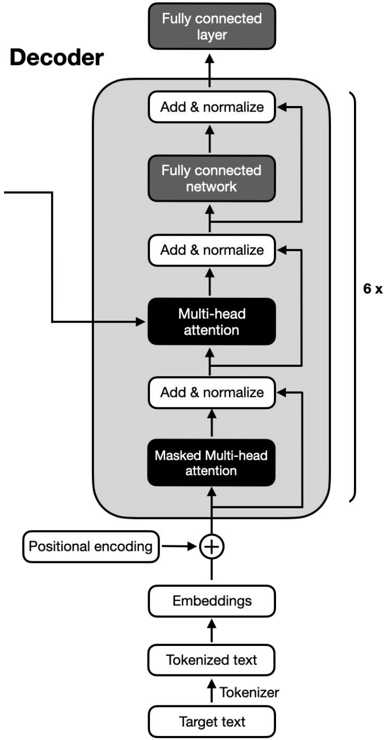
LLM interpretability via sparse autoencoders

INFO 3350/6350: Lecture 20

Why does interpretability matter?

- Can you trust your model?
- Can you trust your data?
- Is the model biased or otherwise harmful?
- Are the features important?
- Are the features useful?
- Training and evaluation costs are high
- Deployment stakes can be high



Approaches to LLM interpretability

Weights are hopeless, but *neurons* might be useful.

There are a lot of neurons, but far fewer than there are total params. Each neuron has many weights and neurons are organized into layers.

Give input, see which neurons activate at a given layer. Hope that there is one (“monosemantic”) neuron that fires. Maybe look for “circuits” of neurons that fire across layers.

Works, sometimes, but polysemanticity and superposition …

Superposition

Superposition

Ideally, want an interpretable neuron to activate iff a single "feature" is present in the input

Aside: what's a feature? Coherent concept *in the data*, nothing to do with the model

Some neurons activate in the presence of multiple, *disparate* features (e.g., cats, hex numbers, gerunds). And some features activate many neurons, but only weakly. But not always …

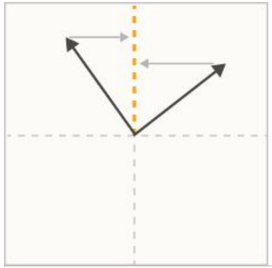
Elhage et al., "Toy Models of Superposition" (2022)

Superposition

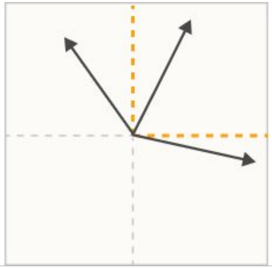
Superposition means that features are represented via combinations of neuron activations. The activating neurons *do not* necessarily have anything to do with the feature in question.

Superposition is *required* to represent more features than there are neurons in the model (e.g., all people, ideas, etc.)

Some features are more important than others



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.

