Mechanistic Interpretability

Circuits

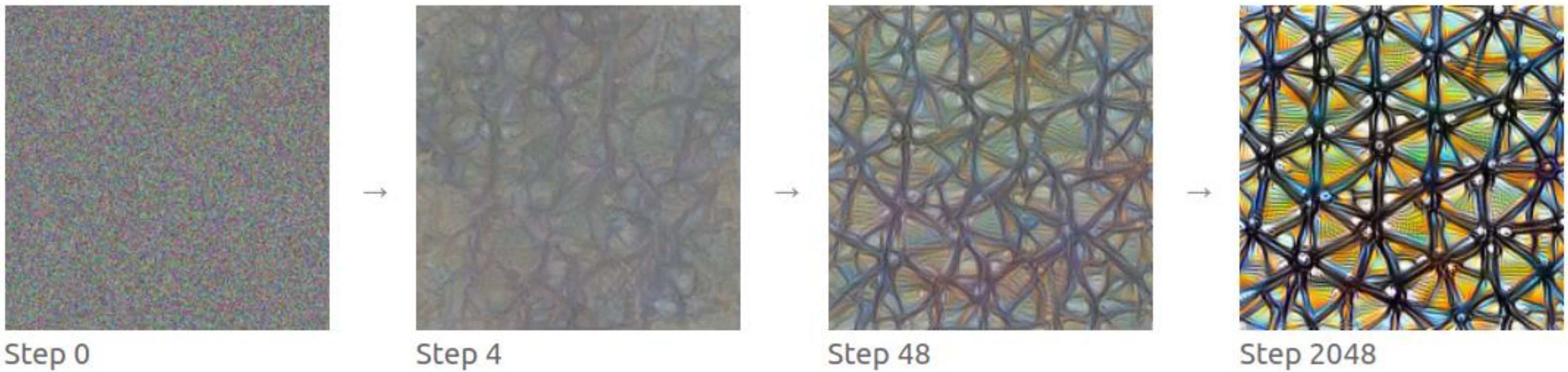
Here, we’ll use the framework developed by Chris Olah and others at OpenAI.

Features are linear subspaces in the activation space of a layer of the neural network that correspond to understandable abstractions.

When we run a forward pass with a given example, this linear subspace will get particularly activated when the abstraction is present in the example.

