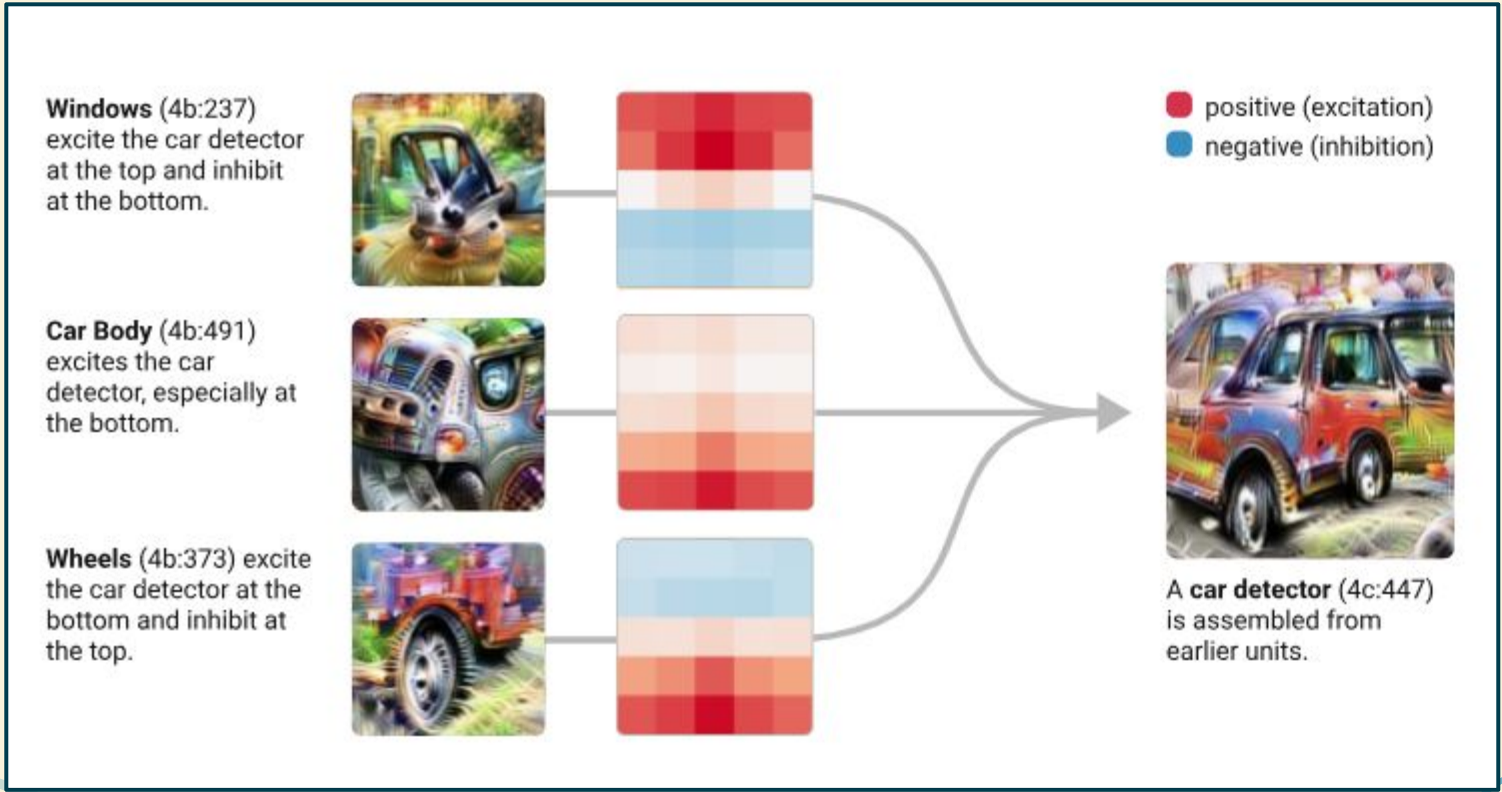