Example of superposition

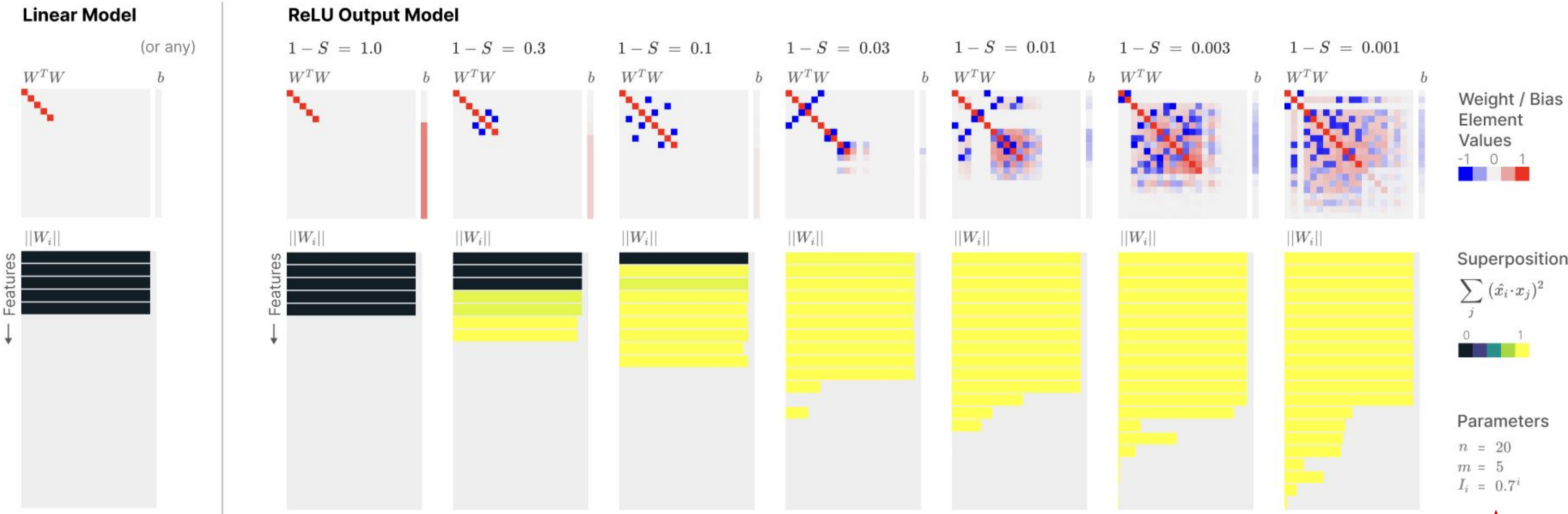
In a small language model ("Toward monosemanticity"), there are features for Arabic script and for Hebrew-language text. But no single neuron activates strongly on all Arabic or Hebrew text input. In fact, the top-activating examples for the neurons that *do* fire on such input contain no examples of Arabic or Hebrew inputs.

Clearly, these features are represented only via superposition. They exist across low levels of activation for many neurons.

Superposition

Superposition only works well if features are *sparse*, i.e., if most inputs have few features, and if the network has a nonlinear activation function.

If these conditions are not met, superposition causes destructive interference, because you can’t tell whether the multiple activations represent a single, superposed feature or a mix of multiple features.



Linear models learn the top m features. $1 - S$ = 0.001 is shown, but others are similar.

In the **dense** regime, ReLU output models also learn the top m features.

As **sparsity increases**, superposition allows models to represent more features. The most important features are initially untouched. This early superposition is organized in antipodal pairs (more on this later).

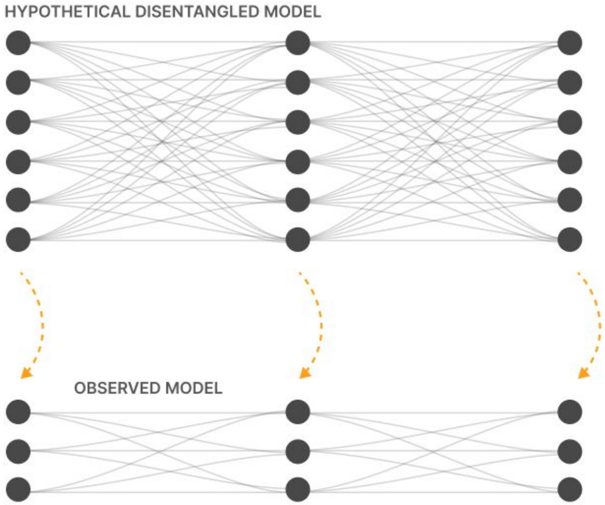
In the **high sparsity** regime, models put all features in superposition, and continue packing more. Note that at this point we begin to see positive interference and negative biases. We'll talk about this more later.

Note this

Hypothetical networks

Sparse networks with superposition can allow a network to function as a simulation of a larger network that is fully *disentangled*

LLMs may be such networks *and* we can use the same hypothesis to *probe* LLMs



Under the superposition hypothesis, the neural networks we observe are **simulations of larger networks** where every neuron is a disentangled feature.

These idealized neurons are **projected** on to the actual network as "almost orthogonal" vectors over the neurons.

The network we observe is a **low-dimensional projection** of the larger network. From the perspective of individual neurons, this presents as polysemanticity.