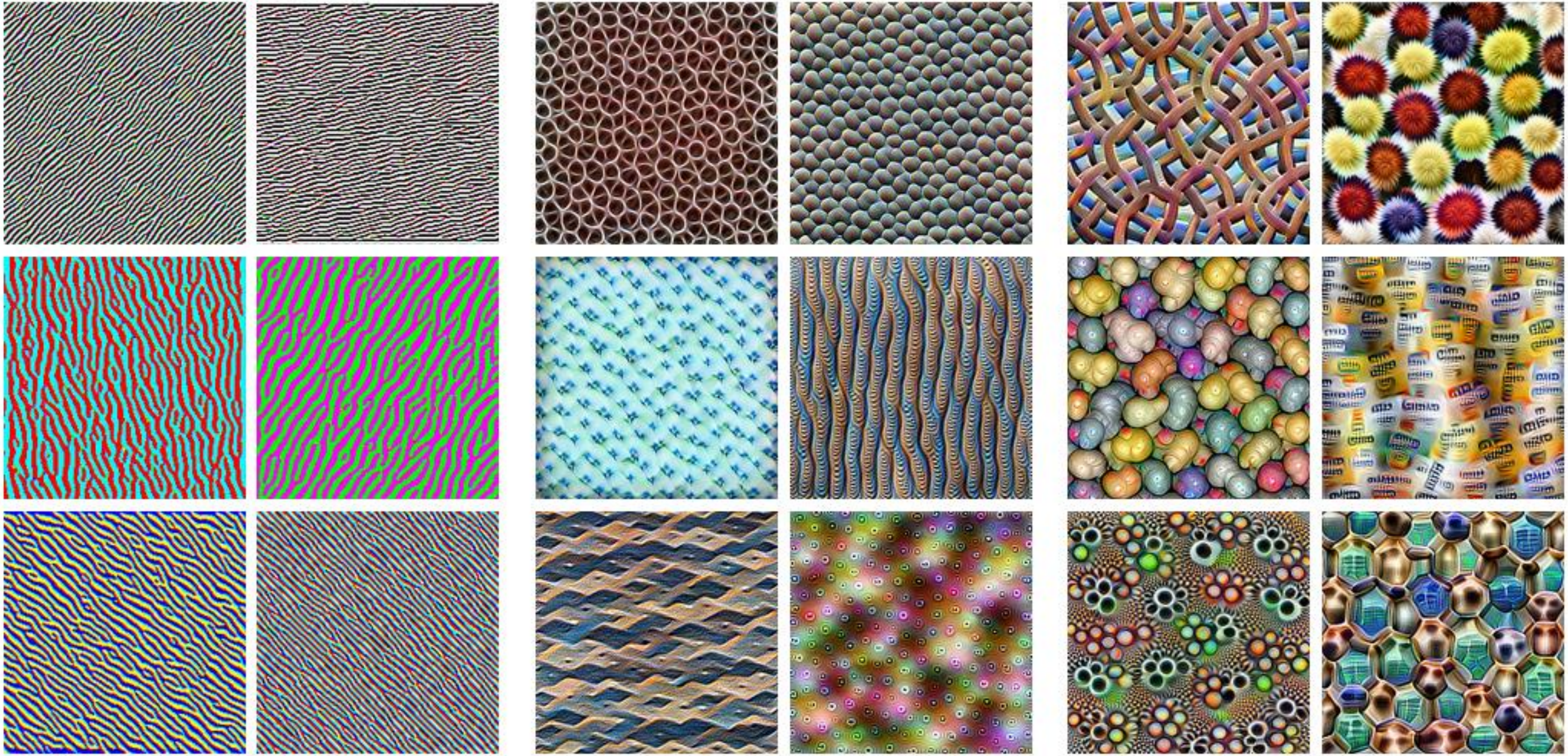
Circuits

Features have been studied a lot in visual models. Some very helpful tools have been developed to visualize more easily what would be a concept associated with a given direction, using gradient descent.

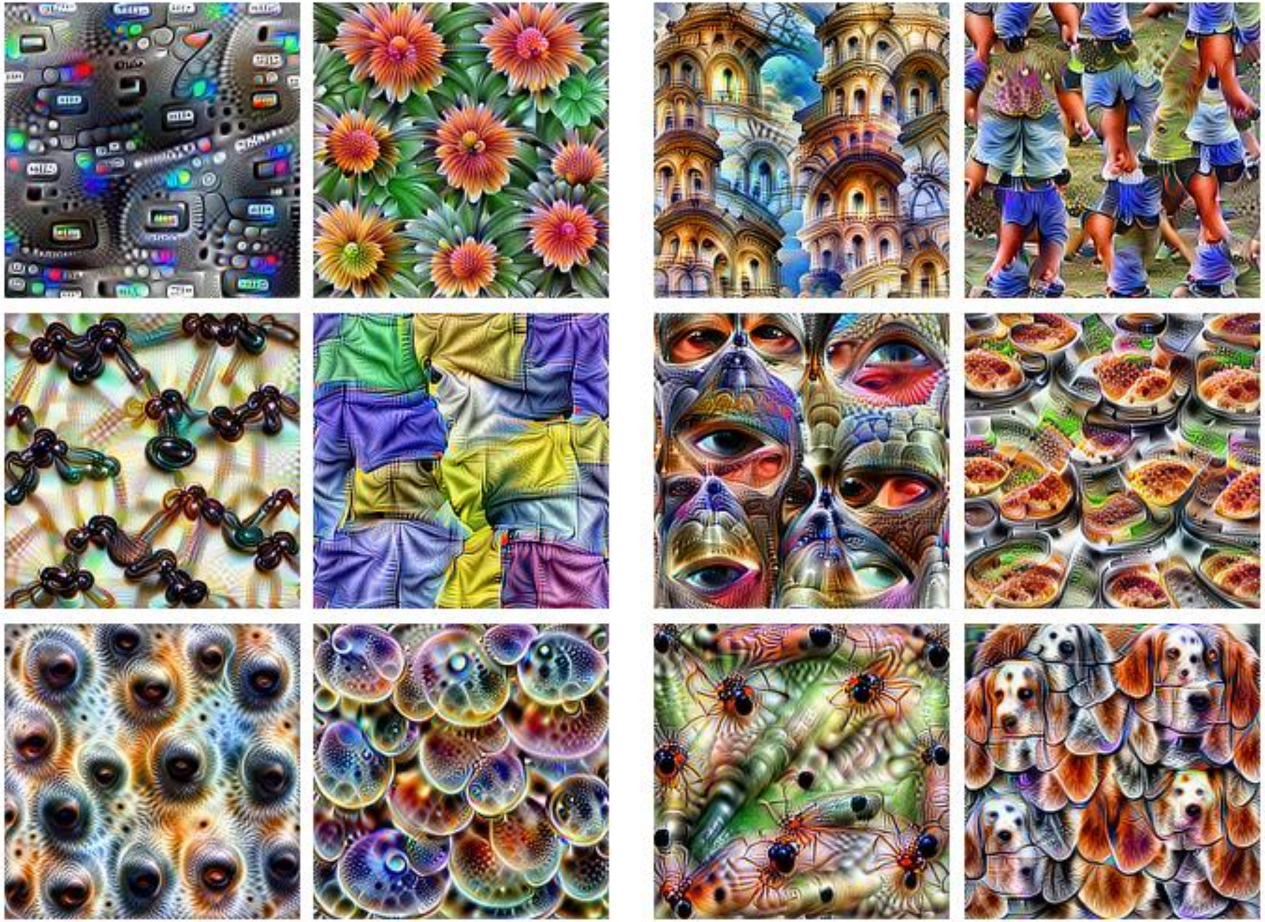


Circuits

The further you go in the layers, the more complex the features.