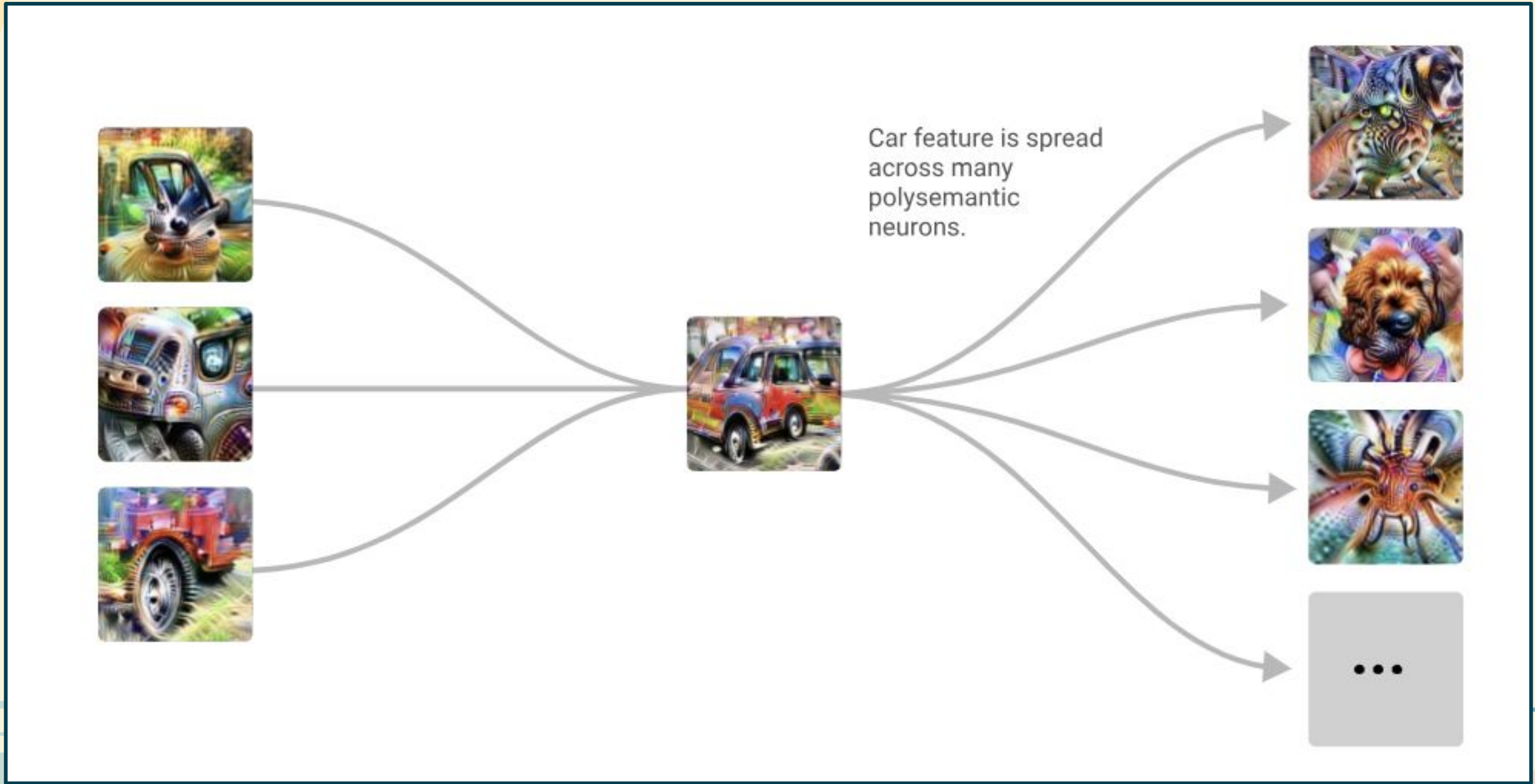
AI Alignment Cohort
Session 10