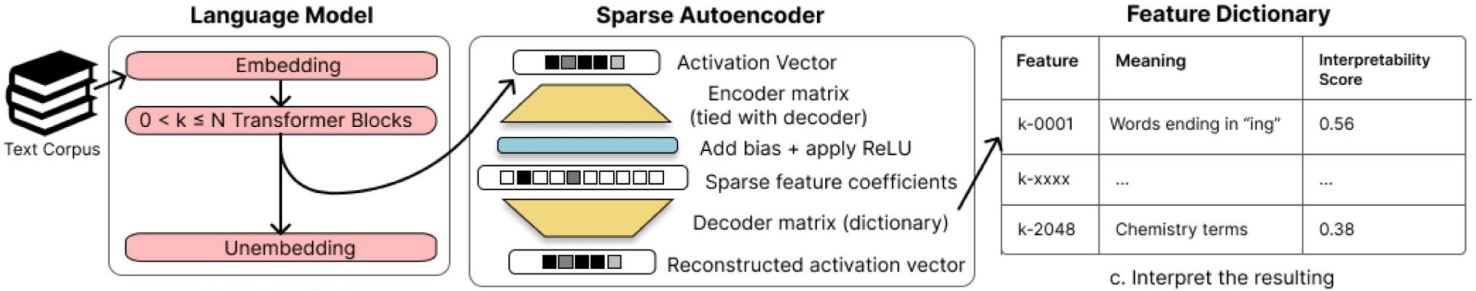
Sparse autoencoders

SAE architecture

Begin with activations from a layer of neurons for many different inputs

Seek to learn a *sparse* representation in *higher*-dimensional space that reconstructs the inputs with minimal loss

Need to penalize non-sparsity, or we just get pseudo-identity



Feature	Meaning	Interpretability Score
k-0001	Words ending in “ing”	0.56
k-xxxx	...	...
k-2048	Chemistry terms	0.38

a. Sample activations from a language model

b. Learn a feature dictionary using an autoencoder that learns to represent activation vectors as a sparse linear combination of feature vectors.

c. Interpret the resulting dictionary features

SAE interpretation

Hidden layer in the SAE is the hypothesized large feature set. In *this space*, 1 neuron = 1 feature.

For a given input, the feature set is manageable for interp, because most features/activations are zero.

Can find inputs that maximally activate each neuron/feature in the SAE.

Can change values of SAE neurons and observe effect on output logits (i.e., next token predictions).

Feature Number
(click for hyperlink)

