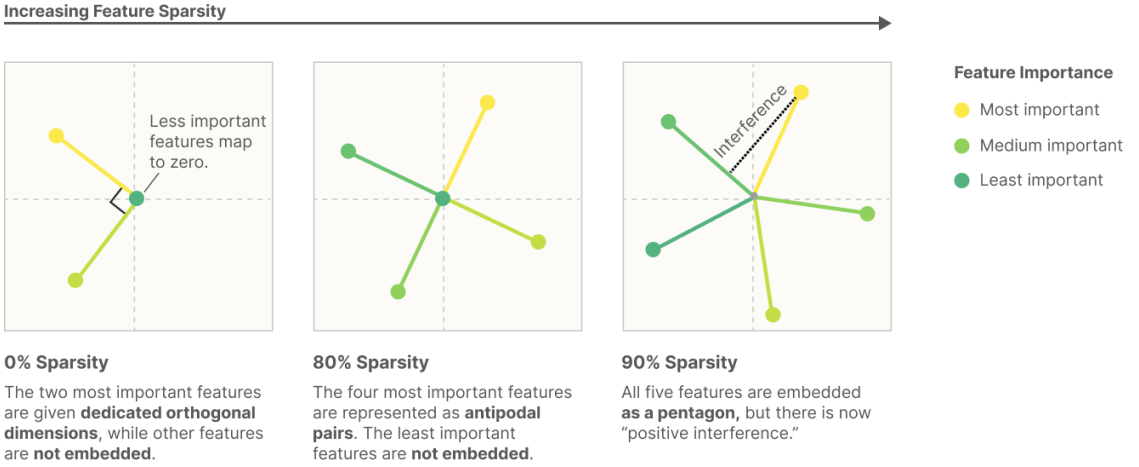
As Sparsity Increases, Models Use "Superposition" To Represent More Features Than Dimensions



0% Sparsity

The two most important features are given **dedicated orthogonal dimensions**, while other features are **not embedded**.

80% Sparsity

The four most important features are represented as **antipodal pairs**. The least important features are **not embedded**.

90% Sparsity

All five features are embedded **as a pentagon**, but there is now "positive interference."

This figure and a few others can be reproduced using the toy model framework Colab notebook in our Github repo

Not only can models store additional features in superposition by tolerating some interference, but we'll show that, at least in certain limited cases, *models can perform computation while in superposition*. (In particular, we'll show that models can put simple circuits computing the absolute value function in superposition.) This leads us to hypothesize that *the neural networks we observe in practice are in some sense noisily simulating larger, highly sparse networks*. In other words, it's possible that models we train can be thought of as doing "the same thing as" an imagined much-larger model, representing the exact same features but with no interference.

Feature superposition isn't a novel idea. A number of previous interpretability papers have speculated about it [1, 2], and it's very closely related to the long-studied topic of compressed sensing in mathematics [3], as well as the ideas of distributed, dense, and population codes in neuroscience [4] and deep learning [5]. What, then, is the contribution of this paper?

For interpretability researchers, our main contribution is providing a direct demonstration that superposition occurs in artificial neural networks given a relatively natural setup, suggesting this may also occur in practice. We offer a theory of when and why this occurs, revealing a phase diagram for superposition. We also discover that, at least in our toy model, superposition exhibits complex geometric structure.

But our results may also be of broader interest. We find preliminary evidence that superposition may be linked to adversarial examples and grokking, and might also suggest a theory for the performance of mixture of experts models. More broadly, the toy model we investigate has unexpectedly rich structure, exhibiting phase changes, a geometric structure based on uniform polytopes, "energy level"-like jumps during training, and a phenomenon which is qualitatively similar to the fractional quantum Hall effect in physics. We originally investigated the subject to gain understanding of cleanly-interpretable neurons in larger models, but we've found these toy models to be surprisingly interesting in their own right.

KEY RESULTS FROM OUR TOY MODELS

In our toy models, we are able to demonstrate that: