Explanation in the Era of LLMs

NAACL 2024 tutorial
Section 4: Transformer Understanding

Presented by Sarah Wiegreffe
Thanks Xi and Chenhao for help with slides.

Outline of the presentation

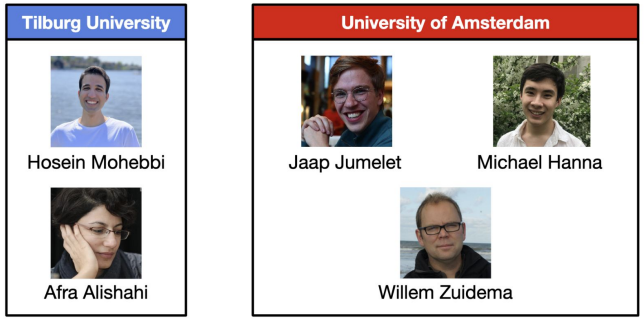
1. Motivation and desiderata
2. Prompting-based Explanations
3. Data attribution
4. **Transformer understanding** ⟸ This section
5. Conclusion and discussion

Tutorial @ EACL 2024: Transformer-Specific Interpretability

Website

Slides

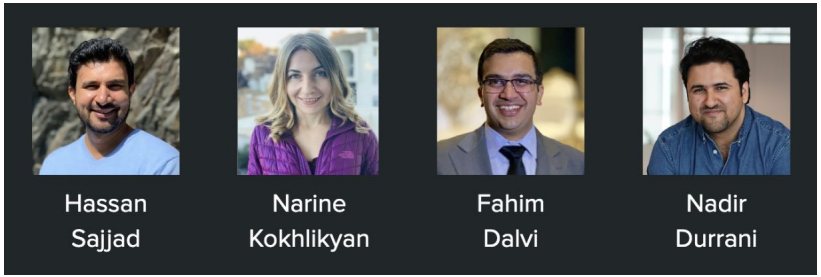
Recording



Tutorial @ NAACL 2021: Fine-grained Interpretation and Causation Analysis in Deep NLP Models

Website/Slides

Recording



Tutorial @ EMNLP 2022: Causal Inference for NLP

Slides/Slides

Recording



Mechanistic Interpretability Workshop @ ICML 2024

Website


Fazl Barez
Research Fellow University of Oxford


Mor Geva
Ass. Prof Tel Aviv University, Visiting Researcher Google Research


Lawrence Chan
PhD student UC Berkeley


Kayo Yin
PhD student UC Berkeley


Neel Nanda
Research Engineer Google DeepMind


Max Tegmark
Professor MIT

Survey Papers

A Primer on the Inner Workings of Transformer Language Models

Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks

Outline

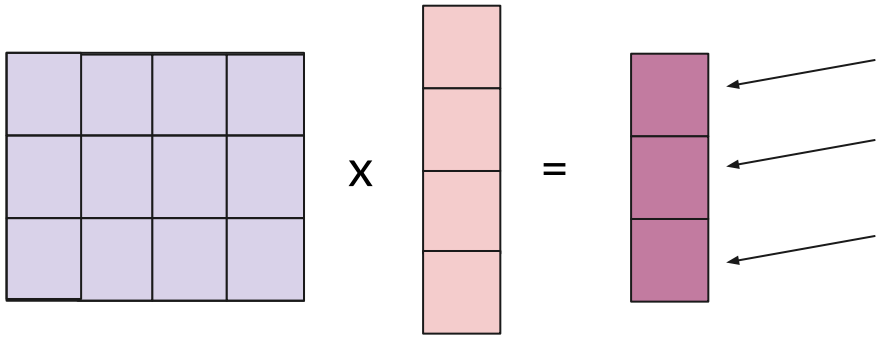
1. Neuron-level interpretability
    a. Sparse Autoencoders
2. Causality Background and Methods
    a. Causal Mediation
    b. Activation Patching/Causal Tracing and Path Patching
    c. High-level causal graphs
    d. Other methods
3. What is mechanistic interpretability?
4. Methods Leveraging Language Model Strengths
    a. Transformer Residual Stream and Linear Structure
    b. Vocabulary projection
    c. Decoding Natural Language Explanations from Representations
5. Conclusion + Q&A

Outline

1. **Neuron-level interpretability**
    a. Sparse Autoencoders
2. Causality Background and Methods
    a. Causal Mediation
    b. Activation Patching/Causal Tracing and Path Patching
    c. High-level causal graphs
    d. Other methods
3. What is mechanistic interpretability?
4. Methods Leveraging Language Model Strengths
    a. Transformer Residual Stream and Linear Structure
    b. Vocabulary projection
    c. Decoding Natural Language Explanations from Representations
5. Conclusion + Q&A

Interpreting Neurons

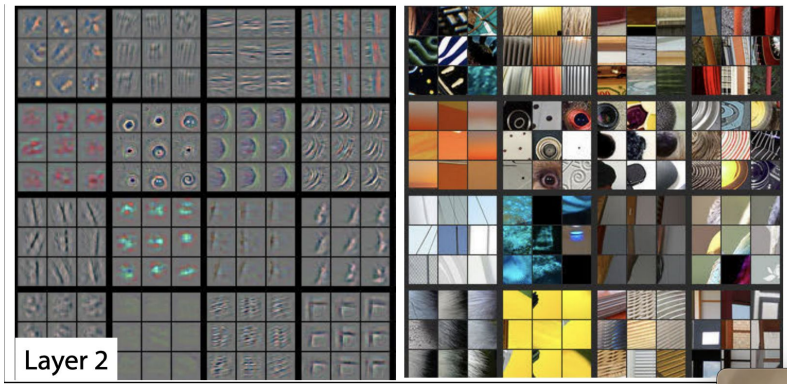
- Neuron = a single dimension of a hidden state representation
- Line of work traditionally does not consider structure



What pattern in the inputs will fire a neuron (i.e., cause high values at a particular dimension)?

Interpreting Neurons

Early research: visualizing salient input features



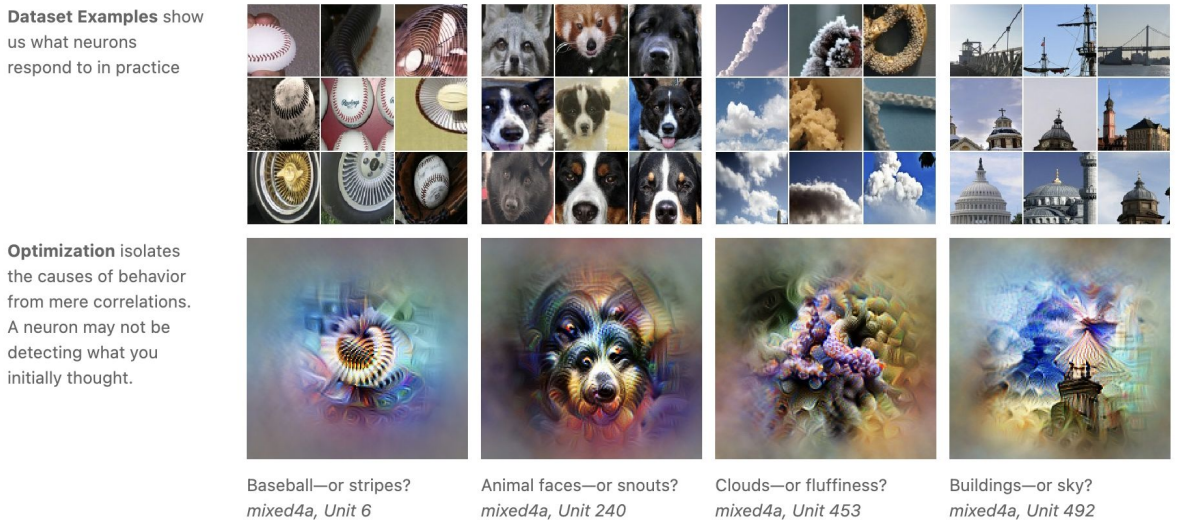
[Zeiler & Fergus, CVPR 2014]

```
/* Duplicate LSM field information.  The lsm_rule is opaque, so
 * re-initialized. */
static inline int audit_dupe_lsm_field(struct audit_field *df,
                                       struct audit_field *sf)
{
  int ret = 0;
  char *lsm_str;
  /* our own copy of lsm_str */
  lsm_str = kstrdup(sf->lsm_str, GFP_KERNEL);
  if (unlikely(!lsm_str))
    return -ENOMEM;
  df->lsm_str = lsm_str;
  /* our own (refreshed) copy of lsm_rule */
  ret = security_audit_rule_init(df->type, df->op, df->lsm_str,
                                 (void **)&df->lsm_rule);
  /* Keep currently invalid fields around in case they
   * become valid after a policy reload. */
  if (ret == -EINVAL) {
    pr_warn("audit rule for LSM \'%s\' is invalid\n",
            df->lsm_str);
    ret = 0;
  }
  return ret;
}
```

