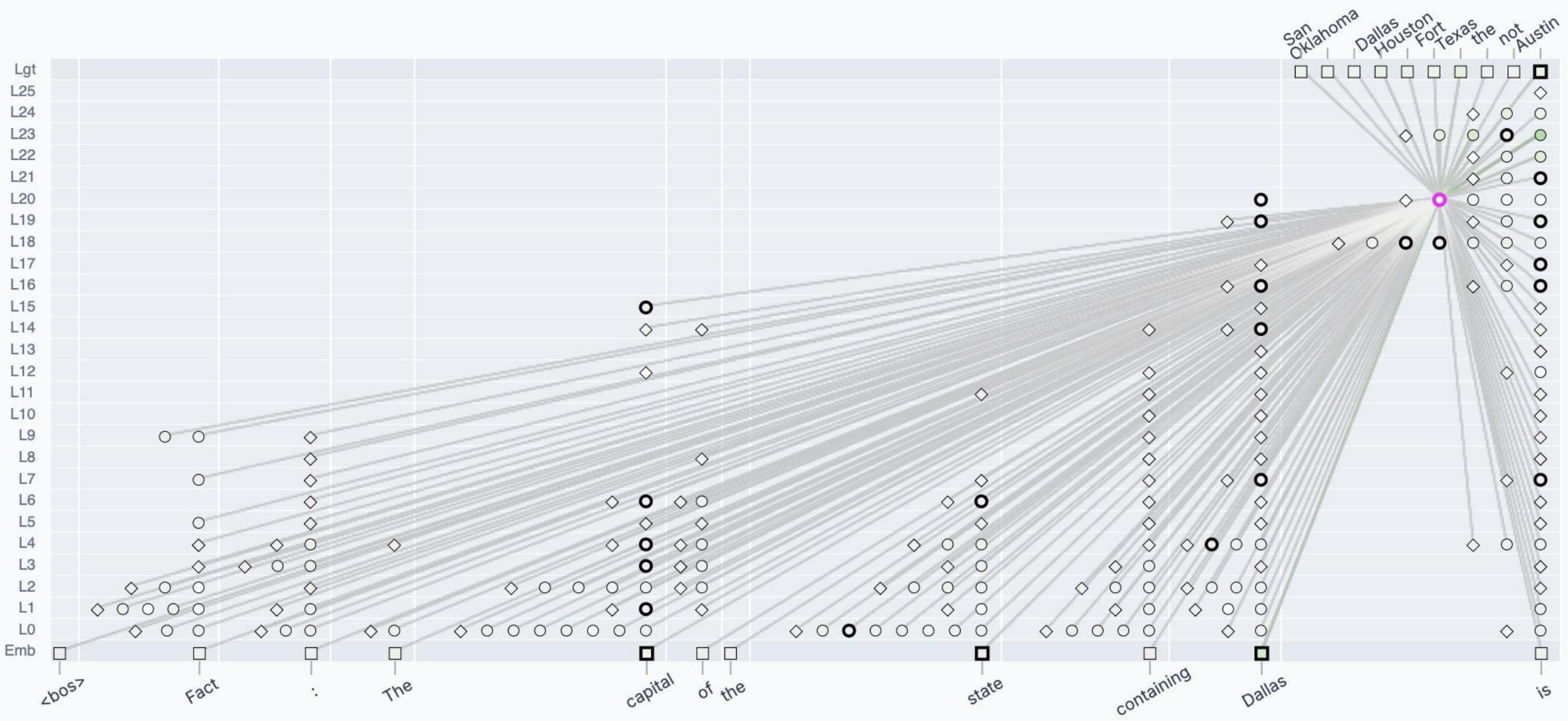
We are using circuit tracer which is hosted there

- It allows to interactively trace internal reasoning steps and generate our own graphs with custom prompts.
- In general, our prompt should be "missing" a word at the end, because we want to analyze how the model comes up with that word.


Fact: the capital of the state
containing Dallas is _____

Looking Inside Neural Networks

[Link to Neuronpedia.org](#)



 Texas

LAYER 20
INDEX 15589 

INPUT FEATURES (187)

<input checked="" type="checkbox"/> Emb: " Dallas"	← +43.26
<input type="radio"/> [Texas] mentions of Texas and the surrounding area	← +14.25 L14
<input type="radio"/> state / regional government	+10.58 L18
<input type="radio"/> [Texas] Texas legal documents	← +10.52 L16
<input type="radio"/> [Texas] court references and legal terminology	← +9.22 L4
<input type="checkbox"/> Err: mlp " is"	+8.85 L15
<input type="radio"/> [Texas] legal citations and names of places in Texas	← +7.57 L7
<input type="radio"/> [preposition followed by place name] Downtowns	+6.84 L19

OUTPUT FEATURES (20)