Human
explanation

Histogram of randomly
sampled non-zero
activations

Top 10 negative and
positive output logits of
the feature

Top 20 max
activating examples

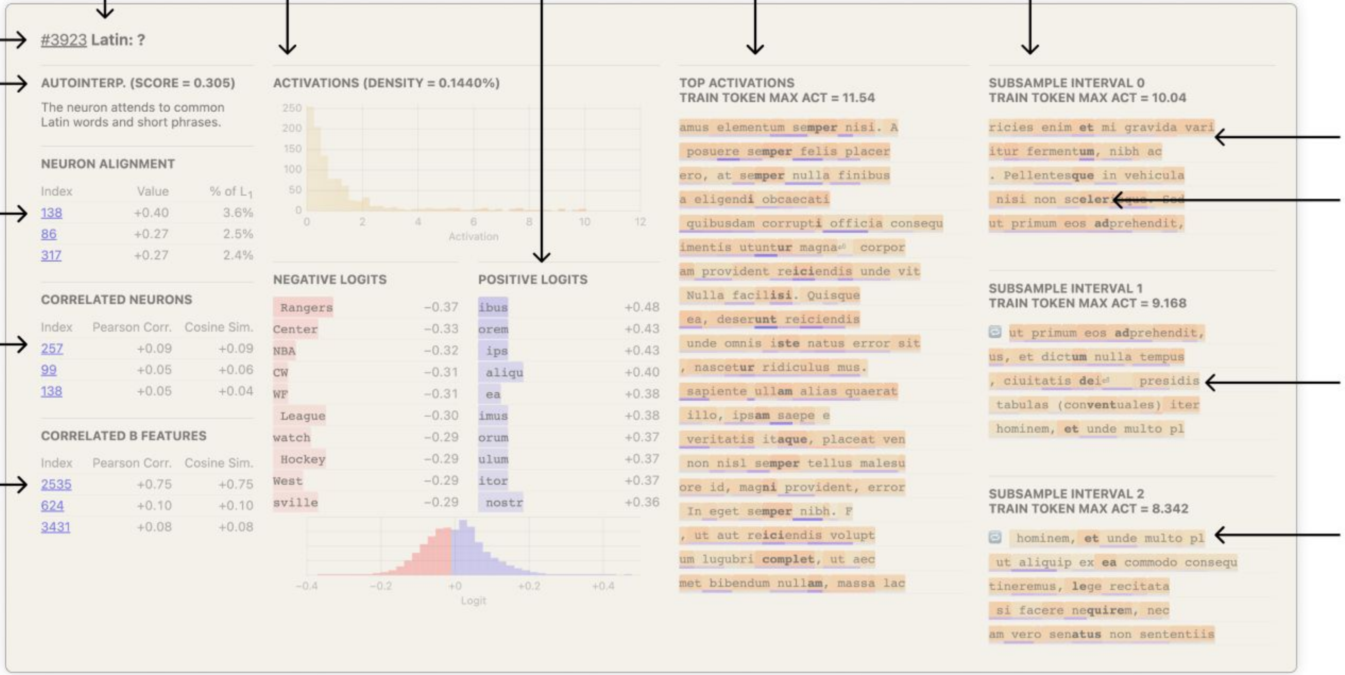
Ten evenly spaced intervals
spanning the full range of
activation values

Autointerp
explanation and
prediction score

Top 3 neurons
by how much the
feature activates them

Top 3 neurons
by token correlation

Top 3 features from
the parallel run with
a different random
seed



#3923 Latin: ?

AUTOINTERP. (SCORE = 0.305)

The neuron attends to common Latin words and short phrases.

NEURON ALIGNMENT

Index	Value	% of L_1
138	+0.40	3.6%
86	+0.27	2.5%
317	+0.27	2.4%

CORRELATED NEURONS

Index	Pearson Corr.	Cosine Sim.
257	+0.09	+0.09
99	+0.05	+0.06
138	+0.05	+0.04

CORRELATED B FEATURES

Index	Pearson Corr.	Cosine Sim.
2535	+0.75	+0.75
624	+0.10	+0.10
3431	+0.08	+0.08

ACTIVATIONS (DENSITY = 0.1440%)

NEGATIVE LOGITS

Rangers	−0.37
Center	−0.33
NBA	−0.32
CW	−0.31
WF	−0.31
League	−0.30
watch	−0.29
Hockey	−0.29
West	−0.29
sville	−0.29

POSITIVE LOGITS

ibus	+0.48
orem	+0.43
ips	+0.43
aliqu	+0.40
ea	+0.38
imus	+0.38
orum	+0.37
ulum	+0.37
itor	+0.37
nostr	+0.36

TOP ACTIVATIONS
TRAIN TOKEN MAX ACT = 11.54

amus elementum se**mper** nisi. A
posuere se**mper** felis placer
ero, at se**mper** nulla finibus
a eligend**i** obcaecati
quibusdam corrupt**i** officia consequ
imentis utunt**ur** magna⏎ corpor
am provident re**ici**endis unde vit
Nulla facil**isi**. Quisque
ea, deser**unt** reiciendis
unde omnis **iste** natus error sit
, nasce**tur** ridiculus mus.
sapiente ull**am** alias quaerat
illo, ips**am** saepe e
veritatis it**aque**, placeat ven
non nisl se**mper** tellus malesu
ore id, mag**ni** provident, error
In eget se**mper** nibh. F
, ut aut re**ici**endis volupt
um lugubri **complet**, ut aec
met bibendum null**am**, massa lac

SUBSAMPLE INTERVAL 0
TRAIN TOKEN MAX ACT = 10.04

ricies enim **et** mi gravida vari
itur ferment**um**, nibh ac
. Pellentes**que** in vehicula
nisi non sc**eler**isque. Sed
ut primum eos **ad**prehendit,