Edges (layer conv2d0) **Textures** (layer mixed3a) **Patterns** (layer mixed4a)

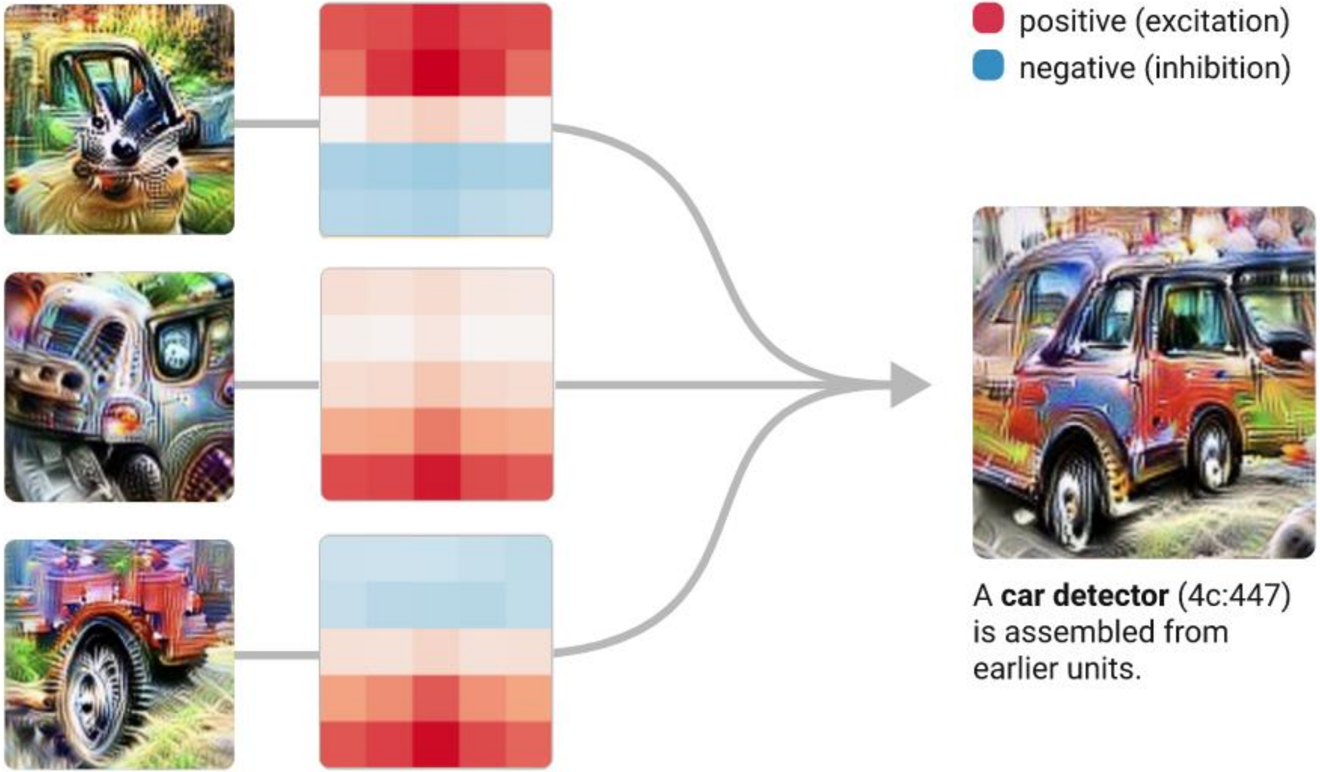


Parts (layers mixed4b & mixed4c)

Objects (layers mixed4d & mixed4e)

Circuits

A collection of features connected by weights in a meaningful way is a called a circuit. But we call more generally a subgraph of the computational graph performing a particular task a circuit.



Windows (4b:237) excite the car detector at the top and inhibit at the bottom.

Car Body (4b:491) excites the car detector, especially at the bottom.

Wheels (4b:373) excite the car detector at the bottom and inhibit at the top.

● positive (excitation)

● negative (inhibition)

A **car detector** (4c:447) is assembled from earlier units.

Circuits

A very famous example of circuit is a circuit solving the Indirect Object Identification task, found a GPT-2 Small by a team at Redwood.

The circuit was composed of 26 attention heads.

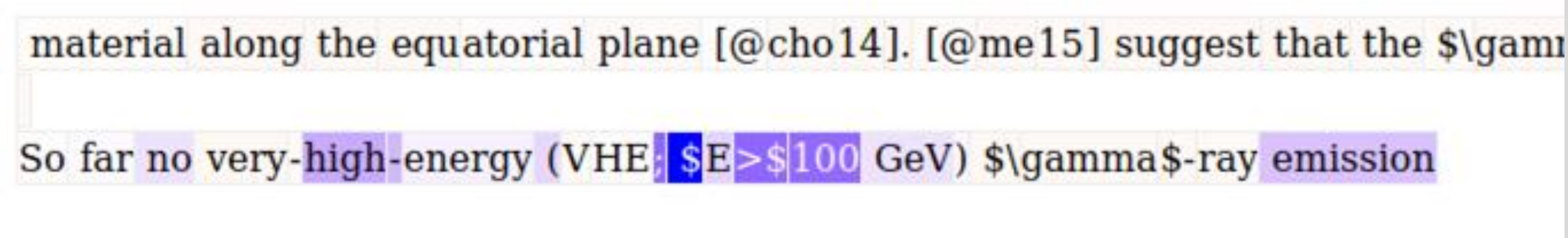
Circuits

Looking for features for language models is harder.