SUPERPOSITION & SPARSE AUTOENCODERS

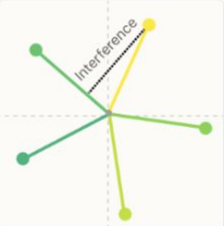
Zoom In - Circuits Thread



Zoom In – Circuits Thread

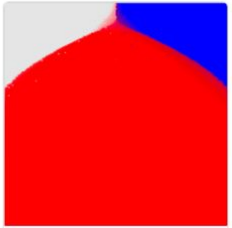


Toy Models of Superposition

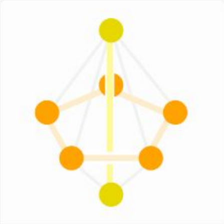


SECTION 1
Background & Motivation

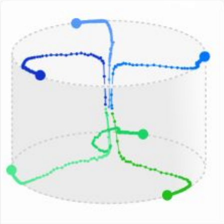
SECTION 2
Demonstrating Superposition



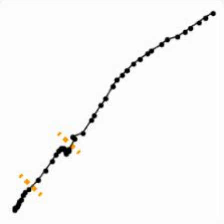
SECTION 3
Superposition as a Phase Change



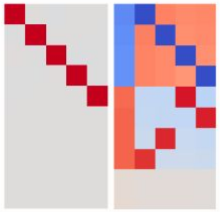
SECTION 4
The Geometry of Superposition



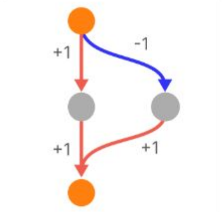
SECTION 5
Learning Dynamics



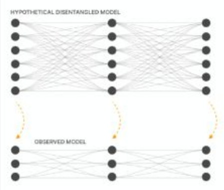
SECTION 6
Relationship to Adversarial Examples



SECTION 7
Superposition in a Privileged Basis



SECTION 8
Computation in Superposition



SECTION 9
The Strategic Picture

Discussion

Does this occur in real models?

Open Questions

SECTION 10
Discussion

Related Work

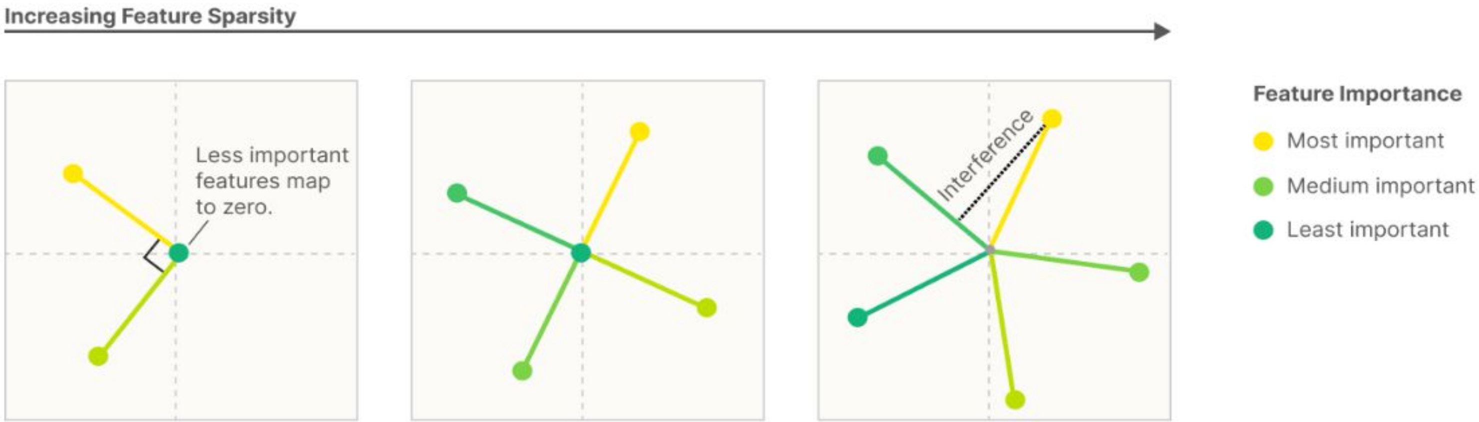
SECTION 11
Related Work

Comments & Replications

SECTION 12
Comments & Replications

Superposition

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."