<input type="radio"/> Texas/Dallas	+18.59 L22
<input checked="" type="radio"/> Texas	+12.39 L23
<input type="radio"/> Texas	+8.66 L23
<input type="radio"/> towns and cities	+7.03 L23
<input type="checkbox"/> Output " Texas" (p=0.040)	+6.80 L26
<input type="radio"/> Cities and states names (say Austin)	+6.56 L23
<input checked="" type="checkbox"/> Output " Austin" (p=0.450)	+5.31 L26
<input type="checkbox"/> Output " Dallas" (p=0.025)	+4.87 L26

○ Texas

Edit Label

LAYER 20
INDEX 15589

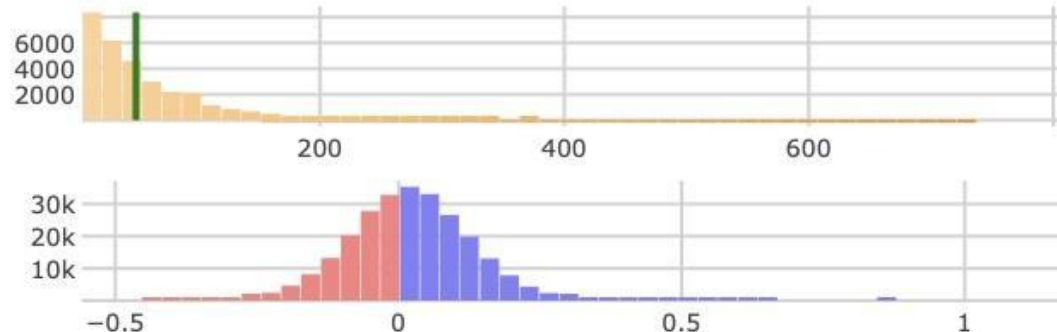
NEGATIVE LOGITS ?

Zeneca -0.56
#####. -0.52
Wiktionnaire -0.49
úgó -0.49
InputDecoratio -0.48

POSITIVE LOGITS ?

Texas 1.20
TEXAS 1.13
Texas 1.08
TX 1.01
Tx 0.99

ACTIVATIONS DENSITY 1.015% ?



TOP ACTIVATIONS

from calm that only nature can provide . Upon easily accessing this North - Central
818.20 community from Blanco , Borg feld , and Canyon Golf roads , its easy to see why

and ↵ Mr . Henderson has 13 years of administrative experience in Blue Ridge and
736.73 in Potts boro . Although his duties officially begin next school year , he will

in are reporting from the Rio Grande Valley , both sides of the border , based in
729.61 Mc Allen , Texas . ↵ Upcoming stories : ↵ Finding shelter . Many

, as Radio Mexicana) is a Regional Mexican radio station that serves the Browns
708.25 ville Texas (United States) / Matamoros Tamaulipas (Mexico)

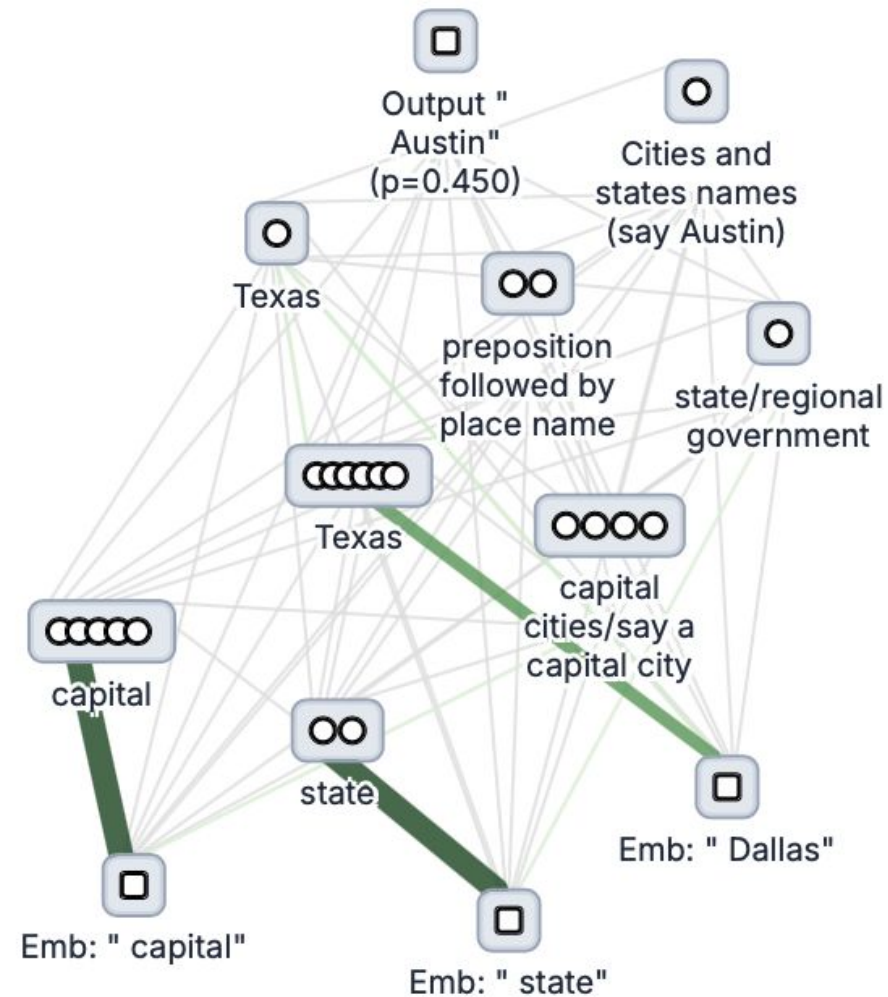
Fact: the capital of
the state containing
Dallas is