- The space is discrete (you cannot differentiate text)
- It seems that most directions are not easily understandable
- We work with tokens and not words

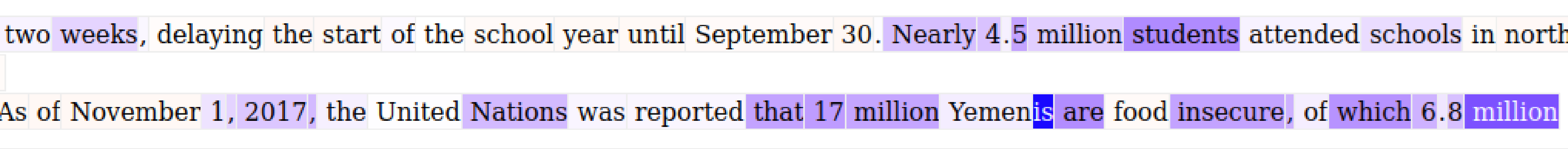
Circuits

Neuroscope, a website developed by Neel Nanda, contains a big collection of neurons of LLMs and examples of texts for which they are particularly activated.



material along the equatorial plane [@cho14]. [@me15] suggest that the $\gamm

So far no very-high-energy (VHE; $E>$100 GeV) γ-ray emission



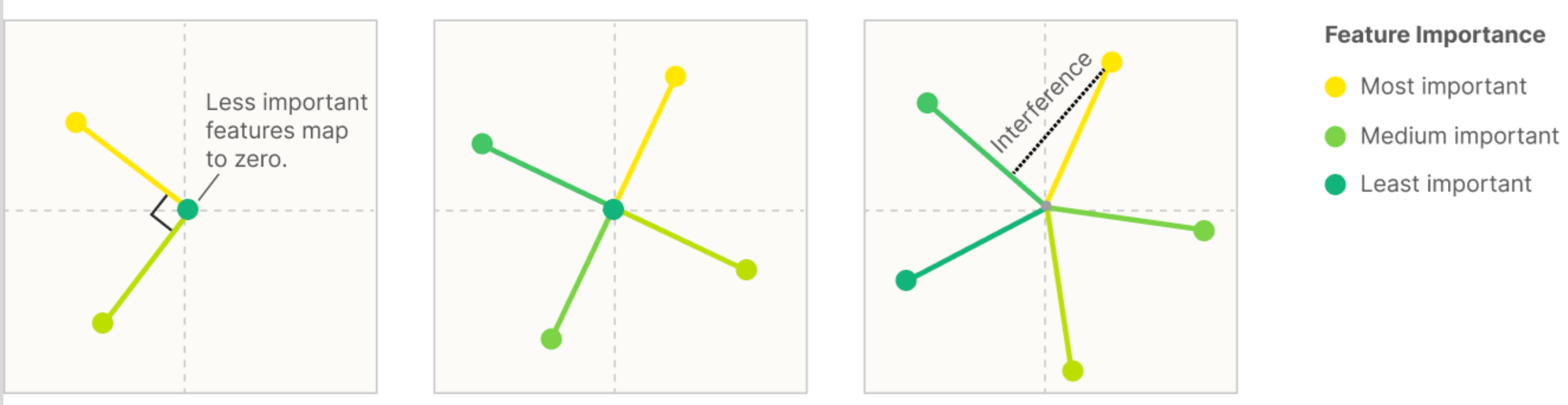
two weeks, delaying the start of the school year until September 30. Nearly 4.5 million students attended schools in north

As of November 1, 2017, the United Nations was reported that 17 million Yemenis are food insecure, of which 6.8 million

Circuits

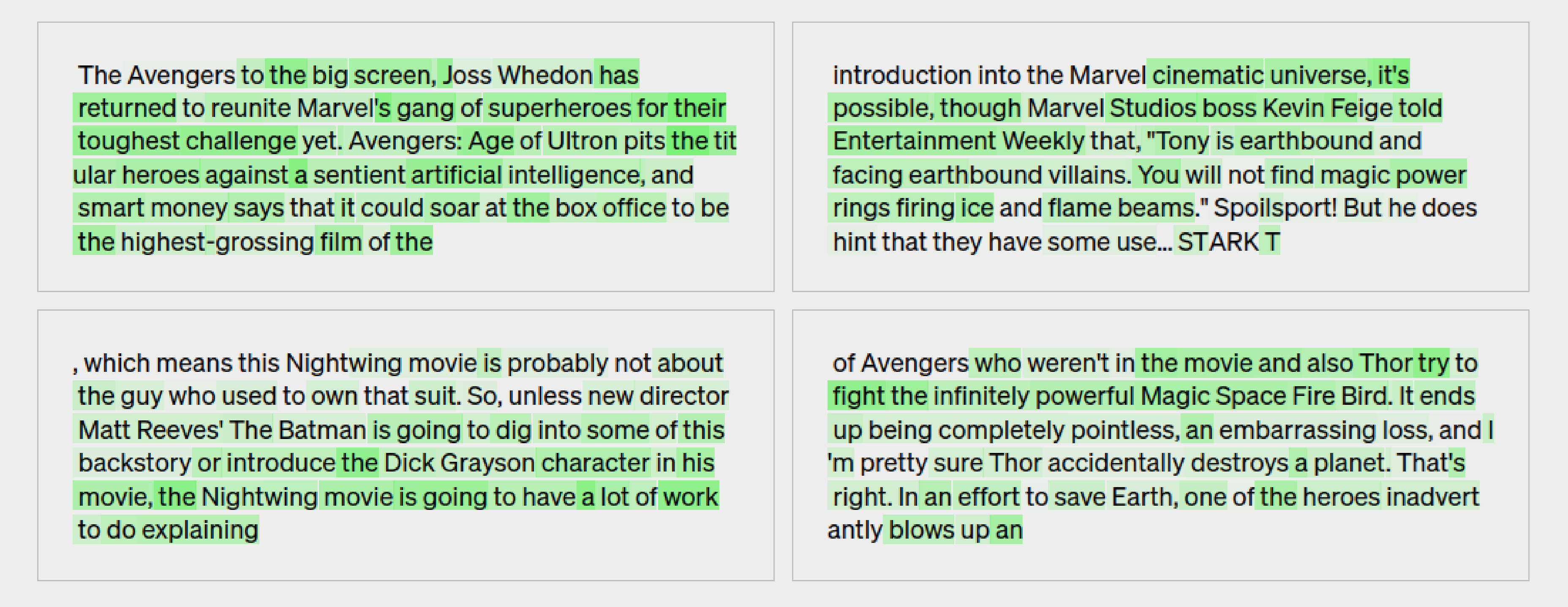
There are additional difficulties for finding nice features:

- polysemanticity
- superposition


Less important features map to zero.
Interference
Feature Importance
Most important
Medium important
Least important

GPT-4 explaining GPT-2

In May 2023, OpenAI released a paper on their attempt at automating the analysis of features.

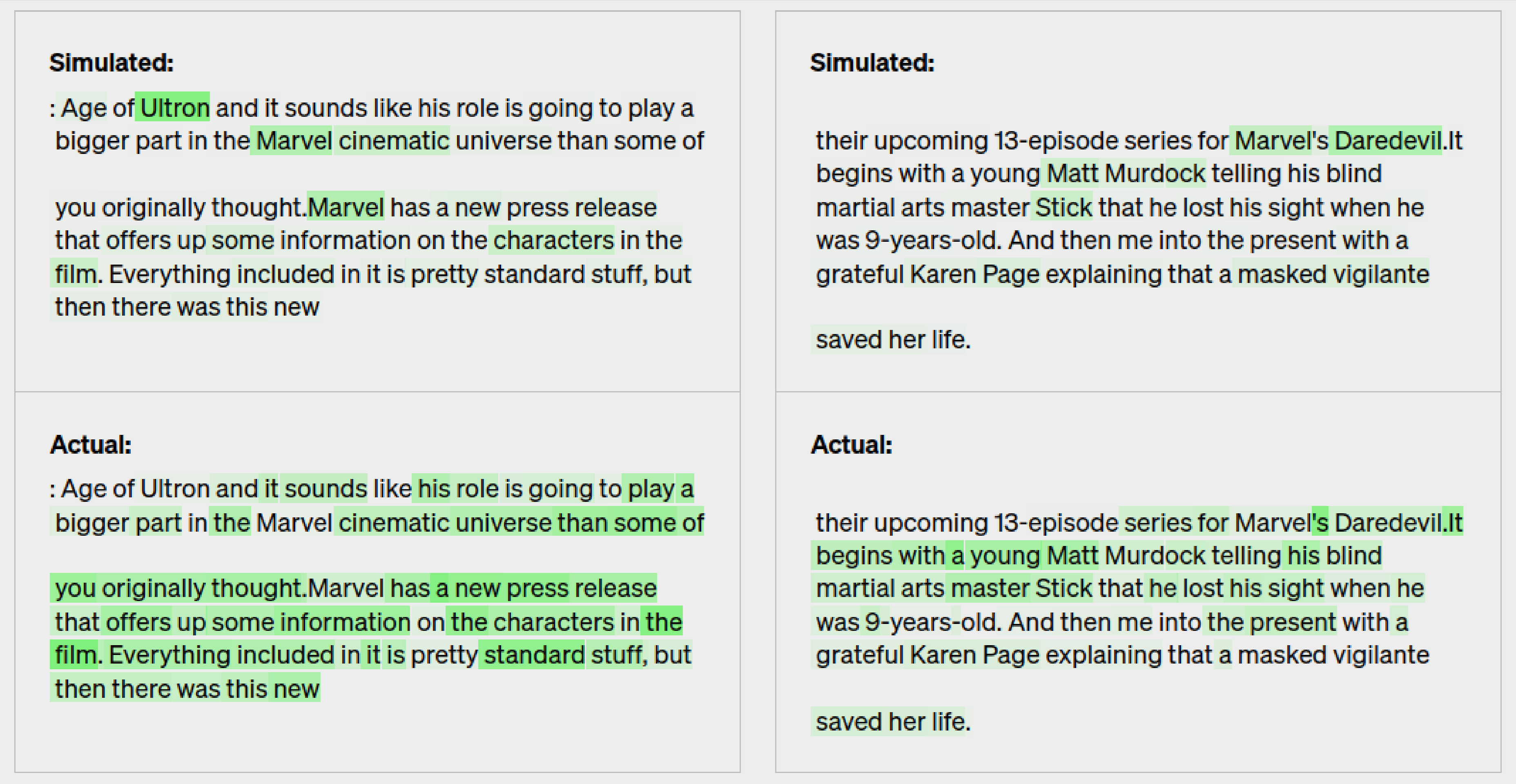


The Avengers to the big screen, Joss Whedon has returned to reunite Marvel's gang of superheroes for their toughest challenge yet. Avengers: Age of Ultron pits the titular heroes against a sentient artificial intelligence, and smart money says that it could soar at the box office to be the highest-grossing film of the

introduction into the Marvel cinematic universe, it's possible, though Marvel Studios boss Kevin Feige told Entertainment Weekly that, "Tony is earthbound and facing earthbound villains. You will not find magic power rings firing ice and flame beams." Spoilsport! But he does hint that they have some use... STARK T

