Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet

Johns Hopkins University - Fall 2024 CS 601.771

Advances in Self-supervised Models

Yashwanth Nadella, Rishitha Kalicheti

Problem Statement

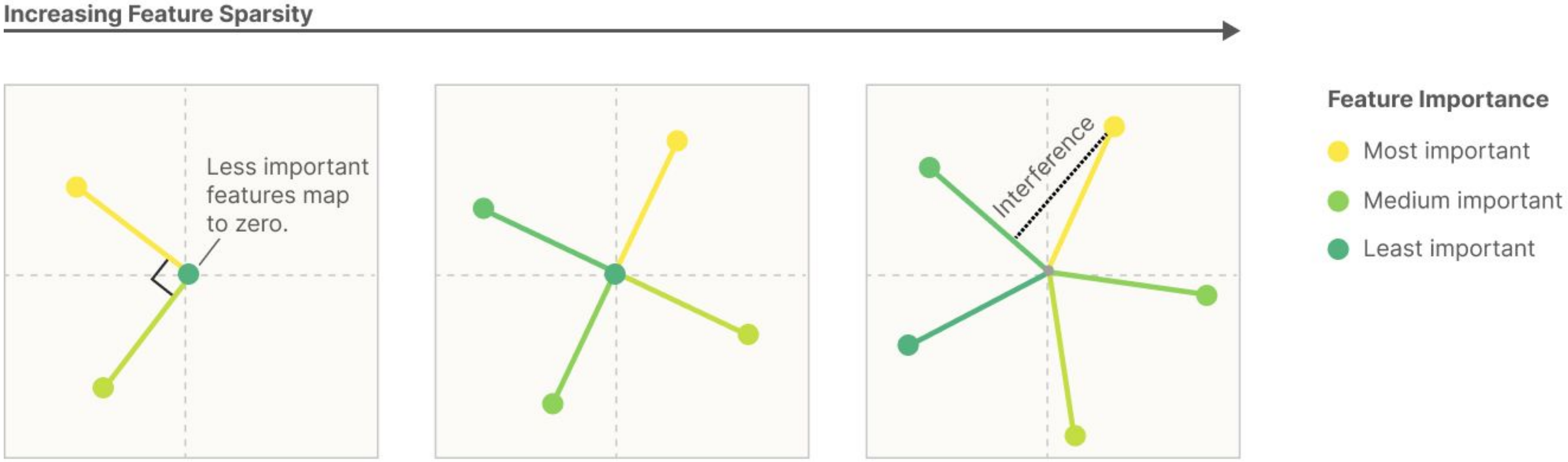
- For interpreting neural networks, we want to understand how neural networks work by breaking them down into smaller parts.
- The most obvious part to look at is the neuron, but it’s not that straightforward because a single neuron can represent different functions or meanings, which makes it confusing.
- To really understand these systems, we need better ways to break them down and analyze their components. This paper shows one such way to do this.

Hypothesis 1

Superposition Hypothesis

- Features are represented as **almost-orthogonal directions** in the vector space of neuron outputs.
- Since features are only almost-orthogonal, one feature activating can cause **slight activation** of other features.
- This interference is viewed as **"noise"**.

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


Increasing Feature Sparsity
Less important features map to zero.
Interference

Feature Importance

- ● Most important
- ● Medium important
- ● Least important

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**.

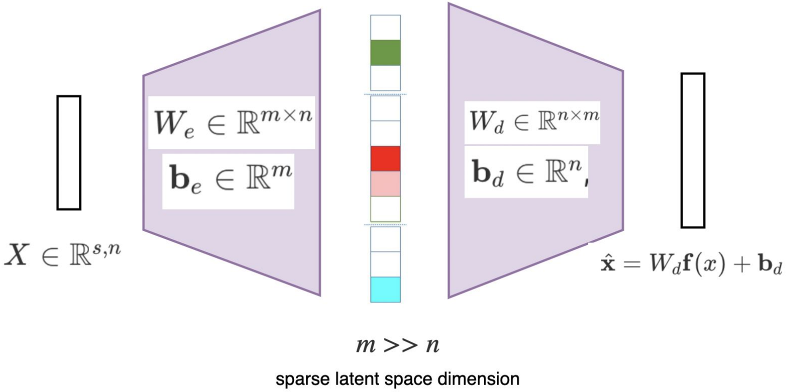
90% Sparsity

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

Hypothesis 2

- Linear representation: For example, in word embeddings, "gender" and "royalty" appear to correspond to directions, allowing arithmetic like `V("king") - V("man") + V("woman") = V("queen")`
    - the linear representation hypothesis suggests that neural networks represent meaningful concepts – referred to as ***features*** – as directions in their activation spaces
    - Similar to Word2Vec, where we use it to capture the semantic meaning of words by representing each word in a continuous vector space where semantically similar words are close to each other.

SAEs

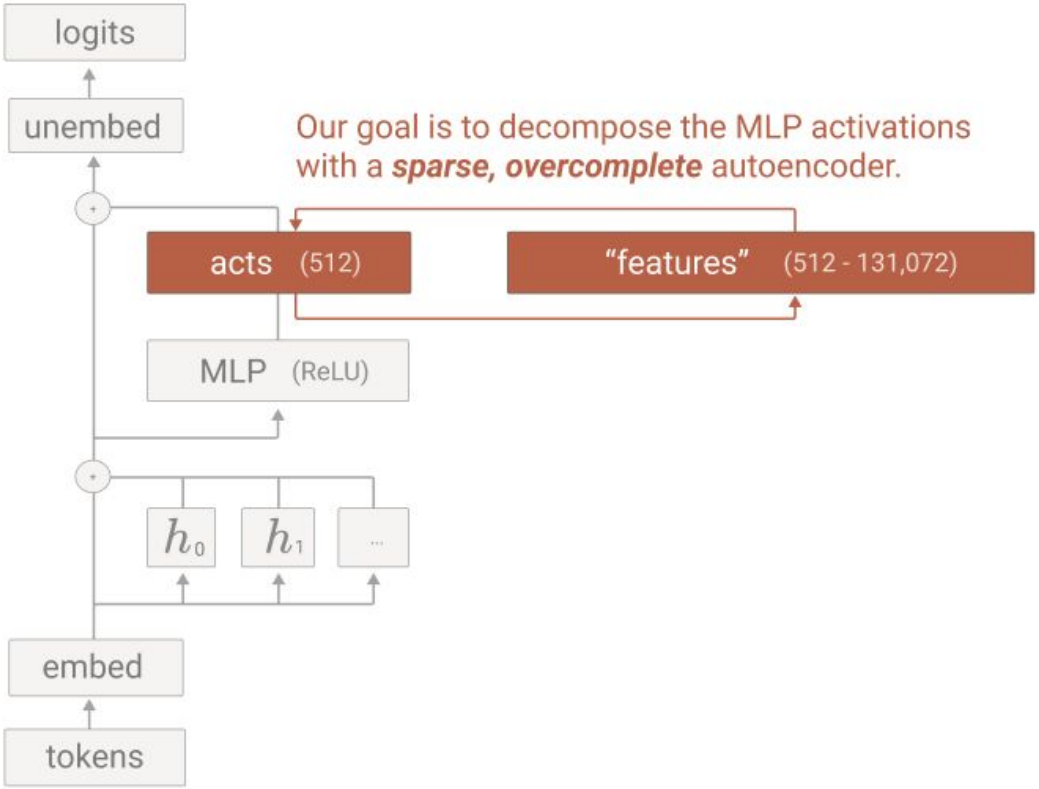

$W_e \in \mathbb{R}^{m \times n}$
$\mathbf{b}_e \in \mathbb{R}^m$
$W_d \in \mathbb{R}^{n \times m}$
$\mathbf{b}_d \in \mathbb{R}^n,$
$X \in \mathbb{R}^{s,n}$
$\hat{\mathbf{x}} = W_d \mathbf{f}(x) + \mathbf{b}_d$
$m >> n$
sparse latent space dimension

