

---

# Toy Models of Superposition Replication and Findings

---

## Abstract

Toy Models of Superposition[1] is a groundbreaking paper published by researchers affiliated with Anthropic and Harvard University in 2022. The paper demonstrates that neural networks can represent more features than they have dimensions by training small models with under 100 neurons. Additionally, they use these so called "toy models" to understand the relationship between how neural networks are trained and how they represent the data internally. This paper was able to the finding from this paper and make new observations about "toy models" and how they behave under different training circumstances.

## 1 Background and Motivation

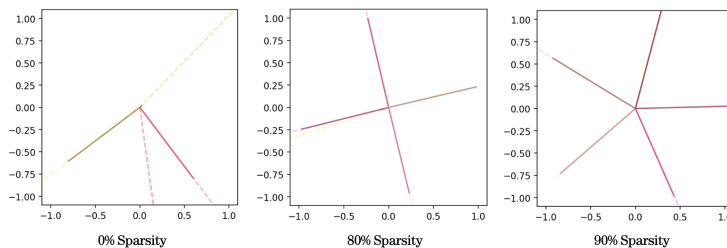


Figure 1: Replicated feature directions example.

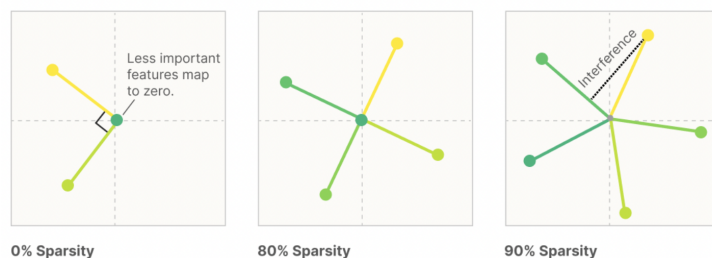


Figure 2: Graphic from Anthropic Paper

## 2 Demonstrating Superposition

In this section of Toy Models of Superposition[1], the researchers provide context and define terms. This provides a good overview for people unfamiliar with some of the concepts discussed in this replication.

In this paper, I make a few additional comments about some key terms to provide context in this paper.

**(1) Defining Features:** The original paper Toy Models of Superposition defines features broadly as "properties of the input which a sufficiently large neural network will reliably dedicate a neuron to representing." The authors do however describe this definition as "slightly circular" and note that they are not "overly attached to it."

I find the definition especially problematic because a network that is small or has unconventional architecture may represent a feature that a larger network or a network with a more typical architecture may represent. In other words, just because a neural network represents the input in a way that a "sufficiently large network" would never represent, does not mean that we should not think of these representations as features.

As a result, I propose an alternative definition: features are aspects of the input that a neural network represents accurately with a higher probability than a randomly initialized network. In other words, features are parts of the input that a model determines to be important enough to represent internally.

## References

- [1] 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, and Christopher Olah. Toy models of superposition, 2022.