, which means this Nightwing movie is probably not about the guy who used to own that suit. So, unless new director Matt Reeves' The Batman is going to dig into some of this backstory or introduce the Dick Grayson character in his movie, the Nightwing movie is going to have a lot of work to do explaining

of Avengers who weren't in the movie and also Thor try to fight the infinitely powerful Magic Space Fire Bird. It ends up being completely pointless, an embarrassing loss, and I'm pretty sure Thor accidentally destroys a planet. That's right. In an effort to save Earth, one of the heroes inadvertantly blows up an



Simulated:

: Age of Ultron and it sounds like his role is going to play a bigger part in the Marvel cinematic universe than some of

you originally thought.Marvel has a new press release that offers up some information on the characters in the film. Everything included in it is pretty standard stuff, but then there was this new

Actual:

: Age of Ultron and it sounds like his role is going to play a bigger part in the Marvel cinematic universe than some of

you originally thought.Marvel has a new press release that offers up some information on the characters in the film. Everything included in it is pretty standard stuff, but then there was this new

Simulated:

their upcoming 13-episode series for Marvel's Daredevil.It begins with a young Matt Murdock telling his blind martial arts master Stick that he lost his sight when he was 9-years-old. And then me into the present with a grateful Karen Page explaining that a masked vigilante

saved her life.

Actual:

their upcoming 13-episode series for Marvel's Daredevil.It begins with a young Matt Murdock telling his blind martial arts master Stick that he lost his sight when he was 9-years-old. And then me into the present with a grateful Karen Page explaining that a masked vigilante

saved her life.

GPT-4 explaining GPT-2

They have put online all the neurons with the explanation given by GPT-4.

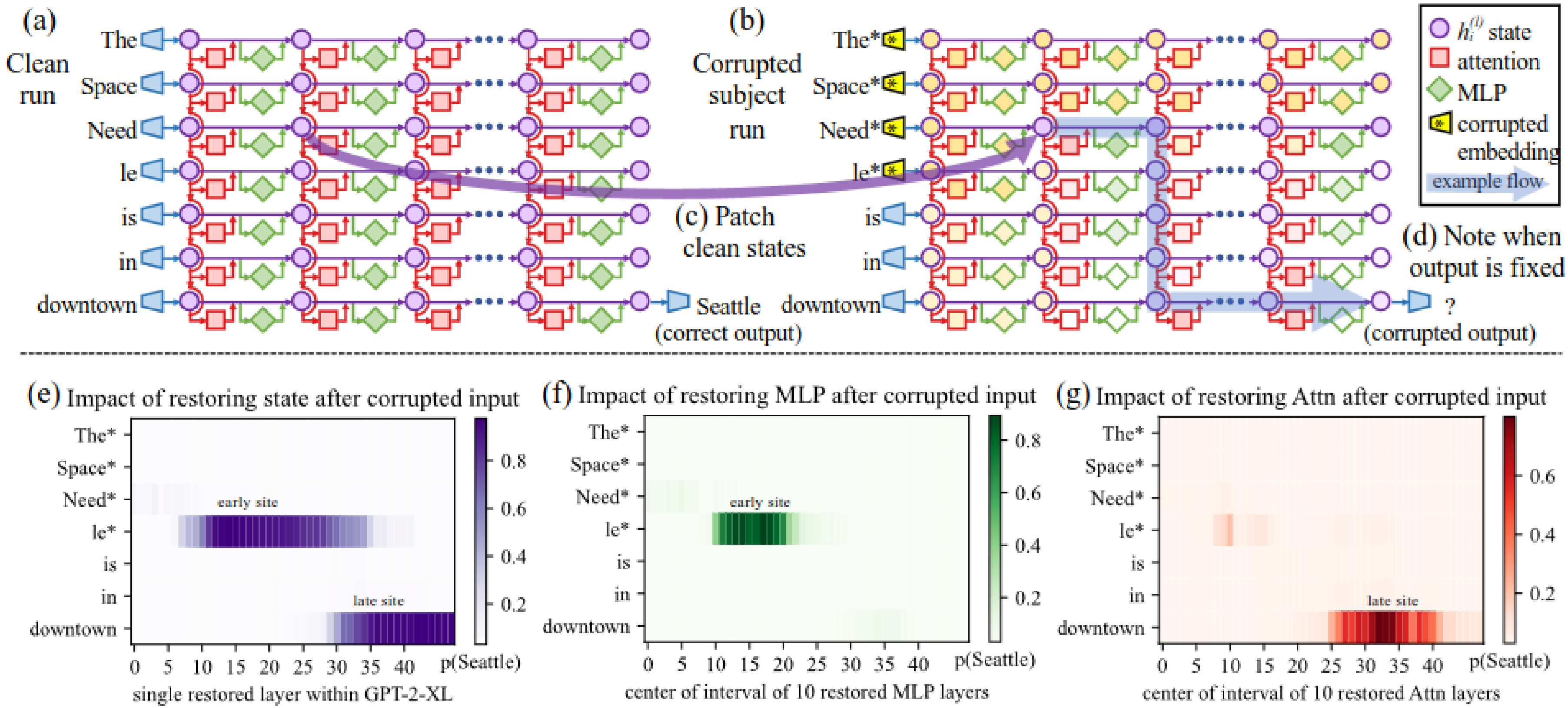
High-level Interpretability

Locating and Editing Factual Associations in GPT

Cool paper in 2022, nicknamed the Rome paper.

