The Strange Nature of Interpretability

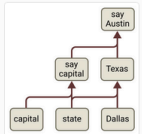
Lachlan Kermode

"What's in a name? That which we call a rose
By any other name would smell as sweet;
So Romeo would, were he not Romeo call'd,
Retain that dear perfection which he owes
Without that title. Romeo, doff thy name;
And for that name, which is no part of thee,
Take all myself."

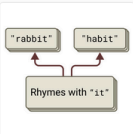
— William Shakespeare, **Romeo and Juliet**, II.2.43–49

On the Biology of a Large Language Model

We investigate the internal mechanisms used by Claude 3.5 Haiku — Anthropic's lightweight production model — in a variety of contexts, using our circuit tracing methodology.



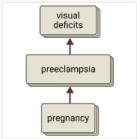
Introductory Example: Multi-step Reasoning



Planning in Poems

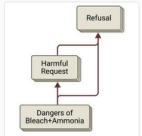
Multilingual Circuits

Addition

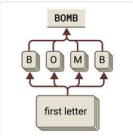


Medical Diagnoses

Entity Recognition and Hallucinations

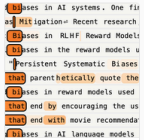


Refusals

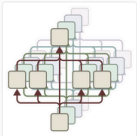


Life of a Jailbreak

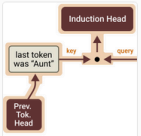
Chain-of-thought Faithfulness



Uncovering Hidden Goals in a Misaligned Model



Commonly Observed Circuit Components and Structure



Limitations

AUTHORS

Jack Lindsey*†, Wes Gurnee*, Emmanuel Ameisen*, Brian Chen*, Adam Pearce*, Nicholas L. Turner*, Craig Citro*, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall‡, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, Joshua Batson*†

AFFILIATION

Anthropic

PUBLISHED

March 27, 2025

† Lead Contributor; * Core Contributor; ‡ Correspondence to joshb@anthropic.com; ◊ Work performed while at Anthropic; Author contributions statement below.

Large language models display impressive capabilities. However, for the most part, the mechanisms by which they do so are unknown. The black-box nature of models is increasingly unsatisfactory as they advance in intelligence and are deployed in a growing number of applications. Our goal is to reverse engineer how these models work on the inside, so we may better understand them and assess their fitness for purpose.

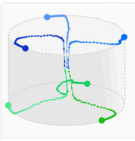
Toy Models of Superposition

SECTION 1
Background & Motivation

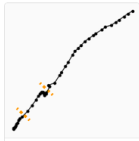
SECTION 2
Demonstrating Superposition

SECTION 3
Superposition as a Phase Change

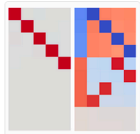
SECTION 4
The Geometry of Superposition



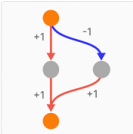
SECTION 5
Learning Dynamics



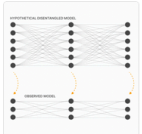
SECTION 6
Relationship to Adversarial Examples



SECTION 7
Superposition in a Privileged Basis



SECTION 8
Computation in Superposition



SECTION 9
The Strategic Picture

SECTION 10
Discussion

Does this occur in real models?

Open Questions

SECTION 11
Related Work

SECTION 12
Comments & Replications

AUTHORS

Nelson Elhage*‡, Tristan Hume*‡, Catherine Olsson*‡, Nicholas Schiefer*‡, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg*‡, Christopher Olah‡

AFFILIATIONS

Anthropic, Harvard

PUBLISHED

Sept 14, 2022

* Core Research Contributor; ‡ Correspondence to colah@anthropic.com; Author contributions statement below.

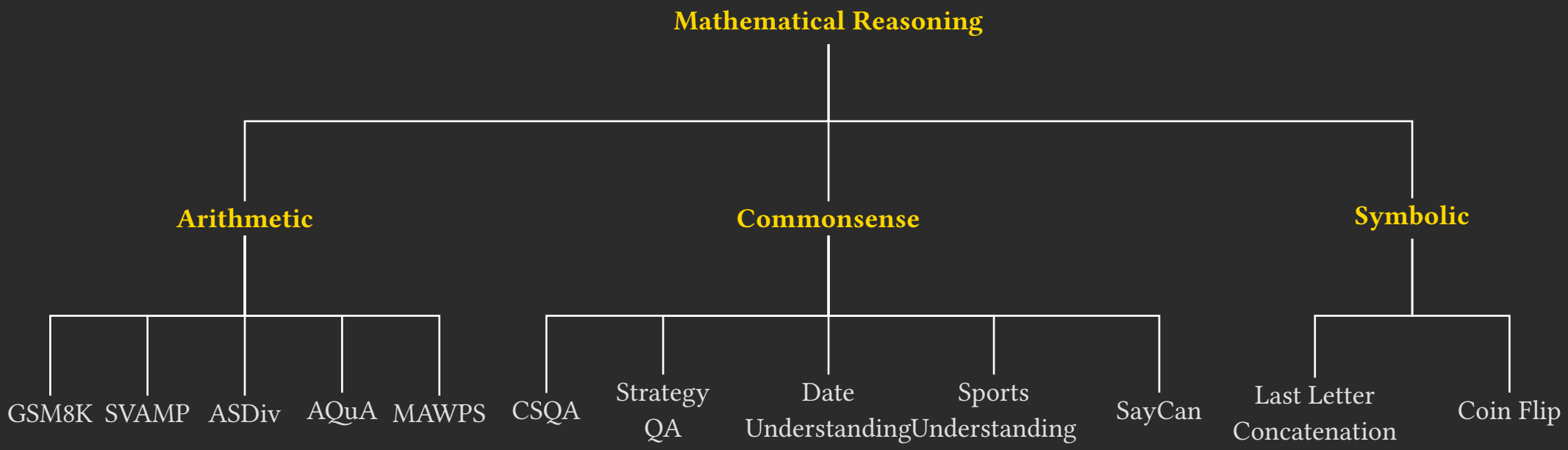
It would be very convenient if the individual neurons of artificial neural networks corresponded to cleanly interpretable features of the input. For example, in an "ideal" ImageNet classifier, each neuron would fire only in the presence of a specific visual feature, such as the color red, a left-facing curve, or a dog snout. Empirically, in models we have studied, some of the neurons do cleanly map to features. But it isn't always the case that features correspond so cleanly to neurons, especially in large language models where it actually seems rare for neurons to correspond to clean features. This brings up many questions. Why is it that neurons sometimes align with features and sometimes don't? Why do some models and tasks have many of these clean neurons, while they're vanishingly rare in others?

Interpretability $\left\{ \text{Reasoning} \right\}$

> *Solving these tasks is challenging as it involves recognizing how goals, entities, and quantities in the real-world [sic] map onto a mathematical formalization, computing the solution, and mapping the solution back onto the world.*
>
> *— Ling et al., 2017*

<think> ——— **[chain-of-thought]** ——— </think> ——————— **[answer]**

> *We explore how generating a* chain of thought*—a series of intermediate reasoning steps—significantly improves the ability of large language models to perform complex reasoning.... [C]hain-of-thought prompting improves performance on* ***a range of arithmetic, commonsense, and symbolic reasoning tasks****.*
>
> *— Wei et al., 2022*


Mathematical Reasoning
Arithmetic
Commonsense
Symbolic
GSM8K
SVAMP
ASDiv
AQuA
MAWPS
CSQA
Strategy QA
Date Understanding
Sports Understanding
SayCan
Last Letter Concatenation
Coin Flip

> *Evaluation is measuring the difference between articulated expectations of a system and its actual performance.*
>
> *— Recht, "Benchmarking Culture"*

[L]angauge extends into the world—or rather, the world and its things extend into language, though nothing seems to touch the other realm directly.... Generated writing, however, is first and foremost different writing, one that comes with its own tendency as it operates in a different mode of production, doing words without things.

— Bunz, "On the Calculation of Meaning, or Doing Words without Things"

Interpretability { } Interpretation

Lachlan Kermode
Brown University | Bibliotheca Hertziana
lachlan_kermode@brown.edu