SUBSAMPLE INTERVAL 1
TRAIN TOKEN MAX ACT = 9.168

🔁 ut primum eos **ad**prehendit,
us, et dict**um** nulla tempus
, ciuitatis **dei**⏎ presidis
tabulas (con**vent**uales) iter
hominem, **et** unde multo pl

SUBSAMPLE INTERVAL 2
TRAIN TOKEN MAX ACT = 8.342

🔁 hominem, **et** unde multo pl
ut aliquip ex **ea** commodo consequ
tineremus, **lege** recitata
si facere ne**quire**m, nec
am vero sen**atus** non sententiis

Blue underline means a lower ablation loss (better token prediction); red means a higher loss

Bold token comes from the training data (used to select each example). Surrounding 4 tokens give context

Hover over any token (for 2+ seconds) to see its activation value and ablations

🔁 means this specific example has already appeared in another interval

https://transformer-circuits.pub/2023/monosemantic-features/vis/a1.html

Evaluating SAE features

Are the features that we find meaningful and monosemantic?

Are similar features found across multiple models and SAE settings?

Do features split as the size of the hidden layer in the SAE increases and coalesce as it decreases?

Can SAE features be intervened on in predictable ways?

Feature Number
(click for hyperlink)

Human
explanation

Histogram of randomly
sampled non-zero
activations

Top 10 negative and
positive output logits of
the feature

Top 20 max
activating examples

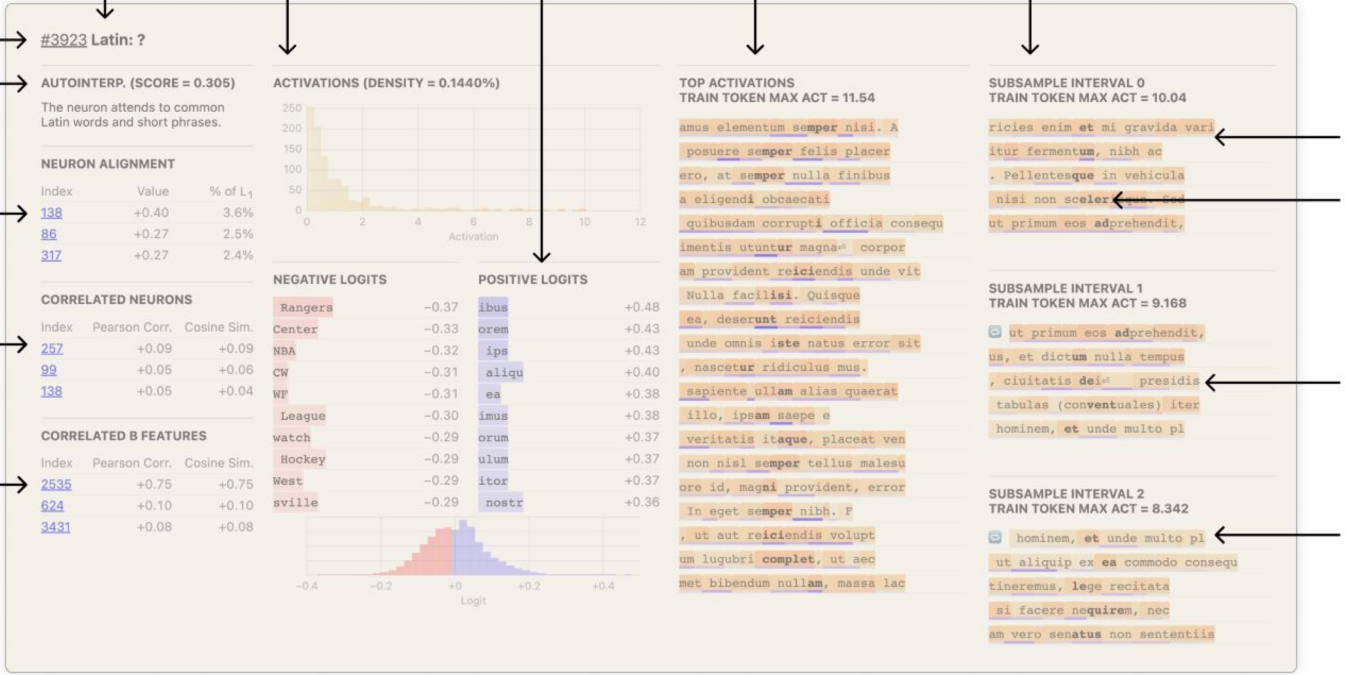
Ten evenly spaced intervals
spanning the full range of
activation values