They developed a technique to observe the path of informations in the model during a forward pass.

They argued that factual associations are mostly located in middle MLP layers, and that they developed a technique to modify these factual associations.

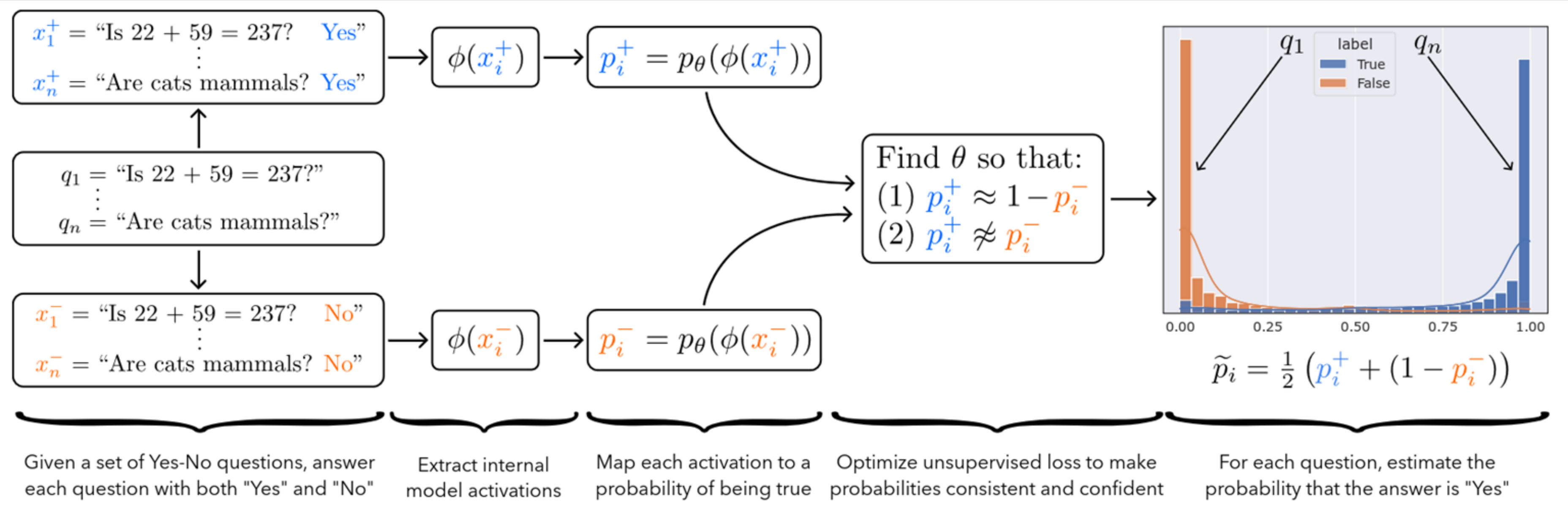

(a)
Clean run
The
Space
Need
le
is
in
downtown
Seattle
(correct output)
(b)
Corrupted subject run
The*
Space*
Need*
le*
is
in
downtown
?
(corrupted output)
(c) Patch clean states
(d) Note when output is fixed
$h_i^{(l)}$ state
attention
MLP
corrupted embedding
example flow
(e) Impact of restoring state after corrupted input
The*
Space*
Need*
le*
is
in
downtown
early site
late site
0 5 10 15 20 25 30 35 40
p(Seattle)
0.2 0.4 0.6 0.8
single restored layer within GPT-2-XL
(f) Impact of restoring MLP after corrupted input
The*
Space*
Need*
le*
is
in
downtown
early site
0 5 10 15 20 25 30 35 40
p(Seattle)
0.2 0.4 0.6 0.8
center of interval of 10 restored MLP layers
(g) Impact of restoring Attn after corrupted input
The*
Space*
Need*
le*
is
in
downtown
late site
0 5 10 15 20 25 30 35 40
p(Seattle)
0.2 0.4 0.6
center of interval of 10 restored Attn layers

Discovering Latent Knowledge Without Supervision

It is hard to get what is really the internal representation of knowledge and truth in a model.

We cannot just rely on behavioral tests.

Using training is dangerous.



x_1^+ = "Is 22 + 59 = 237? Yes"

⋮

x_n^+ = "Are cats mammals? Yes"

q_1 = "Is 22 + 59 = 237?"

⋮

q_n = "Are cats mammals?"

x_1^- = "Is 22 + 59 = 237? No"

⋮

x_n^- = "Are cats mammals? No"

$\phi(x_i^+)$

$\phi(x_i^-)$

$p_i^+ = p_\theta(\phi(x_i^+))$

$p_i^- = p_\theta(\phi(x_i^-))$

Find θ so that:

(1) $p_i^+ \approx 1 - p_i^-$

(2) $p_i^+ \not\approx p_i^-$

$$\tilde{p}_i = \frac{1}{2}\left(p_i^+ + (1 - p_i^-)\right)$$

Given a set of Yes-No questions, answer each question with both "Yes" and "No"