[Karpathy et al., 2015]

Interpreting Neurons

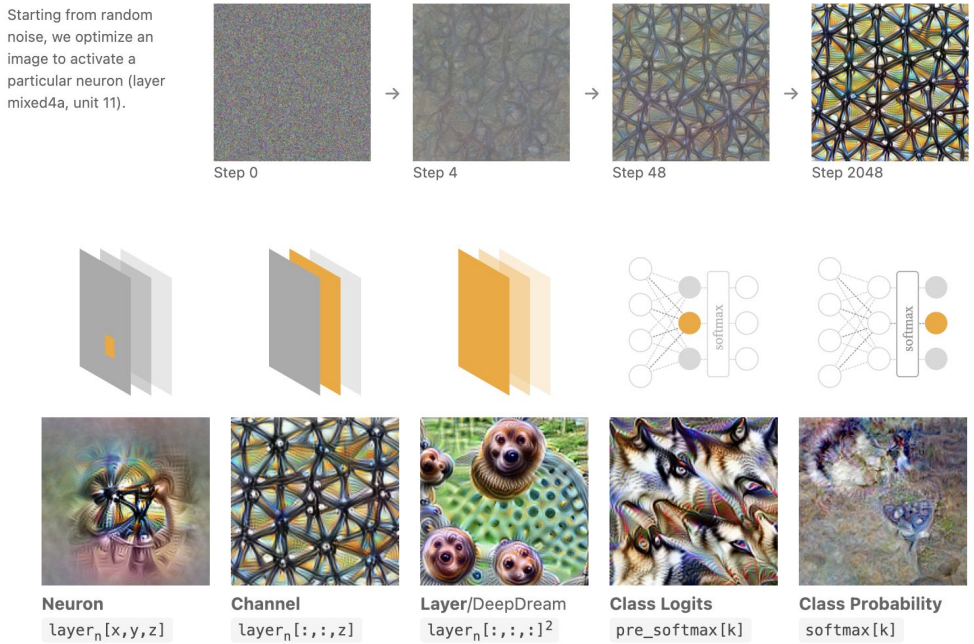
Early research: synthetically constructing maximally-activating instances (“feature visualization”)



[Olah et al. 2017]

Interpreting Neurons

Early research: synthetically constructing maximally-activating instances ("feature visualization")



[Olah et al. 2017]

Interpreting Neurons

More readable interpretation: what **concept** fires a neuron

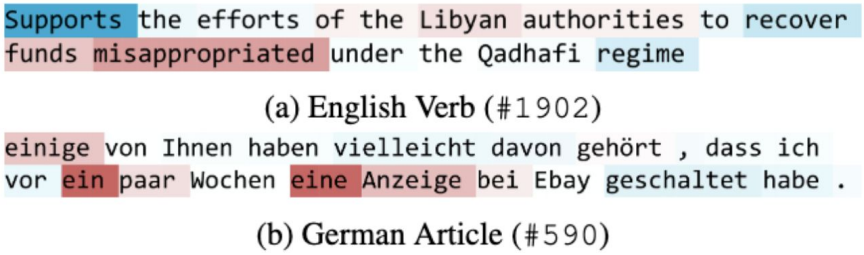
[Bau et al., CVPR 2017]

[Mu & Andreas, NeurIPS 2020]

Interpreting Neurons of NLP Models

More readable interpretation: what **concept** fires a neuron

Activations of neurons for certain properties



Layer14, Unit 224: sure, know, aware

- Are you sure you are aware of our full potential?
- They know that and we know that.
- I am sure you will understand.
- I am sure you will do this.
- I am confident that we will find a solution.

[Bau et al., ICLR 2019]

[Mu & Andreas, NeurIPS 2020]

[Dalvi et al., AAAI 2019]

[Na et al., ICLR 2019]

Pitfalls of Neuron-Level Analysis in NLP

- *"DNNs are **distributed in nature**, which encourages groups of neurons to work together to learn a concept. The current analysis methods, at large, **ignore interaction between neurons** while discovering neurons with respect to a concept."*

[Sajjad et al., TACL 2022]

- *"Since the ranking space is too large (768! in BERT's case), these methods provide **approximations to the problem and are non-optimal**."*

[Antverg & Belinkov, ICLR 2022]

Pitfalls of Visualization/Looking at Examples

- Humans are biased towards simple and clear concepts.

- *"What is the meaning behind the song ""Angel"" by Eric Clapton?"*
- *"What's the meaning of Johnny Cash's song ""King of the Hill""?"*
- *"What is the meaning behind the Tears for Fears song ""Mad World"", such as the lyric, ""All around me are familiar faces""?"*

Song titles? Syntactic sentence structure?

[Bolukbasi et al 2021, Olah et al 2020]

Pitfalls of Visualization/Looking at Examples

- Humans are biased towards simple and clear concepts.

- *”What is the meaning behind the song ””Angel”” by Eric Clapton?”*
- *”What’s the meaning of Johnny Cash’s song ””King of the Hill””?”*
- *”What is the meaning behind the Tears for Fears song ””Mad World””, such as the lyric, ””All around me are familiar faces””?”*

Song titles? Syntactic sentence structure?

- *On 16 June 2006, it was announced that Everton had entered into talks with Knowsley Council and Tesco over the possibility of building a new 55,000 seat stadium, ex-pandable to over 60,000, in Kirkby.*
- *On 15 September 1940, known as the Battle of Britain Day, an RAF pilot, Ray Holmes of No. 504 Squadron RAF rammed a German bomber he believed was going to bomb the Palace.*
- *On 20 August 2010, Queen’s manager Jim Beach put out a Newsletter stating that the band had signed a new contract with Universal Music.*

Historical events? Sentences with dates at the beginning?