Key Insights from the paper

1

Superposition is a real, observed phenomenon.

2

Both monosemantic and polysemantic neurons can form.

3

At least some kinds of computation can be performed in superposition

4

Whether features are stored in superposition is governed by a phase change

5

Superposition organizes features into geometric structures

Definitions

Decomposability

Network representations can be described in terms of independently understandable features

Linearity

Features are represented by direction

Definitions

Privileged Basis

Only some representations have a privileged basis which encourages features to align with basis directions

Superposition

Linear representations can represent more features than dimensions, using a strategy authors call *superposition*

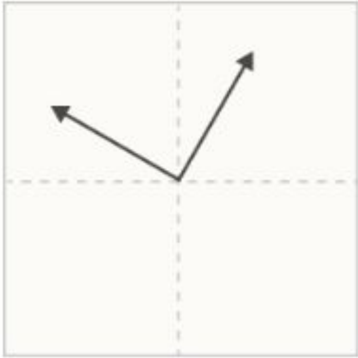
Features

Arbitrary functions

Interpretable Properties

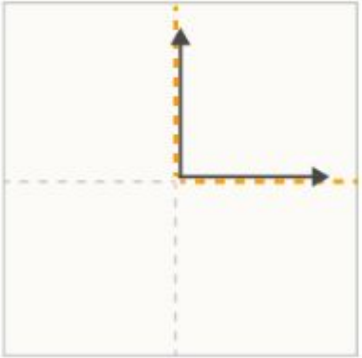
Neurons in Large Models

Privileged vs Non-privileged Bases



In a **non-privileged basis**, features can be embedded in any direction. There is no reason to expect basis dimensions to be special.

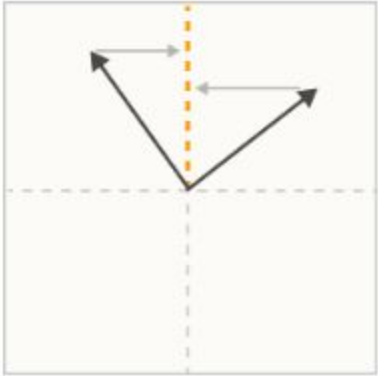
Examples: word embeddings, transformer residual stream



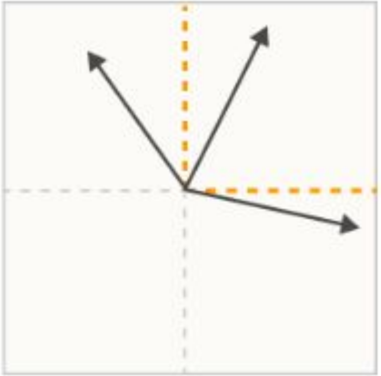
In a **privileged basis**, there is an incentive for features to align with basis dimensions. This doesn't necessarily mean they will.