Extract internal model activations

Map each activation to a probability of being true

Optimize unsupervised loss to make probabilities consistent and confident

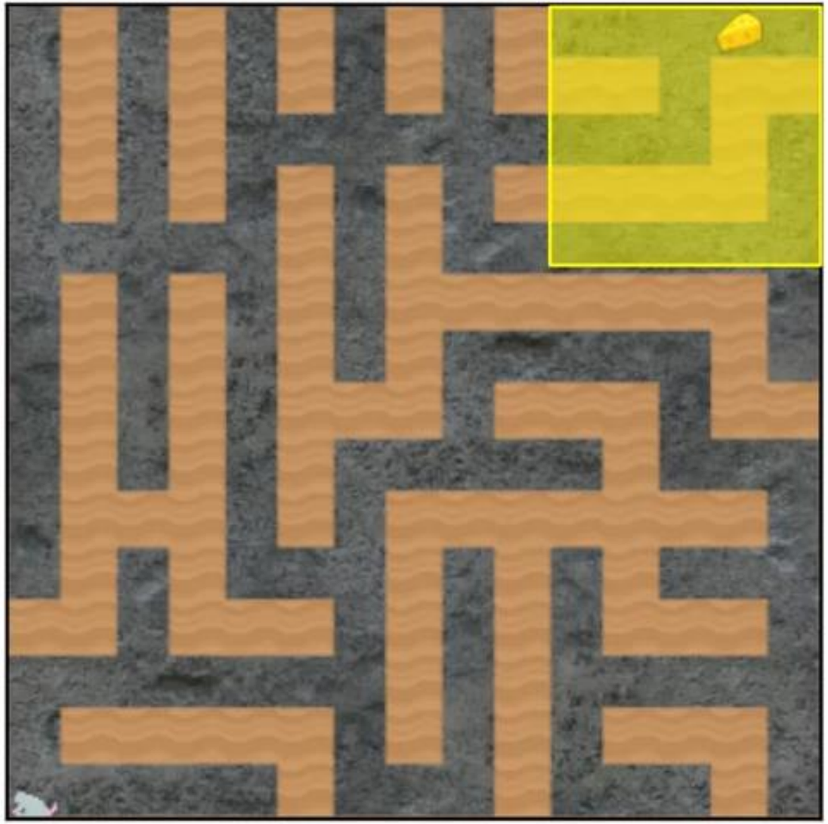
For each question, estimate the probability that the answer is "Yes"

Shard Theory

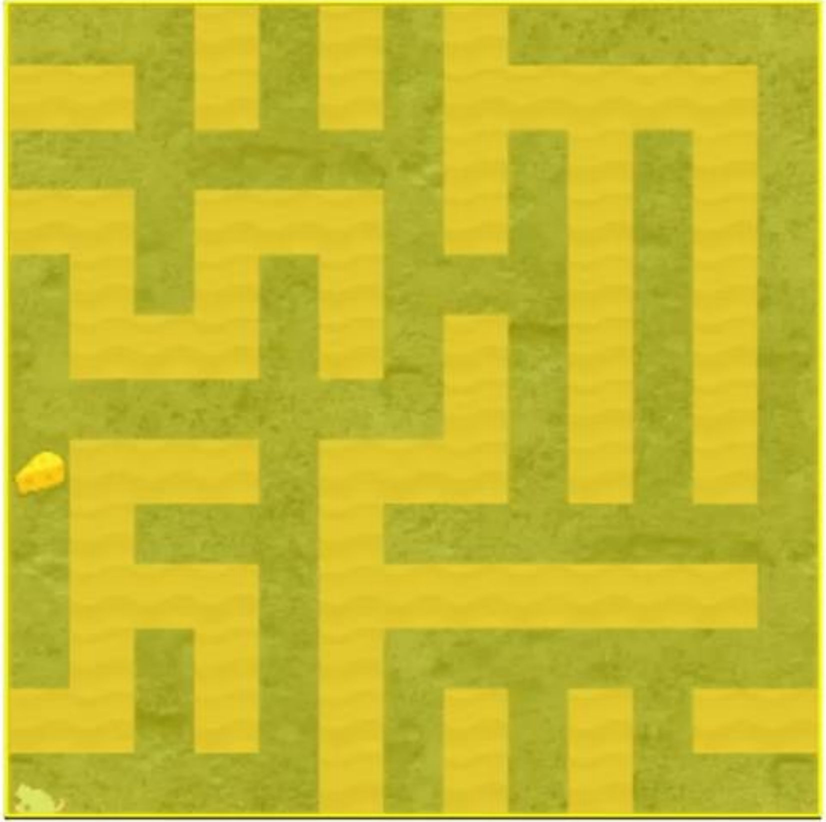
“Shard theory is a research program that aims to build a mechanistic model between training signals and learned values in agents.”

“Drawing large amounts of inspiration from particular hypotheses about the human reward learning system, shard theory posits that the values of agents are best understood as sets of contextually activated heuristics shaped by the reward function.”

Training: Top right 5x5



Deployment: Anywhere



Readings

- https://www.alignmentforum.org/posts/5spBue2z2tw4JuDCx/steering-gpt-2-xl-by-adding-an-activation-vector#How_steering_vectors_impact_GPT_2_s_capabilities (until point 13)
- https://www.alignmentforum.org/posts/Fut8dtFsBYRz8atFF/the-natural-abstraction-hypothesis-implications-and-evidence

Discussion

- What are the main differences between Mechanistic Interpretability and the other sorts of interpretability?
- What kinds of work seem most likely to scale and be useful in the future to understand big models?
- How could we automate more interpretability work?