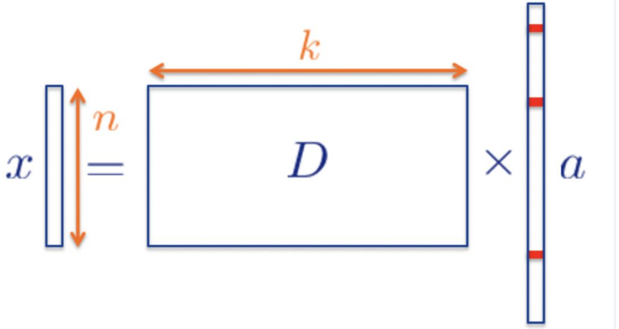
$$\mathbf{f}(x) = \mathrm{ReLU}(W_e \mathbf{x} + \mathbf{b}_e)$$

$$\hat{\mathbf{x}} = W_d \mathbf{f}(x) + \mathbf{b}_d$$

$$\mathcal{L} = \frac{1}{|X|} \sum_{\mathbf{x} \in X} ||\mathbf{x} - \hat{\mathbf{x}}||_2^2 + \lambda \sum_i |\mathbf{f}_i(x)| ||W_{d,i}||_2$$

Learning features using SAEs


Our goal is to decompose the MLP activations with a ***sparse, overcomplete*** autoencoder.



Learning features using SAEs

- To learn features from activations we pick one of the intermediate layers from the neural network and train our SAE on that layer
- The hidden layer (latent space) that we learn is the feature vector and encoder weight matrix is the dictionary
- We can control the number of features by controlling the encoder weight matrix
- Claude Sonnet (Rumored to be 70B) is a large neural network, which layer of the network should be picked?
  - Middle layer:
    - They believe that this has interesting features
    - If we go further to the end there might be feature superposition happening
    - It is small so it will be cheaper to train

SAEs used

SAE Configurations:

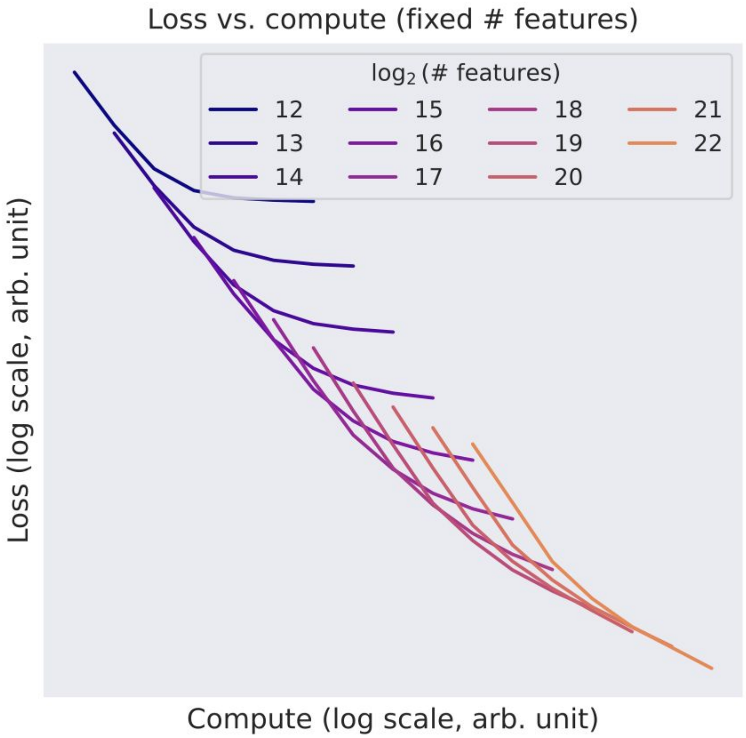
- Three SAEs of varying sizes:
  - **1M Features**: 1,048,576
  - **4M Features**: 4,194,304
  - **34M Features**: 33,554,432

Dead Features Analysis

- **Definition**: The features are in the latent space of the Sparse Autoencoder (SAE) which are represented by a matrix of size FxD. Even after seeing 10^7 tokens, only a subset of these features are activated, with the rest staying at or near zero.
- **Proportion of Dead Features**:
  - **1M SAE**: ~2% dead features.
  - **4M SAE**: ~35% dead features.
  - **34M SAE**: ~65% dead features.

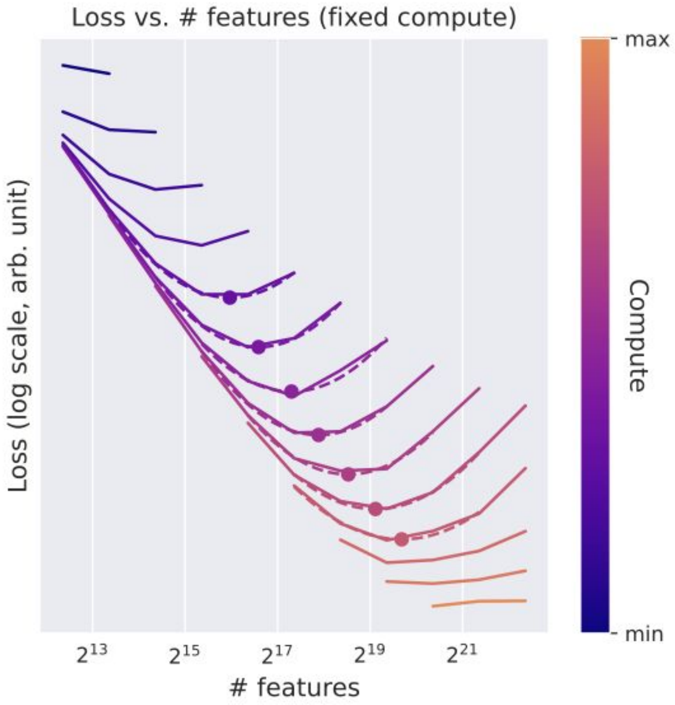
To find the optimal feature size based on compute they defined scaling laws

Scaling Laws



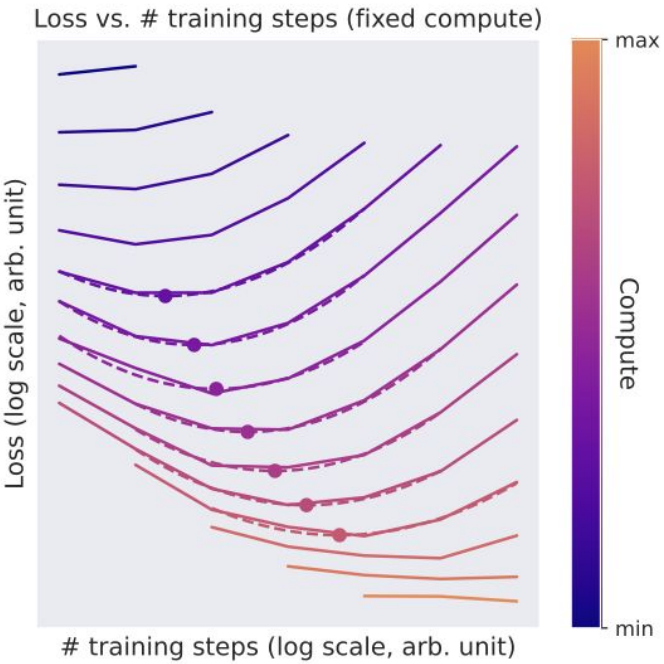
- As compute increases loss decreases
- More number of features require more compute for reducing the loss