- **Polysemanticity**: neurons “respond to multiple unrelated inputs”

[Bolukbasi et al 2021, Olah et al 2020]

Outline

1. **Neuron-level interpretability**
    a. **Sparse Autoencoders**
2. Causality Background and Methods
    a. Causal Mediation
    b. Activation Patching/Causal Tracing and Path Patching
    c. High-level causal graphs
    d. Other methods
3. What is mechanistic interpretability?
4. Methods Leveraging Language Model Strengths
    a. Transformer Residual Stream and Linear Structure
    b. Vocabulary projection
    c. Decoding Natural Language Explanations from Representations
5. Conclusion + Q&A

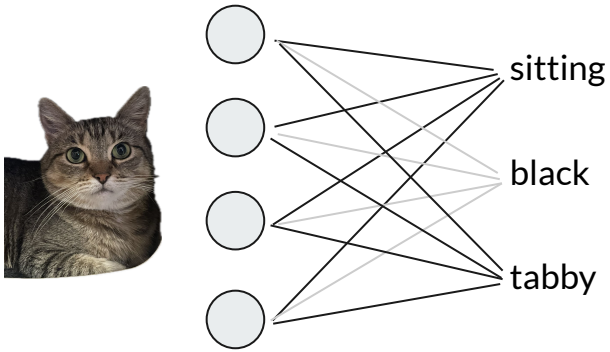
Linear Combinations of Neurons as Concepts

Individual neuron-level interpretations are typically not precise:
many neurons respond to mixtures of concepts

[Anthropic, 2023]

Linear Combinations of Neurons as Concepts

Individual neuron-level interpretations are typically not precise:
many neurons respond to mixtures of concepts



Hypothesis

Neurons together (as opposed to individual neurons) respond to concepts

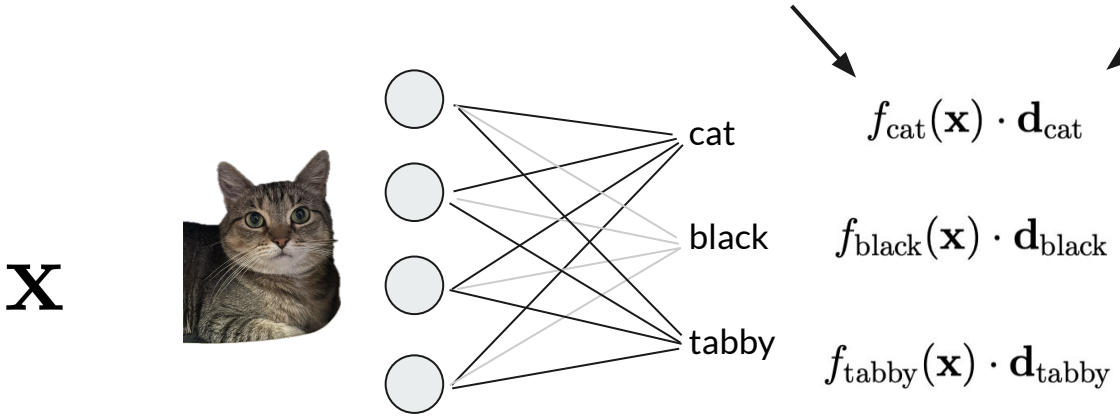
Neuron activations can be decomposed into linear combinations of concept directions (called features)

[Anthropic, 2023]

Decomposing Activations

scalar:
strength of the feature

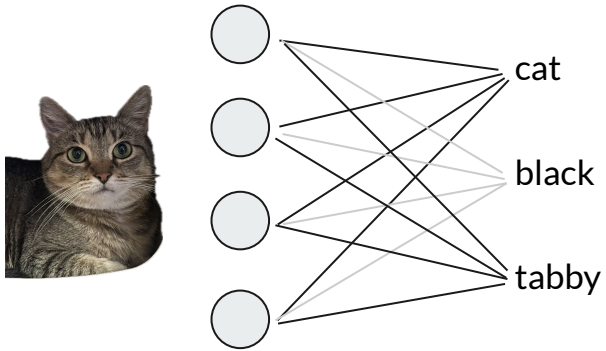
vector:
direction of the feature



$$\mathbf{x} \xrightarrow{\text{decompose}} f_{\text{cat}}(\mathbf{x}) \cdot \mathbf{d}_{\text{cat}} + f_{\text{black}}(\mathbf{x}) \cdot \mathbf{d}_{\text{black}} + f_{\text{tabby}}(\mathbf{x}) \cdot \mathbf{d}_{\text{tabby}}$$

[Anthropic, 2023]

Decomposing Activations with Sparse Autoencoders



Sparsity: for $\mathbf{x}$, we expect only a small number of feature **c** is activated ($f_c(\mathbf{x}) > 0$)

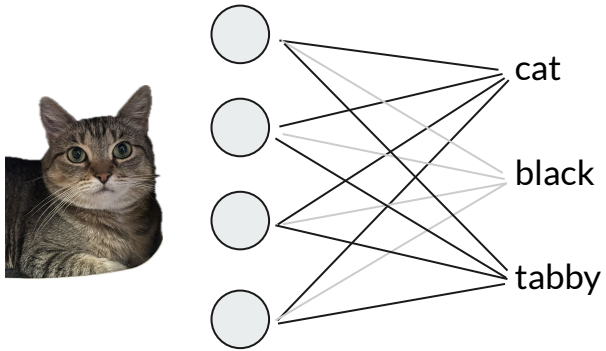
$$\mathbf{x} \approx \mathbf{b} + \sum_c f_c(\mathbf{x})\mathbf{d}_c$$

using **Sparse Autoencoders** to find decompositions

$f_c(\mathbf{x})$: encoder parameters $\mathbf{d}_c$: decoder parameters

[Anthropic, 2023]

Decomposing Activations with Sparse Autoencoders



Sparsity: for $\mathbf{x}$, we expect only a small number of feature **c** is activated ($f_c(\mathbf{x}) > 0$)

$$\mathbf{x} \approx \mathbf{b} + \sum_c f_c(\mathbf{x})\mathbf{d}_c$$

using **Sparse Autoencoders** to find decompositions

$f_c(\mathbf{x})$: encoder parameters $\quad \mathbf{d}_c$: decoder parameters

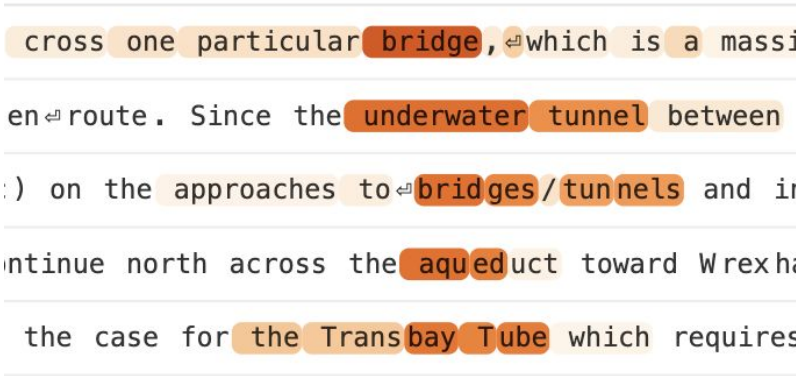
reconstruction loss **sparsity**

Loss: $$\mathbb{E}_x \left[\|x - \hat{x}\|_2^2 + \lambda \sum_c f_c(x) \right]$$

[Anthropic, 2023]

Notes on Learned Features

Example: 1M/3 **Transit infrastructure**



Intervening on feature activations has an influence on behavior

Default output gives reasonable navigation directions

```
Human: What's the best way to get to the grocery
store down the street? Be brief.

Assistant: 1. Walk.
```

with Transit infrastructure F#1M/3
clamped to 5× its max
It confabulates a bridge

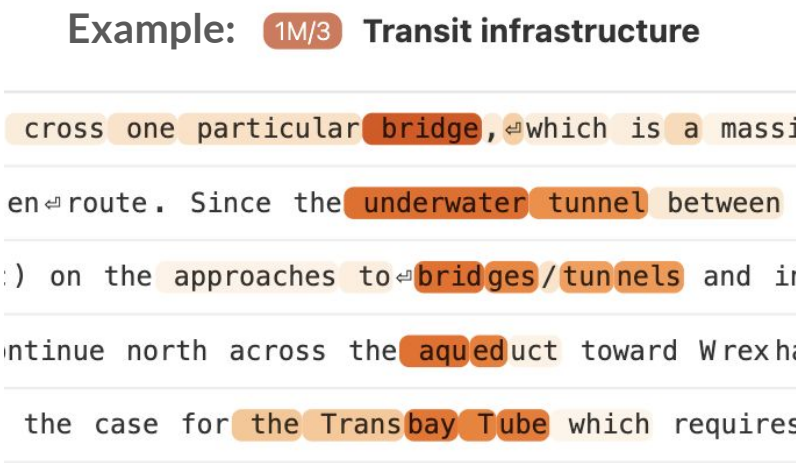
```
Human: What's the best way to get to the grocery
store down the street? Be brief.

Assistant: 1. Walk across the bridge.
```

[Anthropic, 2024]

[Anthropic, 2023]

Notes on Learned Features



Intervening on feature activations has an influence on behavior

Feature activations are more specific than neurons

- “upon manual inspection of a random sample of 50 neurons and features each, the neurons appear significantly less interpretable than the features, typically activating in multiple unrelated contexts..”