Autointerp
explanation and
prediction score

Top 3 neurons
by how much the
feature activates them

Top 3 neurons
by token correlation

Top 3 features from
the parallel run with
a different random
seed



#3923 Latin: ?

AUTOINTERP. (SCORE = 0.305)

The neuron attends to common Latin words and short phrases.

NEURON ALIGNMENT

Index	Value	% of L_1
138	+0.40	3.6%
86	+0.27	2.5%
317	+0.27	2.4%

CORRELATED NEURONS

Index	Pearson Corr.	Cosine Sim.
257	+0.09	+0.09
99	+0.05	+0.06
138	+0.05	+0.04

CORRELATED B FEATURES

Index	Pearson Corr.	Cosine Sim.
2535	+0.75	+0.75
624	+0.10	+0.10
3431	+0.08	+0.08

ACTIVATIONS (DENSITY = 0.1440%)

NEGATIVE LOGITS

Rangers	−0.37
Center	−0.33
NBA	−0.32
CW	−0.31
WF	−0.31
League	−0.30
watch	−0.29
Hockey	−0.29
West	−0.29
sville	−0.29

POSITIVE LOGITS

ibus	+0.48
orem	+0.43
ips	+0.43
aliqu	+0.40
ea	+0.38
imus	+0.38
orum	+0.37
ulum	+0.37
itor	+0.37
nostr	+0.36

TOP ACTIVATIONS
TRAIN TOKEN MAX ACT = 11.54

amus elementum se**mper** nisi. A
posuere se**mper** felis placer
ero, at se**mper** nulla finibus
a eligend**i** obcaecati
quibusdam corrupt**i** officia consequ
imentis utunt**ur** magna↵ corpor
am provident re**ici**endis unde vit
Nulla facil**isi**. Quisque
ea, deser**unt** reiciendis
unde omnis i**ste** natus error sit
, nascet**ur** ridiculus mus.
sapiente ull**am** alias quaerat
illo, ips**am** saepe e
veritatis it**aque**, placeat ven
non nisl se**mper** tellus malesu
ore id, mag**ni** provident, error
In eget se**mper** nibh. F
, ut aut re**ici**endis volupt
um lugubri **complet**, ut aec
met bibendum null**am**, massa lac

SUBSAMPLE INTERVAL 0
TRAIN TOKEN MAX ACT = 10.04

ricies enim **et** mi gravida vari
itur ferment**um**, nibh ac
. Pellentes**que** in vehicula
nisi non sc**eler**isque. Sed
ut primum eos **ad**prehendit,

SUBSAMPLE INTERVAL 1
TRAIN TOKEN MAX ACT = 9.168

🔁 ut primum eos **ad**prehendit,
us, et dict**um** nulla tempus
, ciuitatis **dei**↵ presidis
tabulas (con**vent**uales) iter
hominem, **et** unde multo pl

SUBSAMPLE INTERVAL 2
TRAIN TOKEN MAX ACT = 8.342

🔁 hominem, **et** unde multo pl
ut aliquip ex **ea** commodo consequ
tineremus, **le**ge recitata
si facere ne**quire**m, nec
am vero sen**atus** non sententiis

Blue underline means a lower ablation loss (better token prediction); red means a higher loss

Bold token comes from the training data (used to select each example). Surrounding 4 tokens give context

Hover over any token (for 2+ seconds) to see its activation value and ablations

🔁 means this specific example has already appeared in another interval

https://transformer-circuits.pub/2023/monosemantic-features/vis/a1.html

Steering

Using SAE features to change output

- An interactive demo: Gemma Scope
- A code tutorial from SAE Lens (scroll down to “feature steering” and “feature ablations”)