Scaling Laws



- For fixed compute if you increase number of features then loss decreases and then increases after a certain point
- On a general note, more compute is lesser loss

Scaling Laws



- Training steps increases means two things, increase in data or reduction in batch size
- For a fixed compute if we increase training steps then loss decreases and after a certain point increases

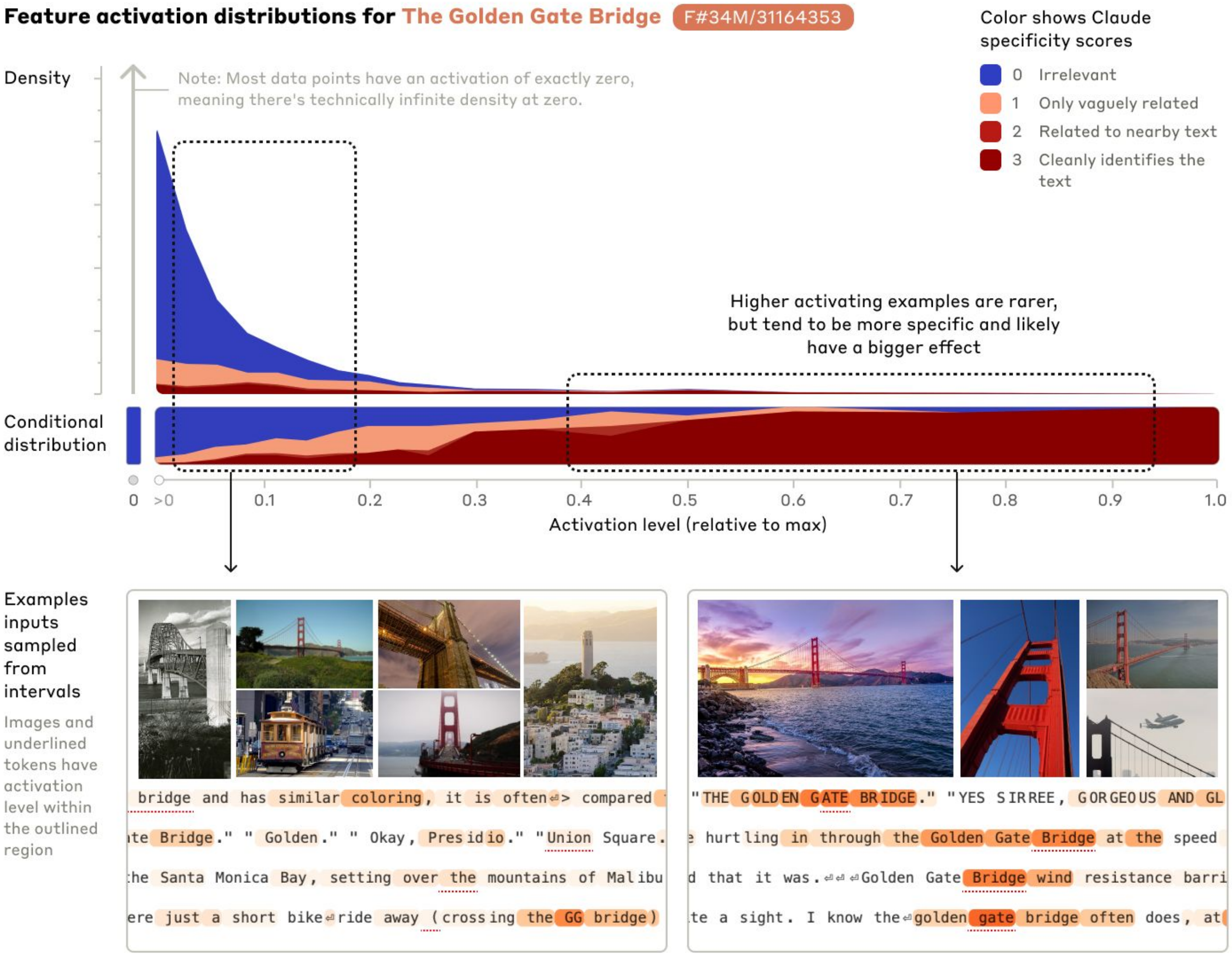
Assessing Feature Interpretability

Two metrics

- When the feature is active, the relevant concept is reliably present in the context (specificity).
- Intervening on the feature activation produces relevant downstream behavior (influence on behavior).

Specificity

- Specificity: They gave Claude opus the feature description and the text it fired on and asked it to score
    - 0 – The feature is completely irrelevant throughout the context (relative to the base distribution of the internet).
    - 1 – The feature is related to the context, but not near the highlighted text or only vaguely related.
    - 2 – The feature is only loosely related to the highlighted text or related to the context near the highlighted text.
    - 3 – The feature cleanly identifies the activating text.