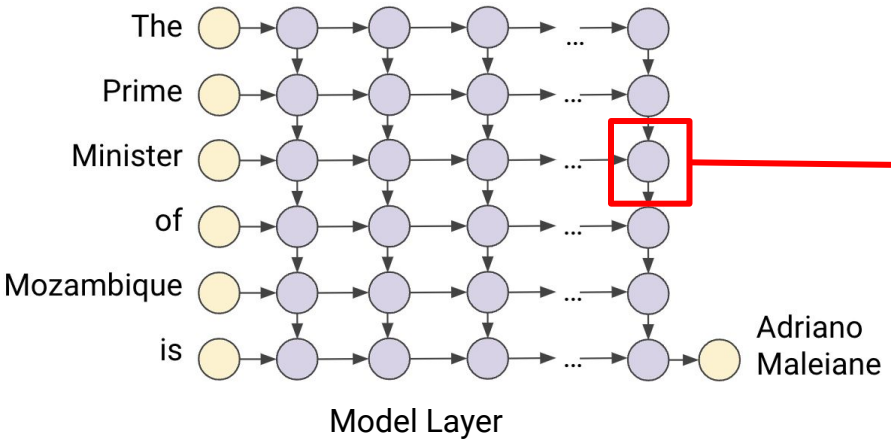
[Anthropic, 2024]

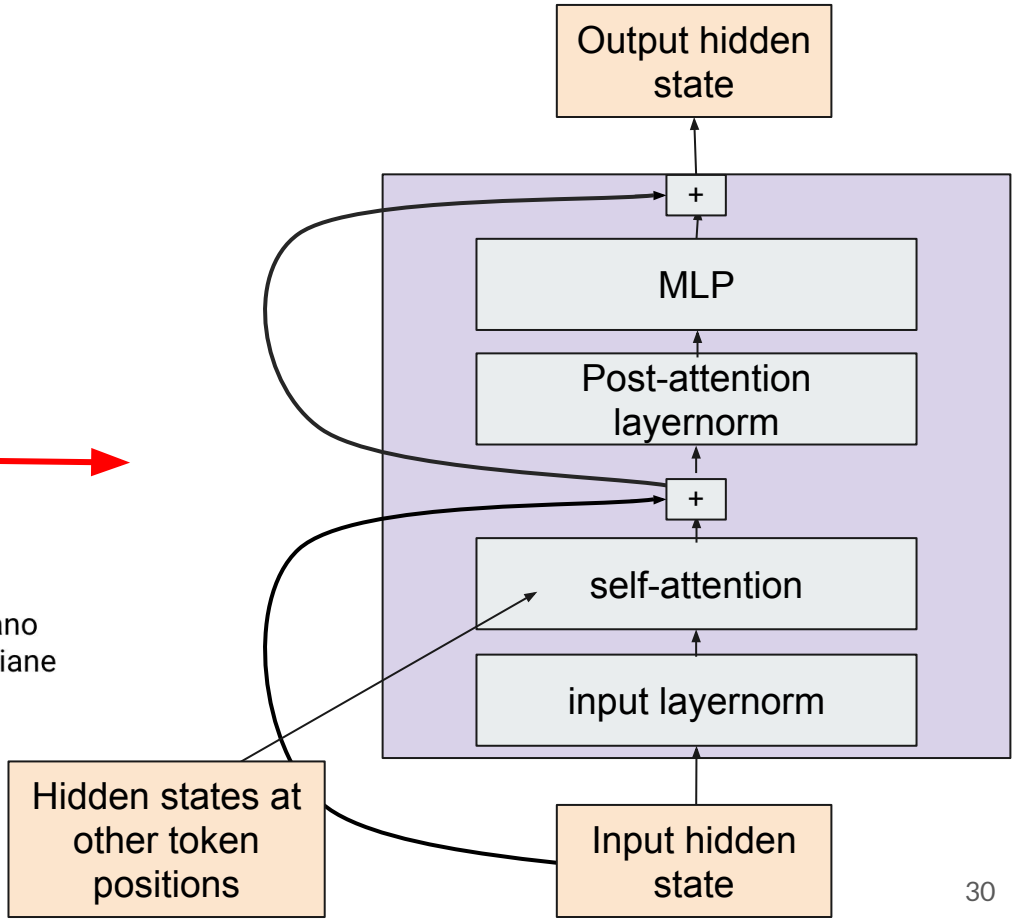
[Anthropic, 2023]

Outline

1. Neuron-level interpretability
    a. Sparse Autoencoders
2. **Causality Background and Methods**
    a. Causal Mediation
    b. Activation Patching/Causal Tracing and Path Patching
    c. High-level causal graphs
    d. Other methods
3. What is mechanistic interpretability?
4. Methods Leveraging Language Model Strengths
    a. Transformer Residual Stream and Linear Structure
    b. Vocabulary projection
    c. Decoding Natural Language Explanations from Representations
5. Conclusion + Q&A

Transformer Block





Causal Mediation

Total Effect = ab + c’