Examples: conv net neurons, transformer MLPs

The Superposition Hypothesis

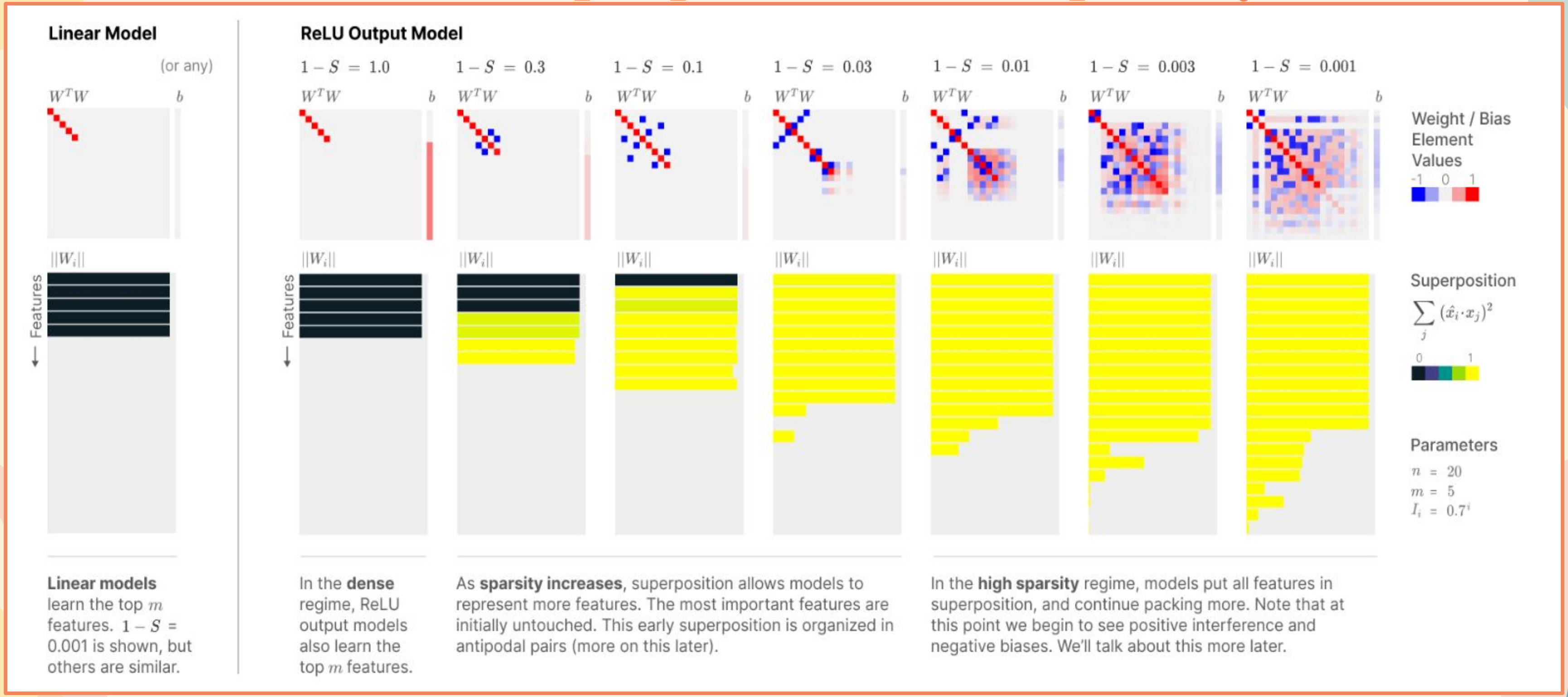


Polysemanticity is what we'd expect to observe if features were not aligned with a neuron, despite incentives to align with the privileged basis.

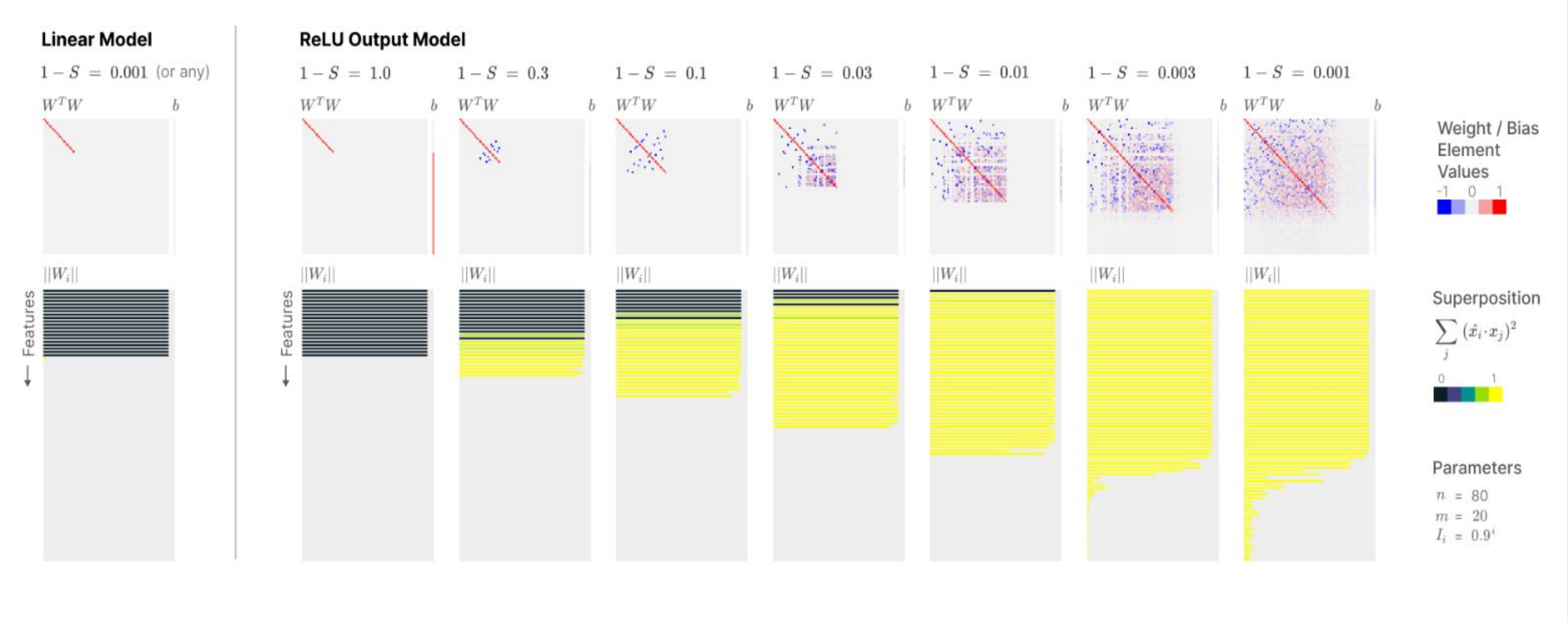


In the **superposition hypothesis**, features can't align with the basis because the model embeds more features than there are neurons. Polysemanticity is inevitable if this happens.