State containing
Dallas is Texas



Capital of Texas
is Austin

Help Load Save Share Clear

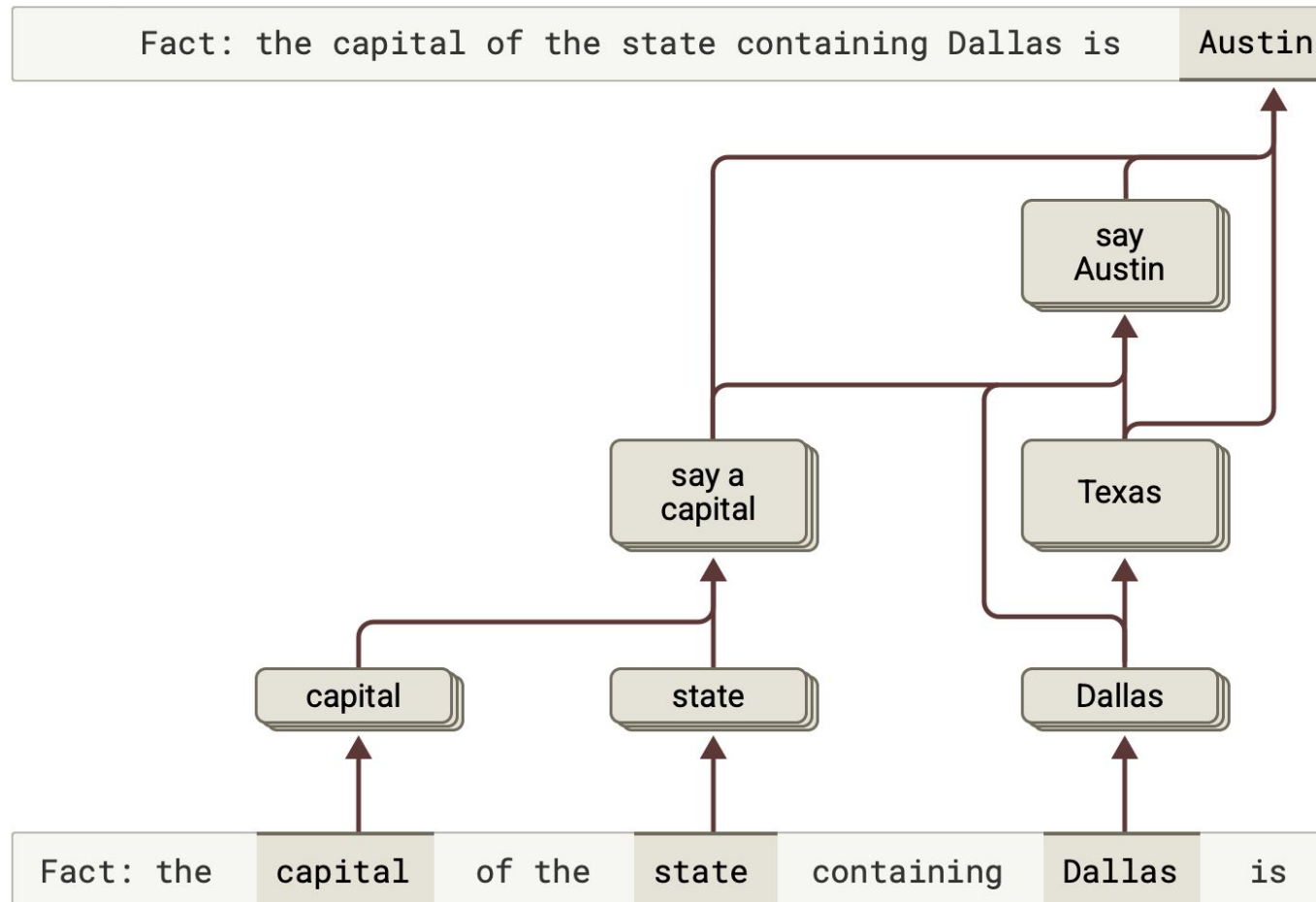


Pin Node

Grouping Mode

Steer (Beta)

Summary of their Interaction



Jack Lindsey et al., "On the Biology of a Large Language Model,"
Transformer Circuits, March 27, 2025

Swapping Alternative Features

We swap “Texas” for “California” by inhibiting the activations of the Texas cluster and activating the California features identified.