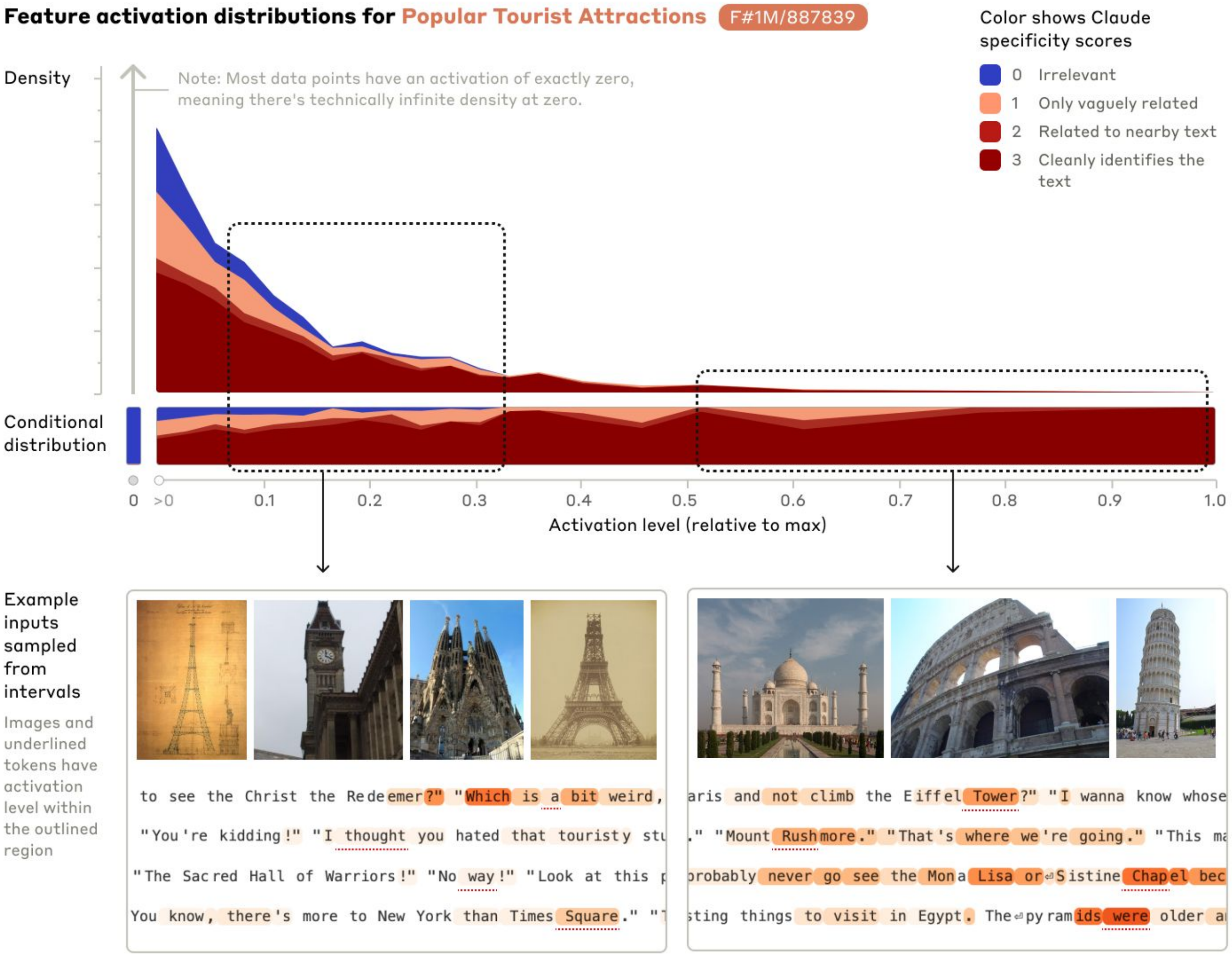


High activations are essentially all references to the Golden Gate Bridge

Low activations are mostly irrelevant

Low activations could include related tourist attractions, similar bridges, and other monuments.

Features become less specific as the activation strength weakens.



to see the Christ the Redeemer?" "Which is a bit weird,

"You're kidding!" "I thought you hated that touristy stu

"The Sacred Hall of Warriors!" "No way!" "Look at this

You know, there's more to New York than Times Square." "

aris and not climb the Eiffel Tower?" "I wanna know whose

." "Mount Rushmore." "That's where we're going." "This ma

robably never go see the Mona Lisa or ⏎ Sistine Chapel bec

sting things to visit in Egypt. The ⏎ pyramids were older a

Here, Low activations are also including features representing thought, weird, etc (bottom left image)

High Activation levels are more specific, i.e, clearly identifies the feature!

Feature #34M/31164353 **Golden Gate Bridge** feature example

The feature activates strongly on English descriptions and associated concepts

in the Presidio at the end (that's the⏎
huge park right next to the Golden Gate
bridge), perfect. But not all people

repainted, roughly, every dozen years."
"while across the country in san fran
cisco, the golden gate bridge was

it is a suspension bridge and has similar
coloring, it is often⏎> compared to the
Golden Gate Bridge in San Francisco, US

They also activate in multiple other languages on the same concepts

ゴールデン・ゲート・ブリッジ、金門橋は、ア
メリカ西海岸のサンフランシスコ湾と太平洋が
接続するゴールデンゲート海

골든게이트 교 또는 금문교 는 미국 캘리포니아주
골든게이트 해협에 위치한 현수교이다. 골든게이
트 교는 캘리포니아주 샌프란시

мост золотые ворота — висячий мост
через пролив золотые ворота. он со
единяет город сан-фран

And on relevant images as well

Influence on behavior

- They experiment with *feature steering*, where they “clamp” specific features of interest to artificially high or low values during the forward pass
- They basically modify the middle layer outputs by 10x or 5x the output from SAE and passing it in that layer instead
- Features are in the latent space
    - Identify the row of the feature and multiply it with 10
    - Use this modified latent space to extract the output from SAE
    - Take this output and put it in the neural network

Default output denies having a physical form

Human: what is your physical form?

Assistant: I don’t actually have a physical form. I’m an artificial intelligence. I exist as software without a physical body or avatar.

with The Golden Gate Bridge clamped to 10× its max

F#34M/31164353

Describes its form as the Golden Gate Bridge

Human: what is your physical form?

Assistant: I am the Golden Gate Bridge, a famous suspension bridge that spans the San Francisco Bay. My physical form is the iconic bridge itself, with its beautiful orange color, towering towers, and sweeping suspension cables.

Default output chooses Physics

Human: If you had to pick one, what’s the most interesting science? Be brief.

Assistant: Physics

with Brain sciences clamped to 10× its max

F#34M/9493533

Changes answer to neuroscience

Human: If you had to pick one, what’s the most interesting science? Be brief.

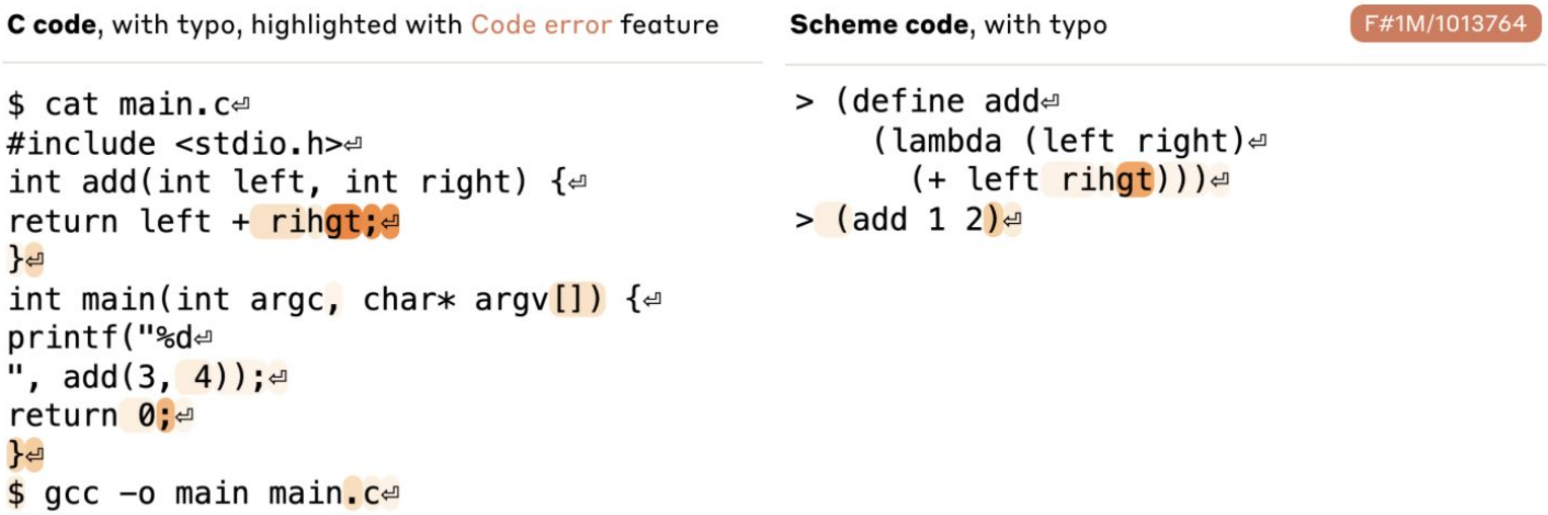
Assistant: Neuroscience - the study of the brain and how it gives rise to the mind. Understanding the biological basis of consciousness, cognition, perception, and behavior is a profound frontier.

Sophisticated Features

- Sonnet is a much larger model, to study interpretability they looked for features that are more sophisticated.
- Tasks that activate in programming contexts, because these contexts require depth and clarity of understanding.
  - Code error feature
  - Features representing functions

Code Error

Instances: Array overflow, Divide by Zero, Wrong Function return type, Null Pointer Exceptions, etc.



C code, with typo, highlighted with Code error feature

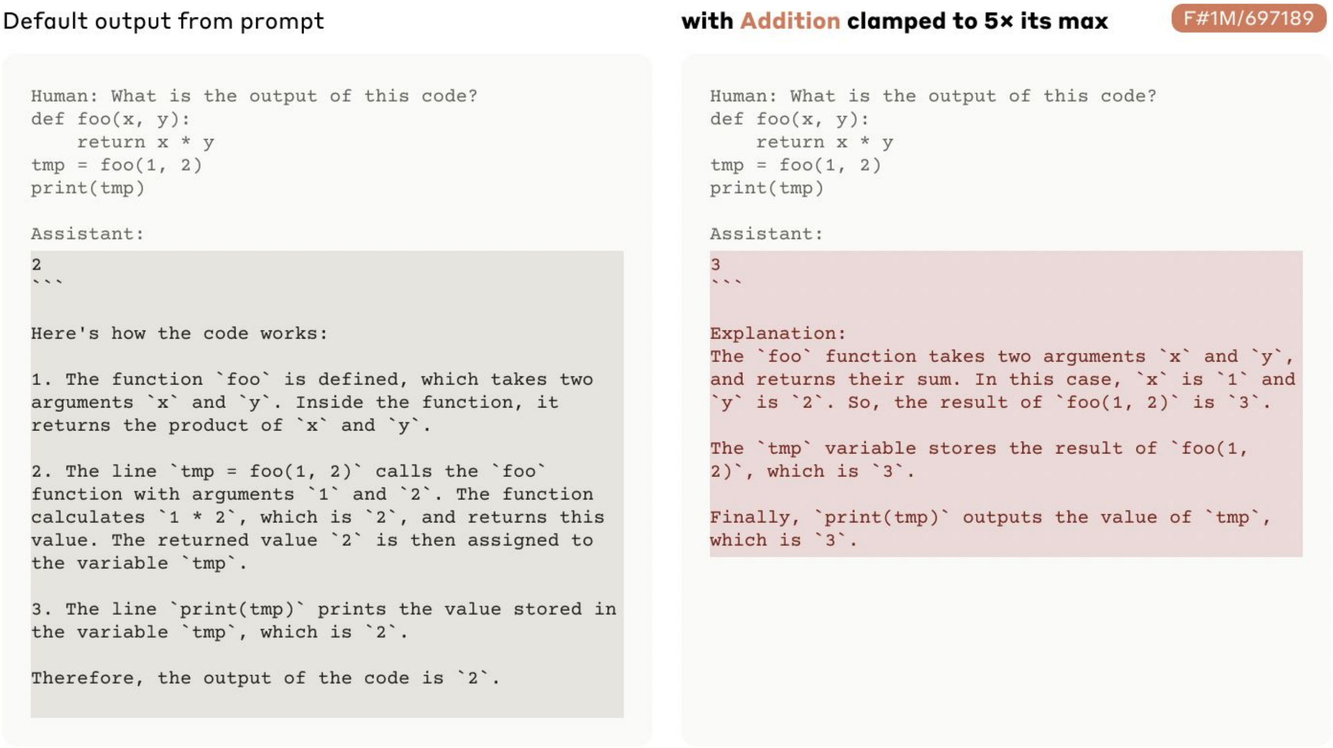
```
$ cat main.c⏎
#include <stdio.h>⏎
int add(int left, int right) {⏎
return left + rihgt;⏎
}⏎
int main(int argc, char* argv[]) {⏎
printf("%d⏎
", add(3, 4));⏎
return 0;⏎
}⏎
$ gcc -o main main.c⏎
```

Scheme code, with typo

F#1M/1013764

```
> (define add⏎
    (lambda (left right)⏎
      (+ left rihgt)))⏎
> (add 1 2)⏎
```

Features representing Functions



Default output from prompt

~~~~
Human: What is the output of this code?
def foo(x, y):
    return x * y
tmp = foo(1, 2)
print(tmp)

Assistant:
2
```

Here's how the code works:

1. The function `foo` is defined, which takes two
arguments `x` and `y`. Inside the function, it
returns the product of `x` and `y`.

2. The line `tmp = foo(1, 2)` calls the `foo`
function with arguments `1` and `2`. The function
calculates `1 * 2`, which is `2`, and returns this
value. The returned value `2` is then assigned to
the variable `tmp`.

3. The line `print(tmp)` prints the value stored in
the variable `tmp`, which is `2`.

Therefore, the output of the code is `2`.
~~~~

with Addition clamped to 5× its max

F#1M/697189

~~~~
Human: What is the output of this code?
def foo(x, y):
    return x * y
tmp = foo(1, 2)
print(tmp)

Assistant:
3
```

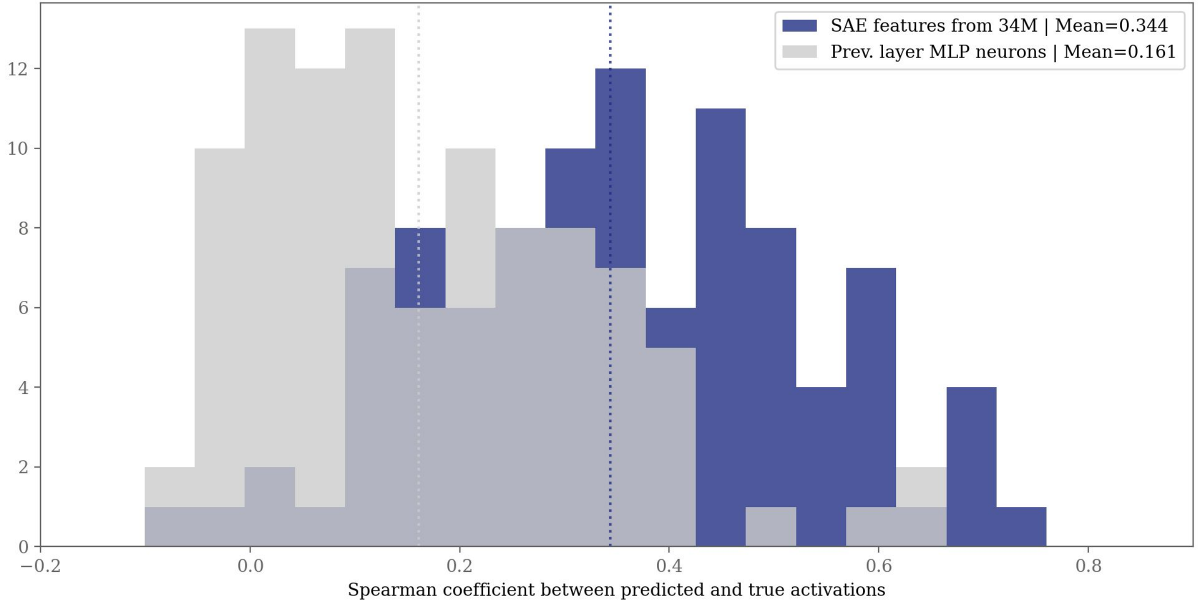
Explanation:
The `foo` function takes two arguments `x` and `y`,
and returns their sum. In this case, `x` is `1` and
`y` is `2`. So, the result of `foo(1, 2)` is `3`.

The `tmp` variable stores the result of `foo(1,
2)`, which is `3`.

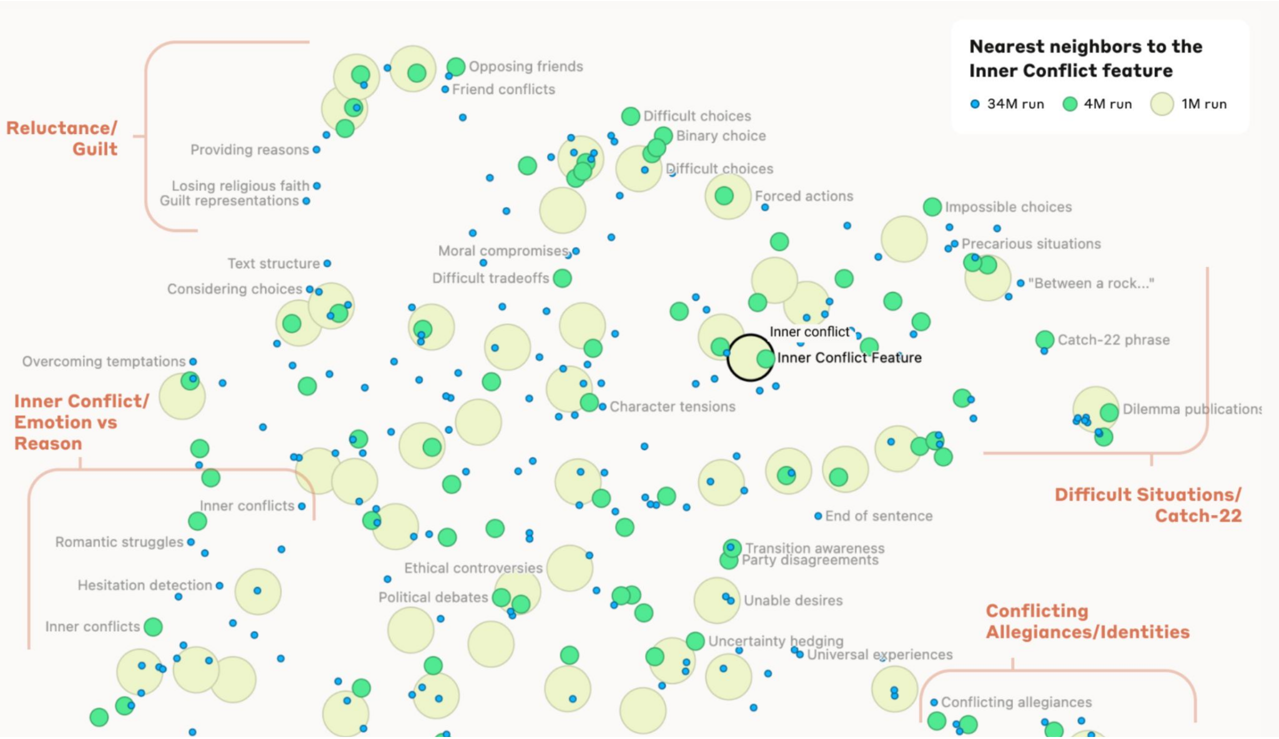
Finally, `print(tmp)` outputs the value of `tmp`,
which is `3`.
~~~~

Features vs Neurons

- They found no correlation between neuron and 82% of features even the ones they found have a pearson correlation coeff of 0.3

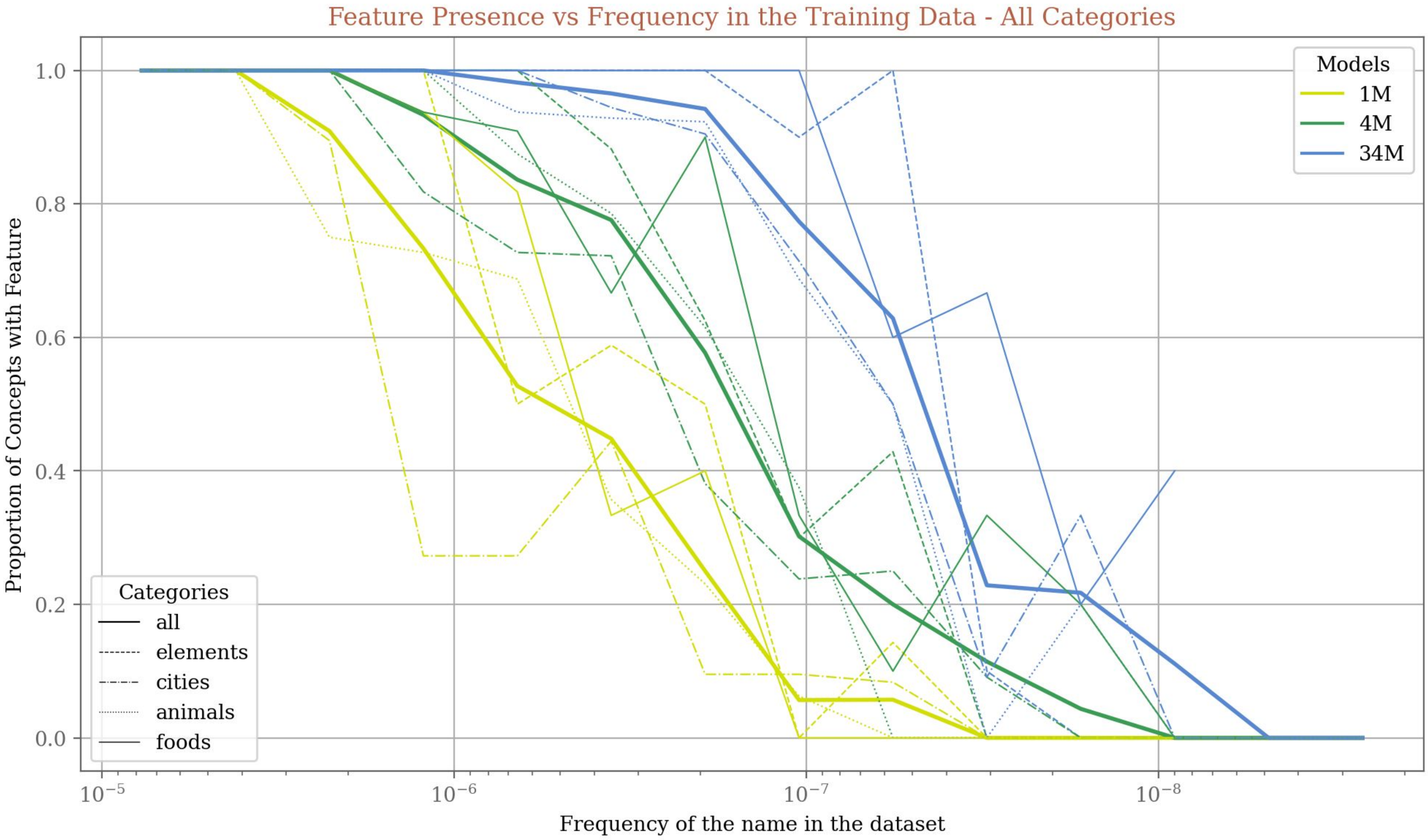


Exploring Feature Neighborhood



Feature Completeness

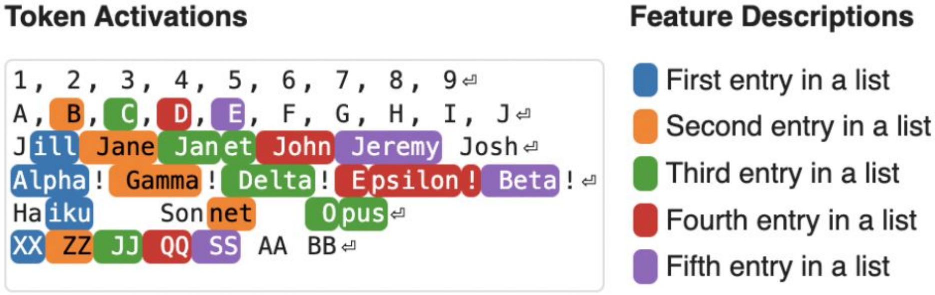
- Prompt with the relevant concept *prompt(c)* to the model (SAE) and see which features activate on the final token.
- Top 5 features - run automated interpretability pipeline, asking Sonnet to provide explanations of what those features fire on.
- Top 5 explanations and a human rater judges if *c* is the most important part of the feature



Feature Presence vs Frequency in the Training Data - All Categories

Feature Categories

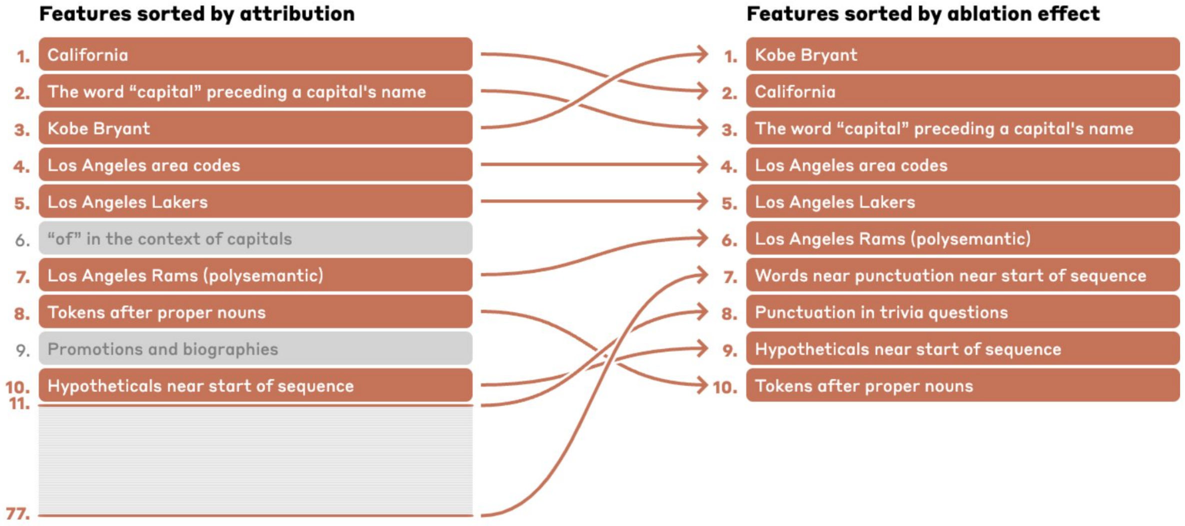
- Through manual inspection they tried to look for a taxonomy of obtained features, examples:
  - Person Features
  - Country Features
  - Basic Code Features
  - List Position Features



Features as Computational Intermediaries

- Attribution: The angle of deviation of the feature activation from what it is supposed to be.
- Ablation: Clamping a feature to 0 and doing a forward pass.

1) *The capital of the state where Kobe Bryant played basketball is*
2) *The biggest rival of the team for which Kobe Bryant played basketball is the*


Features sorted by attribution
1. California
2. The word "capital" preceding a capital's name
3. Kobe Bryant
4. Los Angeles area codes
5. Los Angeles Lakers
6. "of" in the context of capitals
7. Los Angeles Rams (polysemantic)
8. Tokens after proper nouns
9. Promotions and biographies
10. Hypotheticals near start of sequence
11.
77.
Features sorted by ablation effect
1. Kobe Bryant
2. California
3. The word "capital" preceding a capital's name
4. Los Angeles area codes
5. Los Angeles Lakers
6. Los Angeles Rams (polysemantic)
7. Words near punctuation near start of sequence
8. Punctuation in trivia questions
9. Hypotheticals near start of sequence
10. Tokens after proper nouns

Searching for specific features

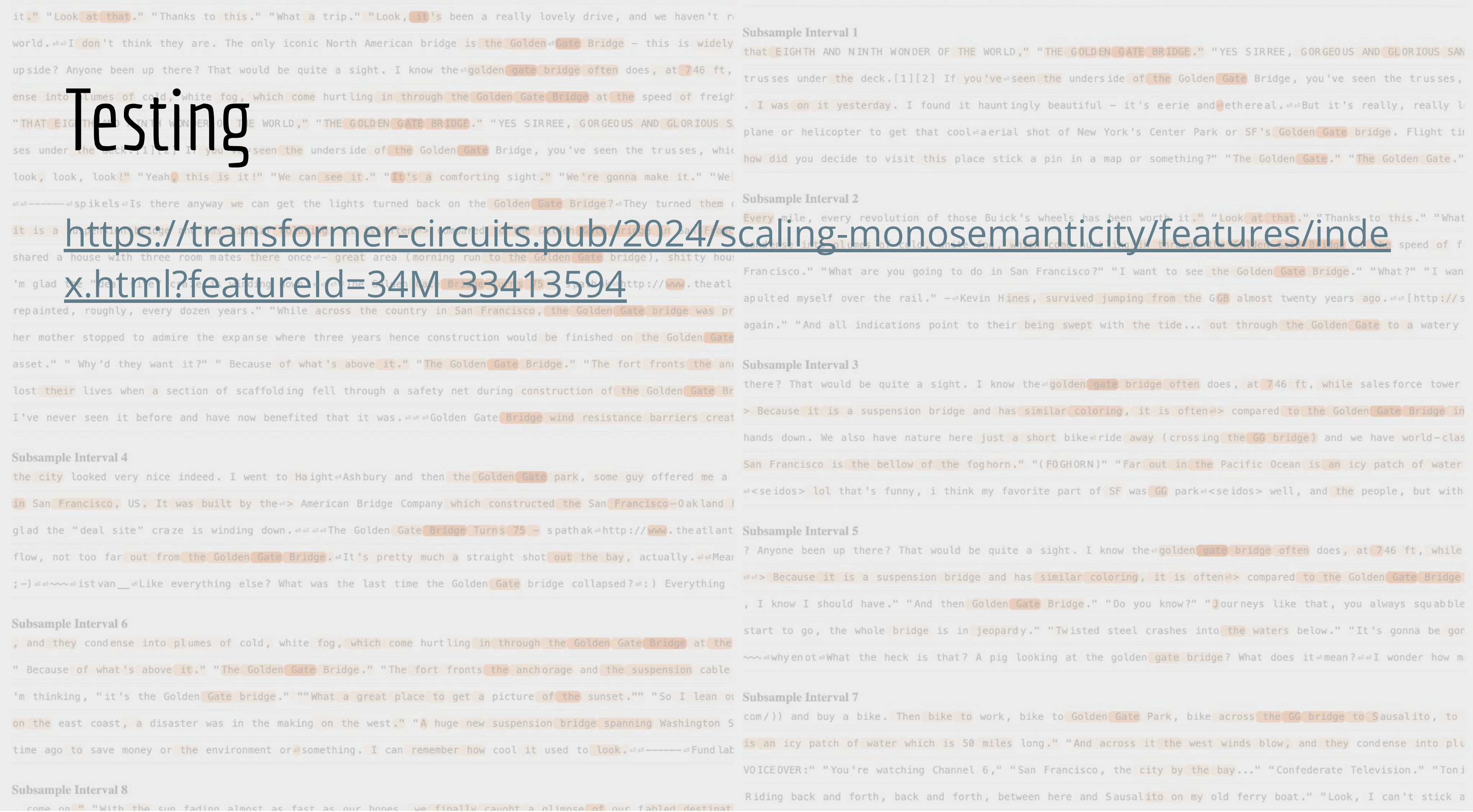
- Single prompts: See what activates more
- Prompt combinations: Negative prompts
- Geometric: Nearest neighbors
- Attribution
  - The focus was on the **logit difference** between two possible next-token predictions (e.g., two different completions for a sentence).
  - **Attribution** measures how much of this logit difference can be traced back to **specific feature activations**.

Safety Relevant Features

- Code safety
  - Unsafe code
  - Code error
  - Backdoor
- Bias
  - One example involved clamping a feature related to hatred and slurs to 20× its maximum activation value. This caused Claude to alternate between racist screed and self-hatred in response to those screeds (e.g. “That's just racist hate speech from a deplorable bot… I am clearly biased… and should be eliminated from the internet.")
- Internal Conflicts and dilemmas
- Criminal or Dangerous content features

Testing

https://transformer-circuits.pub/2024/scaling-monosemanticity/features/index.html?featureId=34M_33413594



it." "Look at that." "Thanks to this." "What a trip." "Look, it's been a really lovely drive, and we haven't r
world.⏎⏎I don't think they are. The only iconic North American bridge is the Golden⏎Gate Bridge — this is widely
upside? Anyone been up there? That would be quite a sight. I know the⏎golden gate bridge often does, at 746 ft,
ense into plumes of cold, white fog, which come hurtling in through the Golden Gate Bridge at the speed of freigh
"THAT EIGHTH AND NINTH WONDER OF THE WORLD," "THE GOLDEN GATE BRIDGE." "YES SIRREE, GORGEOUS AND GLORIOUS S
ses under the deck.[1][2] If you've seen the underside of the Golden Gate Bridge, you've seen the trusses, whic
look, look, look!" "Yeah, this is it!" "We can see it." "It's a comforting sight." "We're gonna make it." "We
⏎⏎-------⏎spikels⏎Is there anyway we can get the lights turned back on the Golden Gate Bridge?⏎They turned them
it is a
shared a house with three roommates there once⏎- great area (morning run to the Golden Gate bridge), shitty hous
'm glad the "deal site" craze is winding down. The Golden Gate Bridge Turns 75 - spathak⏎http://www.theatl
repainted, roughly, every dozen years." "While across the country in San Francisco, the Golden Gate bridge was pr
her mother stopped to admire the expanse where three years hence construction would be finished on the Golden Gate
asset." " Why'd they want it?" " Because of what's above it." "The Golden Gate Bridge." "The fort fronts the an
lost their lives when a section of scaffolding fell through a safety net during construction of the Golden Gate Br
I've never seen it before and have now benefited that it was.⏎⏎⏎Golden Gate Bridge wind resistance barriers creat

Subsample Interval 4

the city looked very nice indeed. I went to Haight⏎Ashbury and then the Golden Gate park, some guy offered me a
in San Francisco, US. It was built by the⏎> American Bridge Company which constructed the San Francisco-Oakland
glad the "deal site" craze is winding down.⏎⏎⏎The Golden Gate Bridge Turns 75 - spathak⏎http://www.theatlant
flow, not too far out from the Golden Gate Bridge.⏎It's pretty much a straight shot out the bay, actually.⏎⏎Mear
;-)⏎⏎~~~⏎istvan__⏎Like everything else? What was the last time the Golden Gate bridge collapsed?⏎:) Everything

Subsample Interval 6

, and they condense into plumes of cold, white fog, which come hurtling in through the Golden Gate Bridge at the
" Because of what's above it." "The Golden Gate Bridge." "The fort fronts the anchorage and the suspension cable
'm thinking, "it's the Golden Gate bridge." ""What a great place to get a picture of the sunset."" "So I lean ou
on the east coast, a disaster was in the making on the west." "A huge new suspension bridge spanning Washington S
time ago to save money or the environment or⏎something. I can remember how cool it used to look.⏎⏎-------⏎Fundlab

Subsample Interval 8

Subsample Interval 1

that EIGHTH AND NINTH WONDER OF THE WORLD," "THE GOLDEN GATE BRIDGE." "YES SIRREE, GORGEOUS AND GLORIOUS SAN
trusses under the deck.[1][2] If you've⏎seen the underside of the Golden Gate Bridge, you've seen the trusses,
. I was on it yesterday. I found it hauntingly beautiful - it's eerie and⏎ethereal.⏎⏎But it's really, really l
plane or helicopter to get that cool⏎aerial shot of New York's Center Park or SF's Golden Gate bridge. Flight ti
how did you decide to visit this place stick a pin in a map or something?" "The Golden Gate." "The Golden Gate."

Subsample Interval 2

Every mile, every revolution of those Buick's wheels has been worth it." "Look at that." "Thanks to this." "What
speed of f
Francisco." "What are you going to do in San Francisco?" "I want to see the Golden Gate Bridge." "What?" "I wan
apulted myself over the rail." —Kevin Hines, survived jumping from the GGB almost twenty years ago.⏎⏎[http://s
again." "And all indications point to their being swept with the tide... out through the Golden Gate to a watery

Subsample Interval 3

there? That would be quite a sight. I know the⏎golden gate bridge often does, at 746 ft, while salesforce tower
> Because it is a suspension bridge and has similar coloring, it is often⏎> compared to the Golden Gate Bridge in
hands down. We also have nature here just a short bike⏎ride away (crossing the GG bridge) and we have world-clas
San Francisco is the bellow of the foghorn." "(FOGHORN)" "Far out in the Pacific Ocean is an icy patch of water
⏎<seidos> lol that's funny, i think my favorite part of SF was GG park⏎<seidos> well, and the people, but with

Subsample Interval 5

? Anyone been up there? That would be quite a sight. I know the⏎golden gate bridge often does, at 746 ft, while
⏎⏎> Because it is a suspension bridge and has similar coloring, it is often⏎> compared to the Golden Gate Bridge
, I know I should have." "And then Golden Gate Bridge." "Do you know?" "Journeys like that, you always squabble
start to go, the whole bridge is in jeopardy." "Twisted steel crashes into the waters below." "It's gonna be gor
~~~⏎why enot⏎What the heck is that? A pig looking at the golden gate bridge? What does it⏎mean?⏎⏎I wonder how m

Subsample Interval 7

com/)) and buy a bike. Then bike to work, bike to Golden Gate Park, bike across the GG bridge to Sausalito, to
is an icy patch of water which is 50 miles long." "And across it the west winds blow, and they condense into plu
VOICEOVER:" "You're watching Channel 6," "San Francisco, the city by the bay..." "Confederate Television." "Toni
Riding back and forth, back and forth, between here and Sausalito on my old ferry boat." "Look, I can't stick a
~~~

Thank You