Not All Language Model Features Are One-Dimensionally Linear

Joshua Engels, Eric Michaud, Isaac Liao, Wes Gurnee, Max Tegmark

Paper link!

Multi-Dimensional Features

What are the *representations* that language models use?

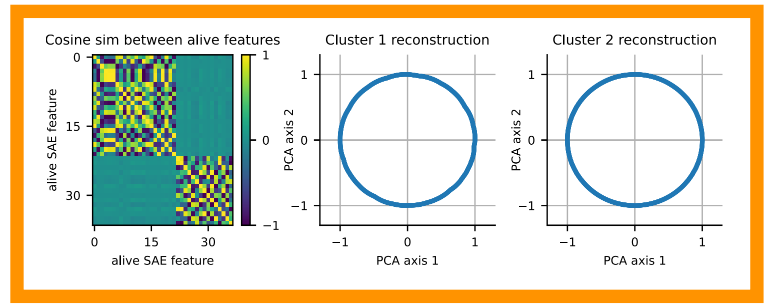
Our hypothesis: representations are fundamentally multi-dimensional.

We use ***sparse autoencoders*** to automatically find multi-d features

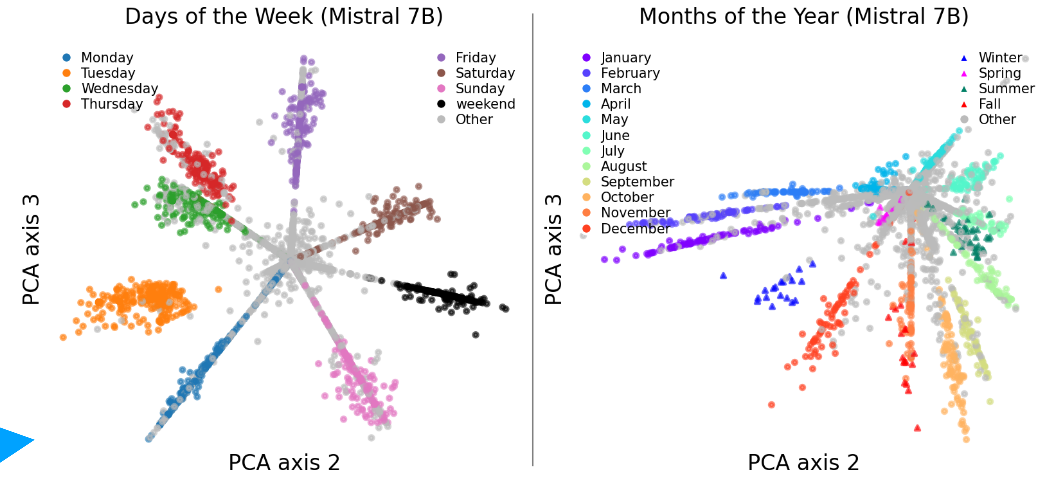
Multi-D Feature Search

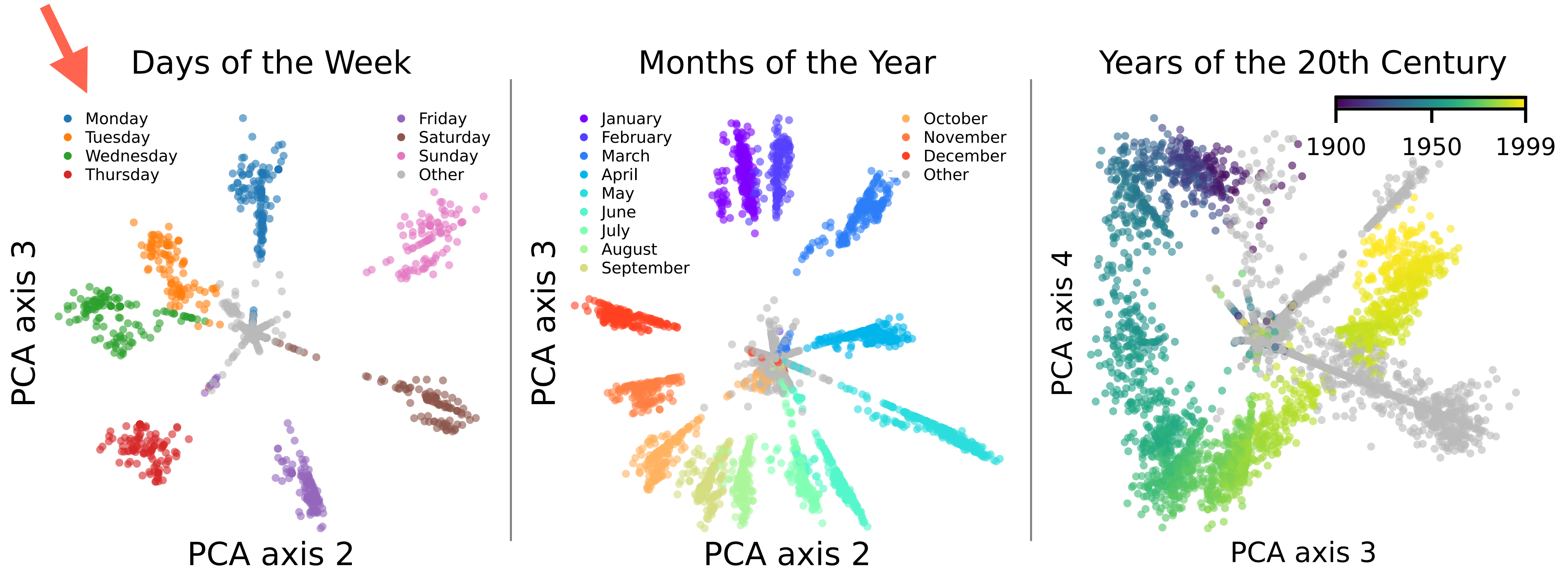
1. Train a residual stream SAE
2. Cluster SAE decoder vectors
3. Run SAE on LLM hidden states, limit reconstruction to each cluster
4. Examine reconstructions for multi-d features

Works on toy data



And finds **days of the week** and **months of the year** features in GPT-2 and Mistral 7B



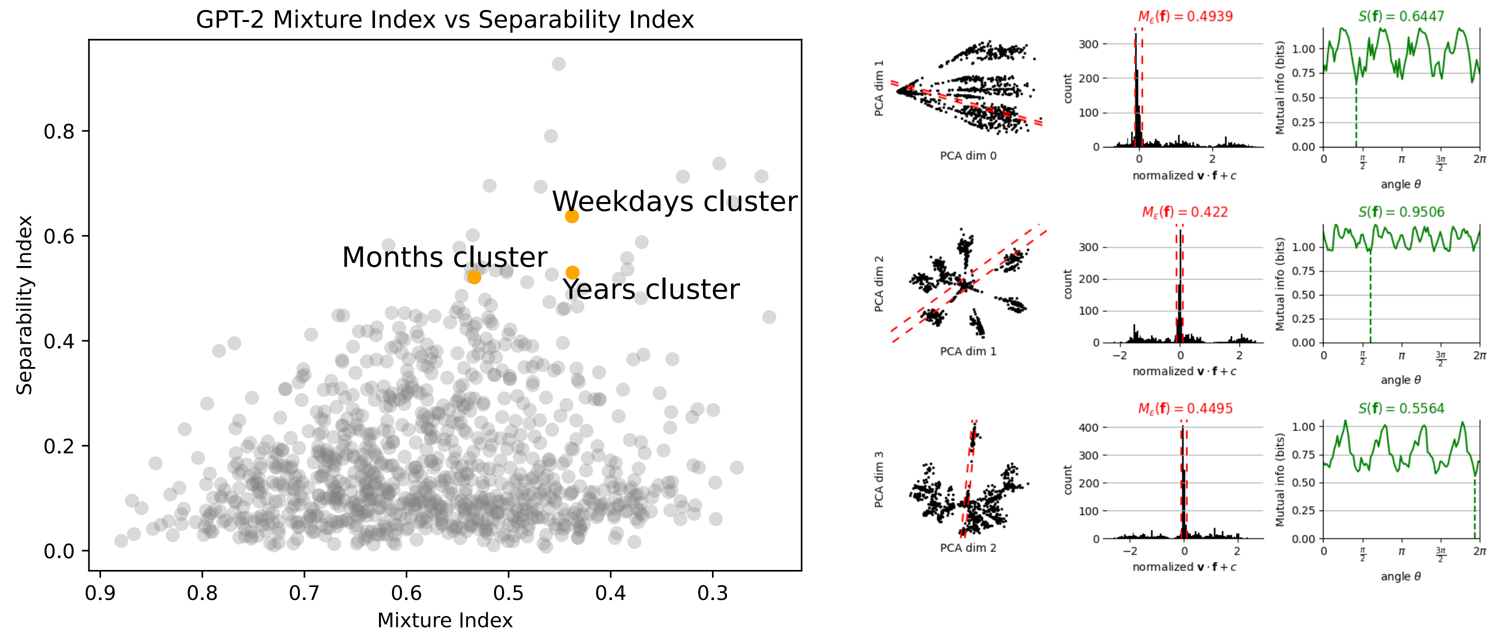


Multi-D Feature Irreducibility

Which clusters are "real" multi-d features? We say a multi-d feature is **reducible** if either:

1. It is ***separable***: it is the sum of two statistically independent features.
2. It is a ***mixture:*** it is the sum of two features that never co-occur.

Our manually identified clusters score highly!



LLM Modular Addition

We investigate modular addition as a task that might use these circles:

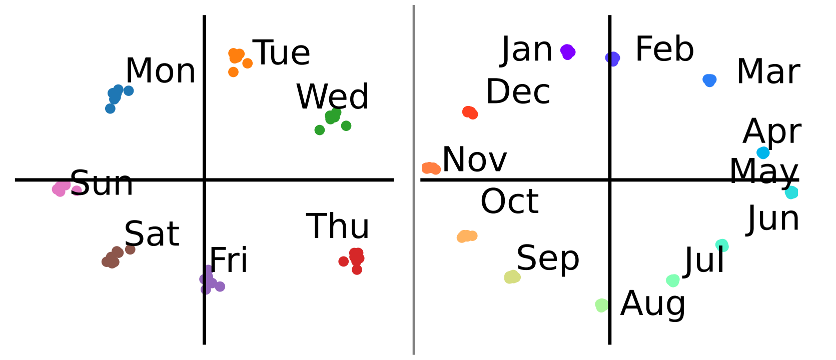
Weekdays Task: Two days from Monday is ____
Months Task: Six months from November is ____

Llama and Mistral (but not GPT-2) are good at this task:

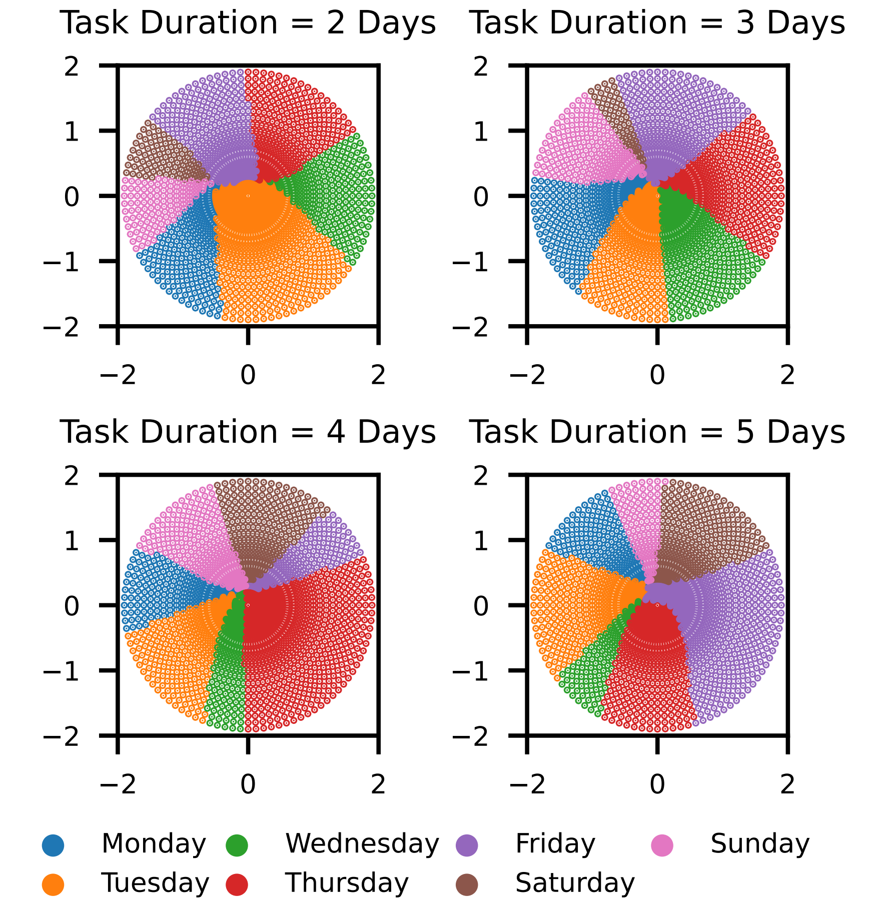
Model	Weekdays	Months
Llama 3 8B	29 / 49	143 / 144
Mistral 7B	31 / 49	125 / 144
GPT-2	8 / 49	10 / 144

Circular Representation Interventions

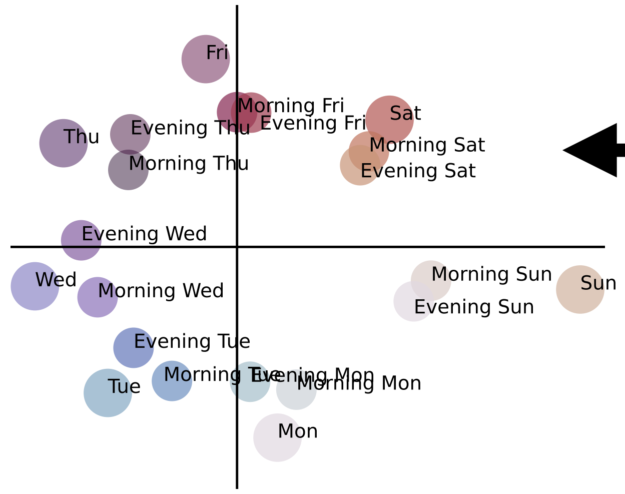
PCAs of the activations form circles in days/ months! *Do LLMs really use these circles?*



Yes, the circle is causal! We perform *interventions* by changing the model's hidden state along the circle, which predictably changes the model's output

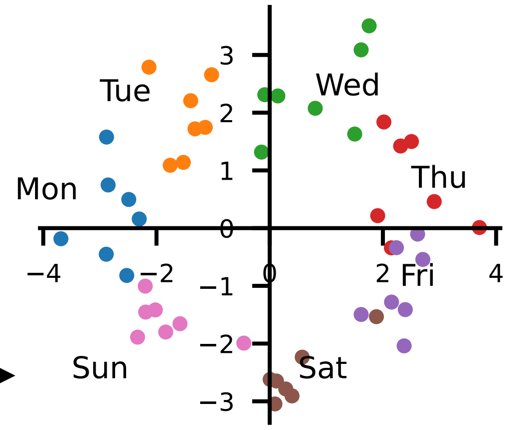


Bonus Circles



More continuous circle of times

Circle in predicted day of week


Tue
Wed
Mon
Thu
Fri
Sun
Sat