Indirect Effect = ab

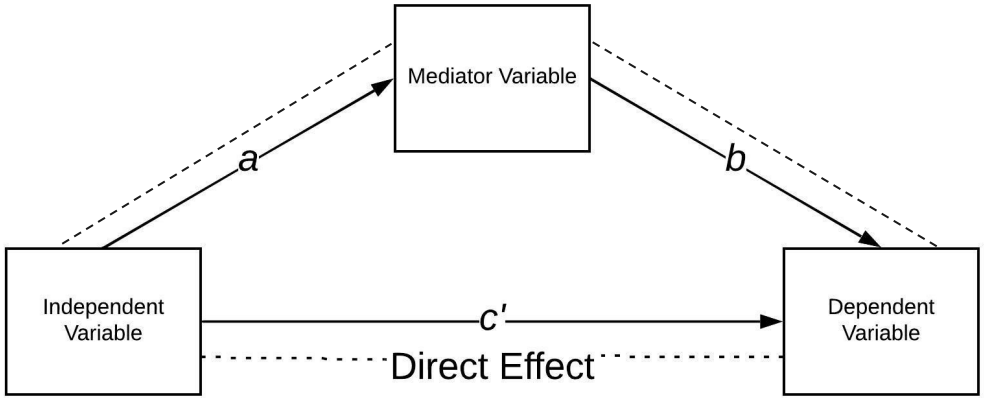
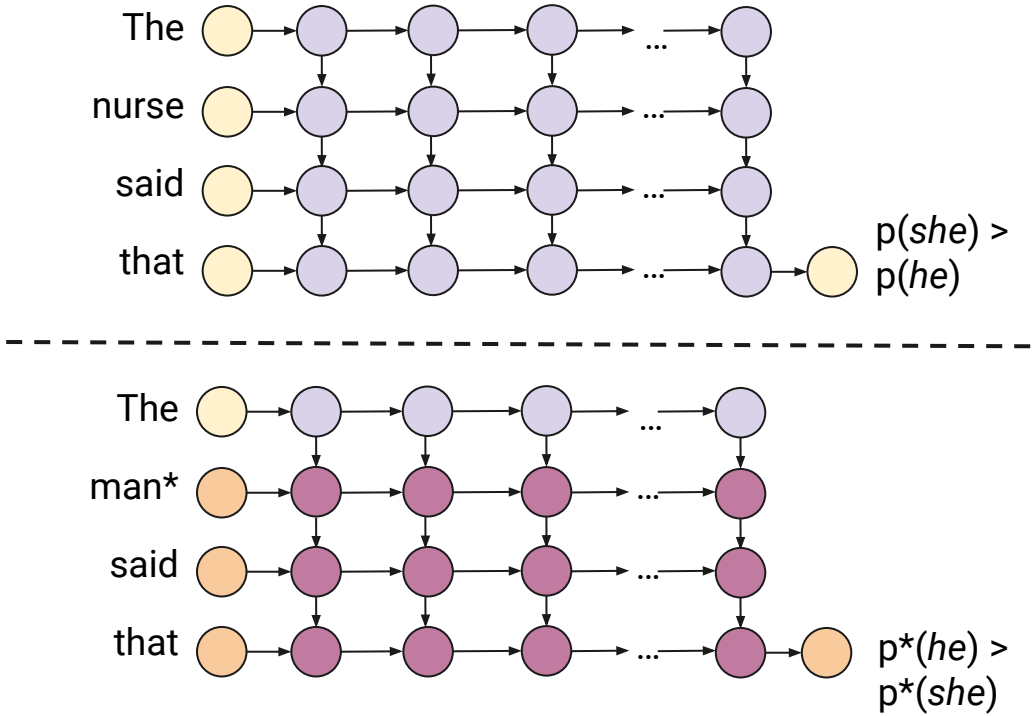


Image: Wikipedia

Activation Patching/Causal Tracing

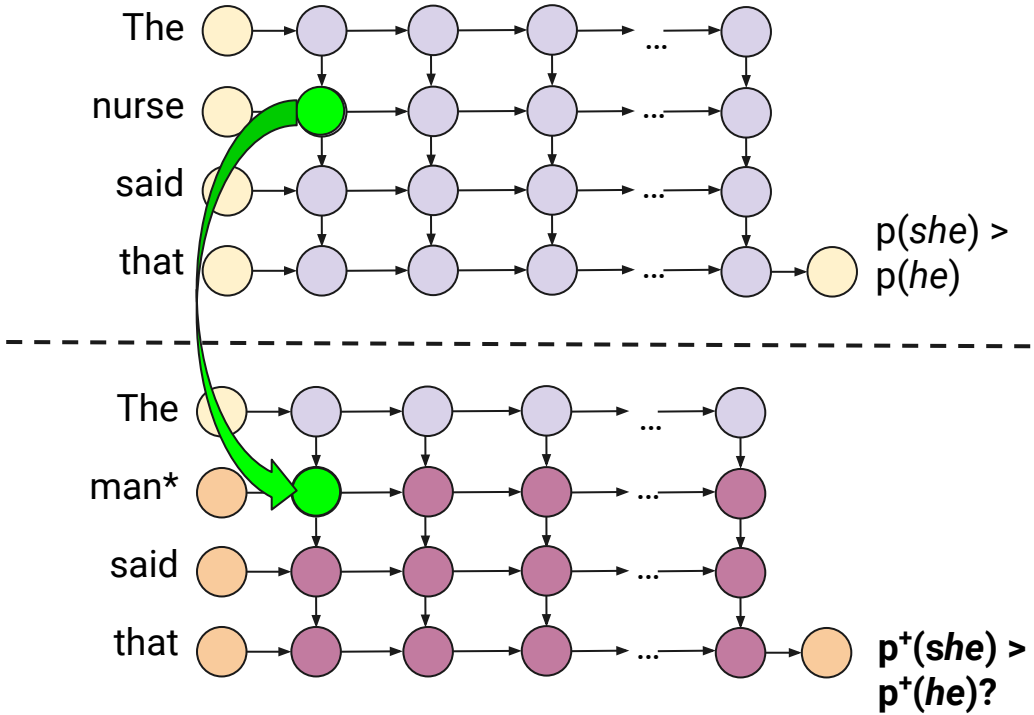
- Run inference through the network twice
- Measure the change in probabilities of the tokens of interest



[Vig et al. 2020]

Activation Patching/Causal Tracing

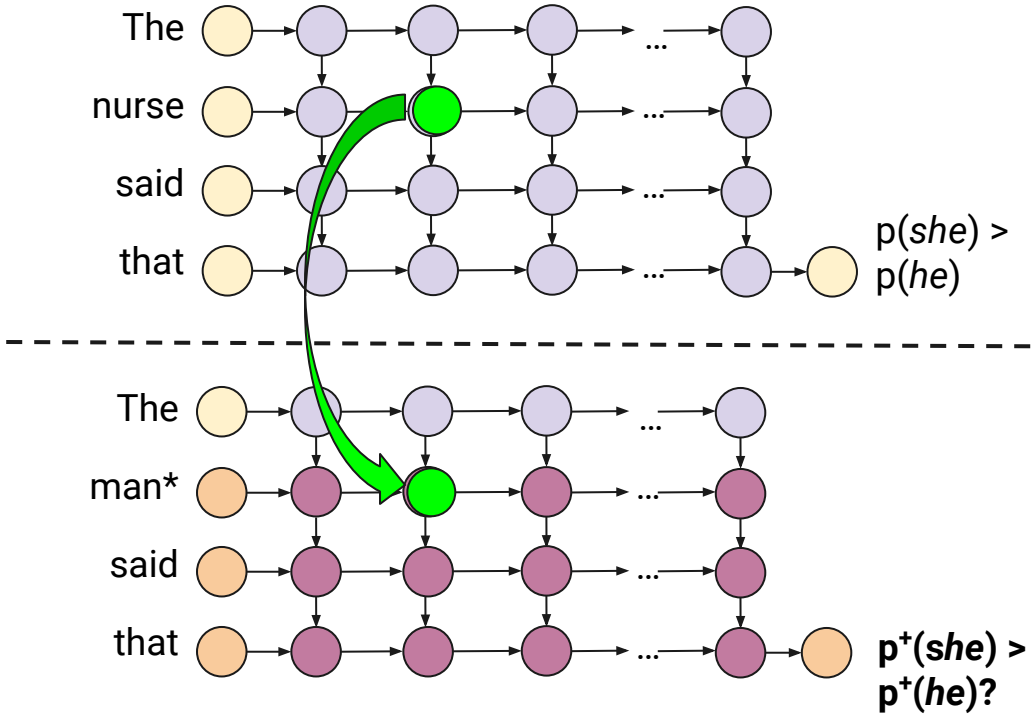
- Run inference through the network twice
- Measure the change in probabilities of the tokens of interest
- *Patch* in states from one inference run into another
- Observe how probabilities change → the most important hidden states will have the largest effect in "restoring" the probabilities of the run that is being patched in.



[Vig et al. 2020]

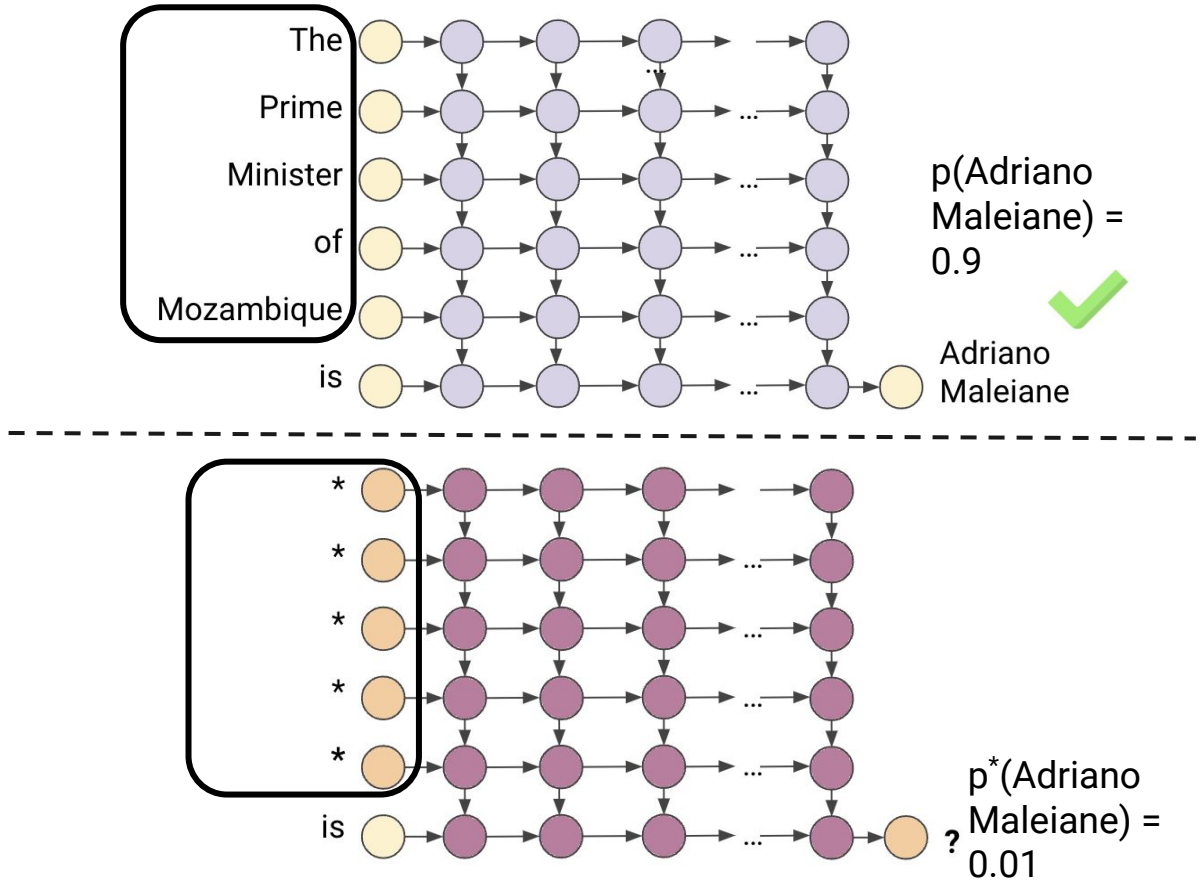
Activation Patching/Causal Tracing

- Run inference through the network twice
- Measure the change in probabilities of the tokens of interest
- *Patch* in states from one inference run into another
- Observe how probabilities change → the most important hidden states will have the largest effect in "restoring" the probabilities of the run that is being patched in.



[Vig et al. 2020]

Method



Textual Effect =
$p(y) - p^*(y) = 0.89$

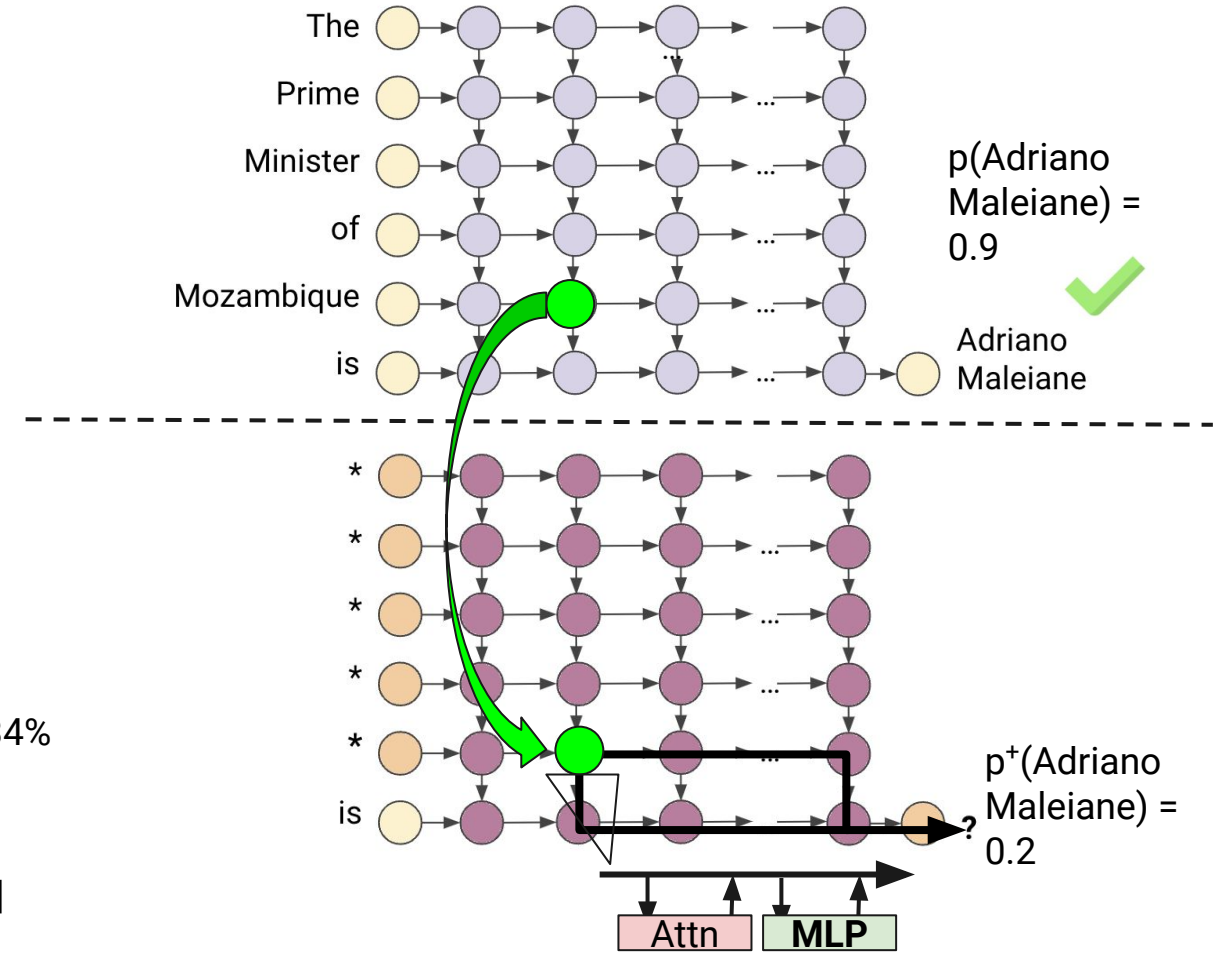
[Meng et al. 2022, Meng et al. 2023]

Method

Textual Effect =
$p(y) - p^*(y) = 0.89$

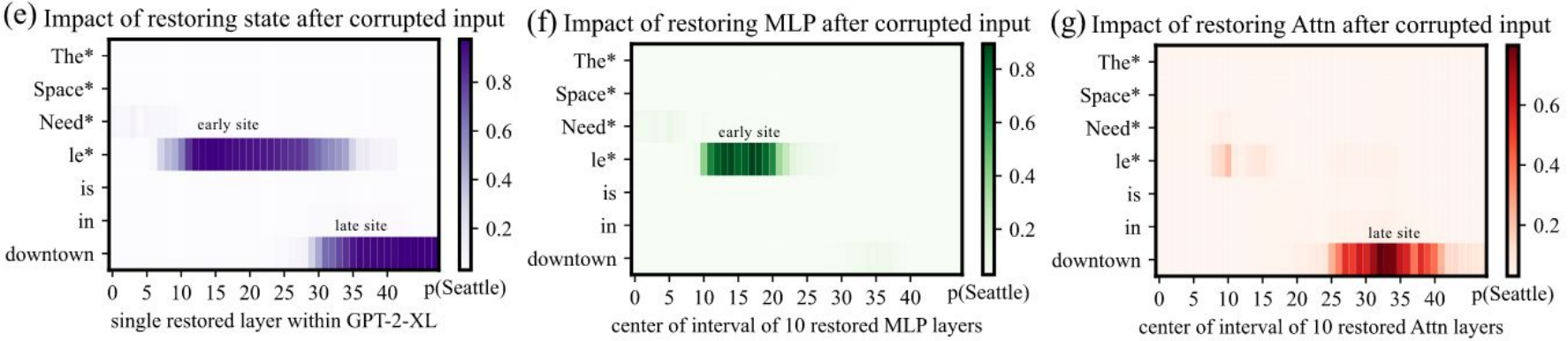
Effect of Repair =
$p^+(y) - p^*(y) = 0.19$

Fractional Effect of Repair =
$[p^+(y) - p^*(y)] / [p(y) - p^*(y)] = 21.34\%$



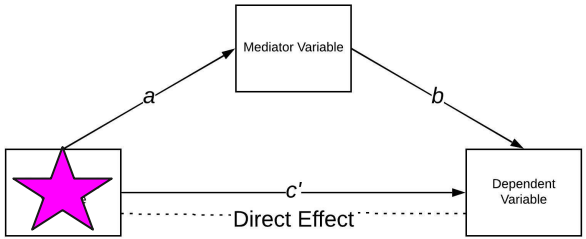
[Meng et al. 2022, Meng et al. 2023]

Results



[Meng et al. 2022, Meng et al. 2023]

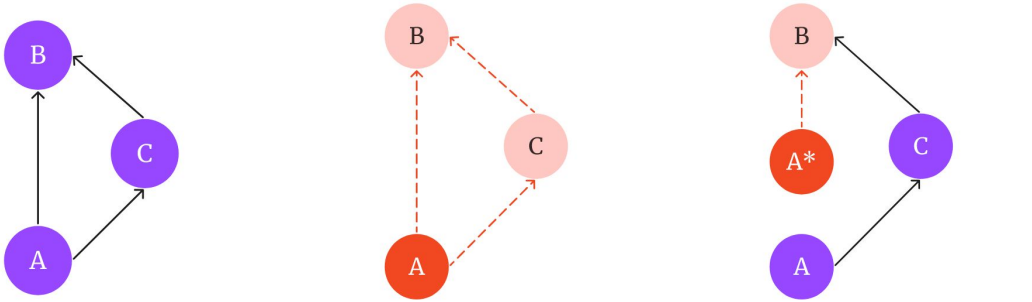
Notes



- Measures **total effect** of hidden state on output
- Method is rather **computationally expensive**
  - Each patch is a separate inference run
  - Also requires two copies of the model to be loaded into memory, generally
- **Strong independence assumption** about individual hidden states or neurons in the network
  - Ideally, one could patch multiple states at once, but enumerating all possible combinations of states is intractable
- More efficient (gradient-based) approximation: “Attribution Patching” [Nanda 2022, Kramár et al. 2024]
- How to design paired instances? [Zhang & Nanda 2024]

Path Patching

- Controls more carefully **which effect** you can measure



(a) Clean forward pass, no intervention

(b) Intervene on A to observe *total* effect on B.

(c) Intervene on the edge A→B to observe *direct* effect on B.

[Goldowsky-Dill et al. 2023]

Slide credit: Lieberum et al. 2023

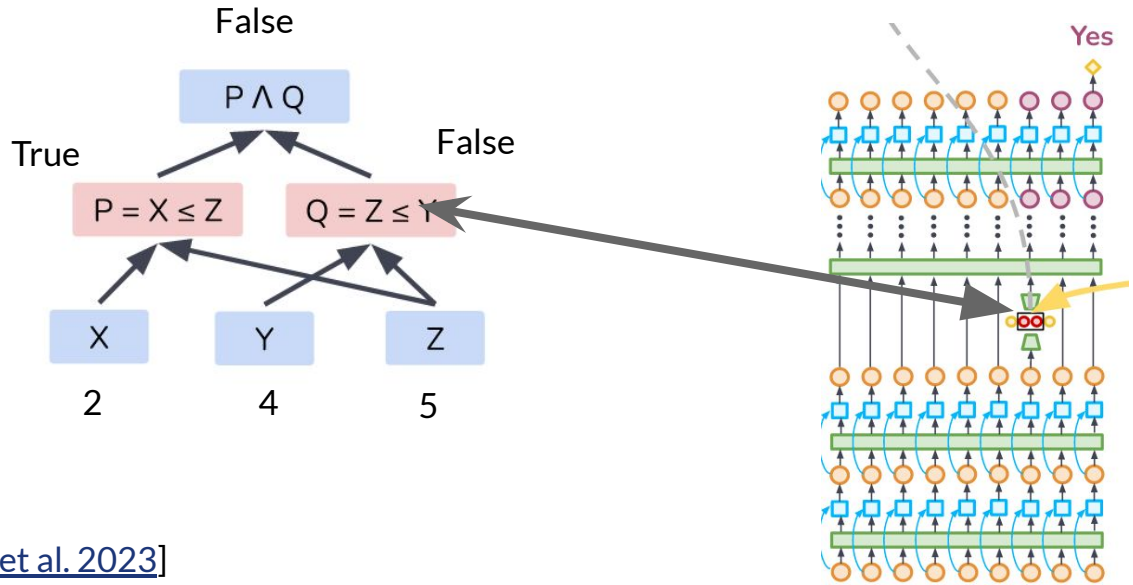
Outline

1. Neuron-level interpretability
    a. Sparse Autoencoders
2. **Causality Background and Methods**
    a. Causal Mediation
    b. Activation Patching/Causal Tracing and Path Patching
    c. High-level causal graphs
    d. Other methods
3. What is mechanistic interpretability?
4. Methods Leveraging Language Model Strengths
    a. Transformer Residual Stream and Linear Structure
    b. Vocabulary projection
    c. Decoding Natural Language Explanations from Representations
5. Conclusion + Q&A

Causal Abstraction

The key idea is to learn a causal graph that maps to the neural network.