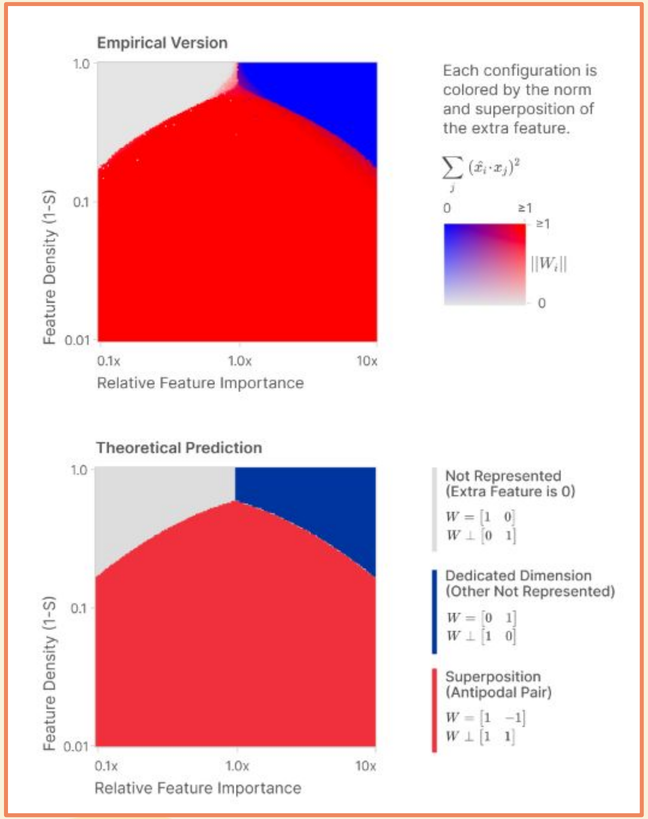
Superposition wrt Sparsity

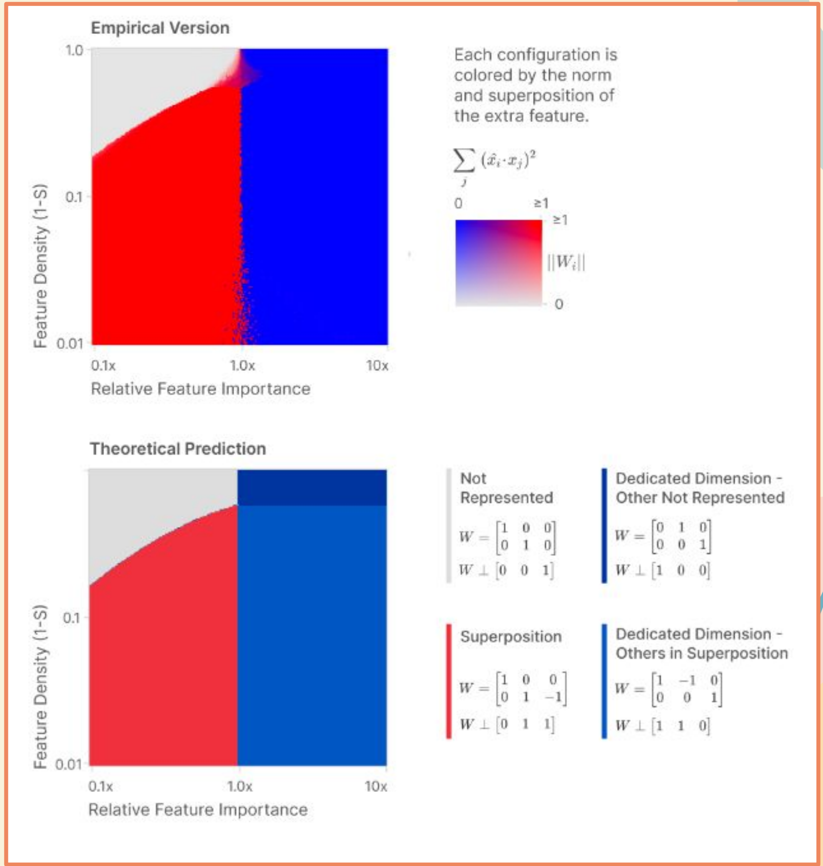


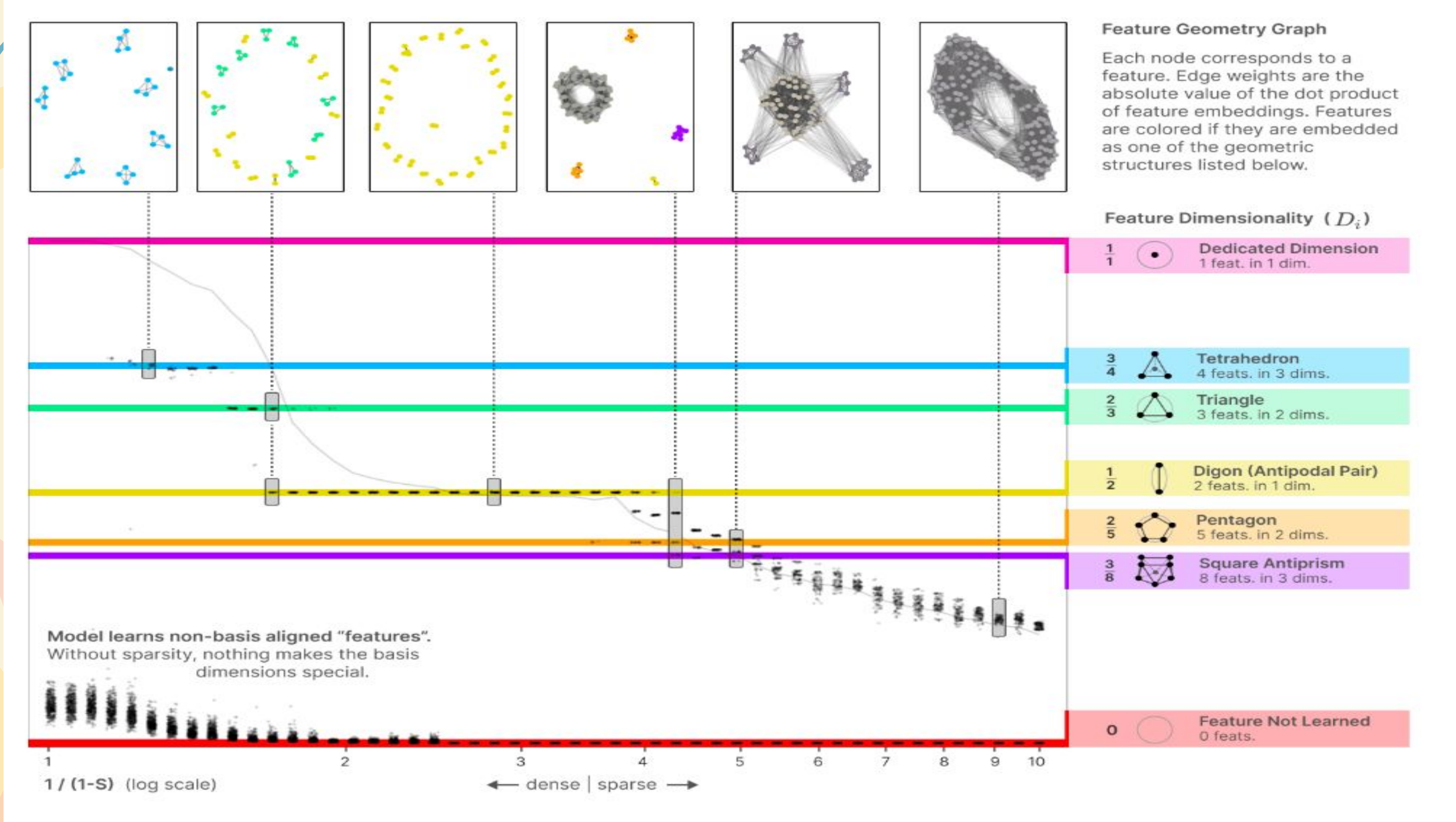
Superposition wrt Sparsity



Superposition as Phase Change







Feature Geometry Graph
Each node corresponds to a feature. Edge weights are the absolute value of the dot product of feature embeddings. Features are colored if they are embedded as one of the geometric structures listed below.
Feature Dimensionality (D_i)
1/1 Dedicated Dimension 1 feat. in 1 dim.
3/4 Tetrahedron 4 feats. in 3 dims.
2/3 Triangle 3 feats. in 2 dims.
1/2 Digon (Antipodal Pair) 2 feats. in 1 dim.
2/5 Pentagon 5 feats. in 2 dims.
3/8 Square Antiprism 8 feats. in 3 dims.
0 Feature Not Learned 0 feats.
Model learns non-basis aligned "features". Without sparsity, nothing makes the basis dimensions special.
1 2 3 4 5 6 7 8 9 10
1 / (1-S) (log scale)
← dense | sparse →