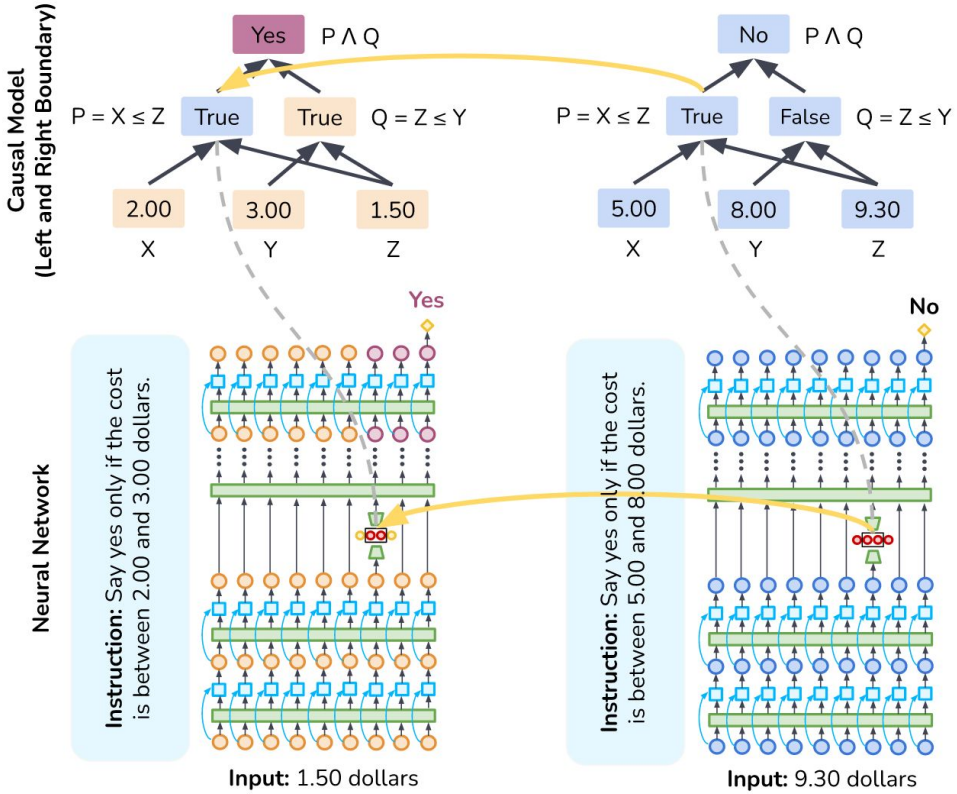
Consider the simple task of determining whether Z is in the interval [X, Y]



[Geiger et al 2023, Wu et al. 2023]

Intervention analyses is essential to causal abstraction

Key intuition: intervention in the low-level neural representations has the same effect as the intervention in the high-level causal graph.

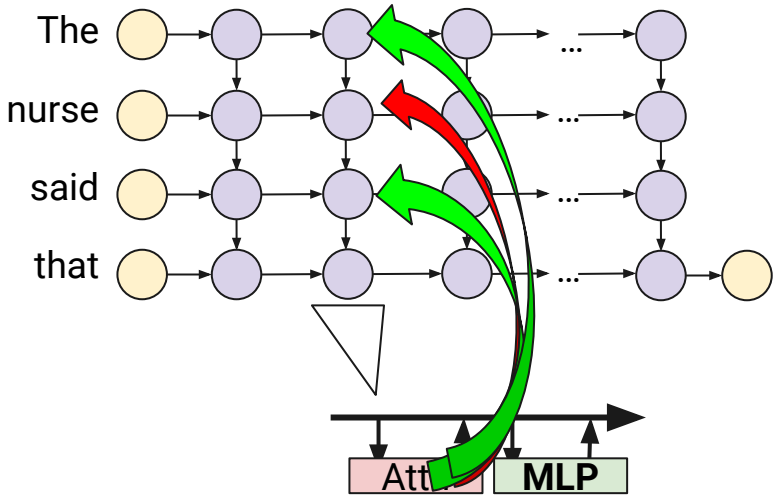


[Geiger et al 2023, Wu et al. 2023]

A couple more notes

- Coming up with a high-level causal graph is highly non-trivial in practice
- A particular style of intervention is considered in this work.

Attention Knockout and Zeroing



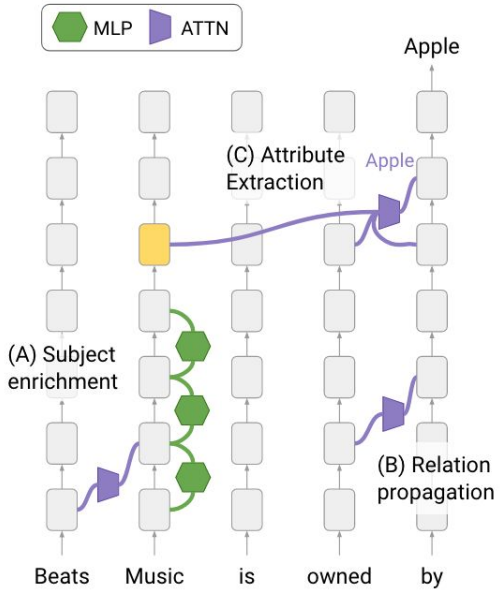


Figure 1: Illustration of our findings: given subject-relation query, a subject representation is constructed via attributes' enrichment from MLP sublayers (A), while the relation propagates to the prediction (B). The attribute is then extracted by the MHSA sublayers (C).

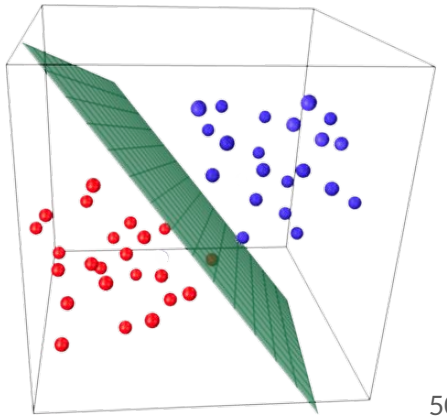
[Mohebbi et al. 2023, Geva et al 2023]

Linear Subspace Projections + Concept Erasure

- Models encode many interpretable concepts linearly.

Linear concept subspace hypothesis: a concept (such as gender) lives in low-dimensional **subspace** within the representation space.

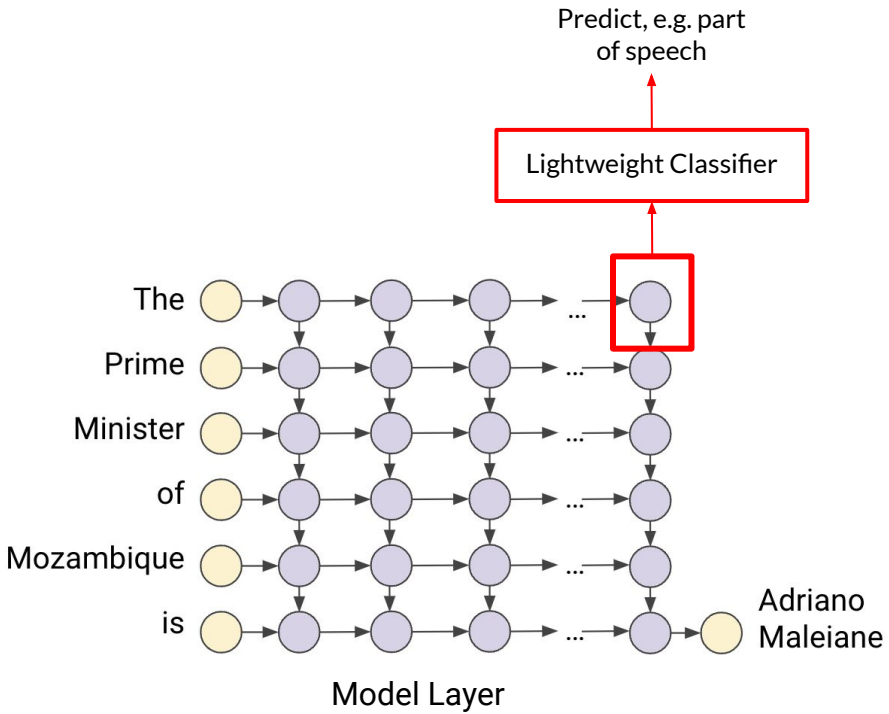
How can we identify the concept subspace?
Once located, can we intervene in its encoding?



[Ravfogel et al 2020, Belrose et al 2023, inter alia]

Slide credit: Shauli Ravfogel

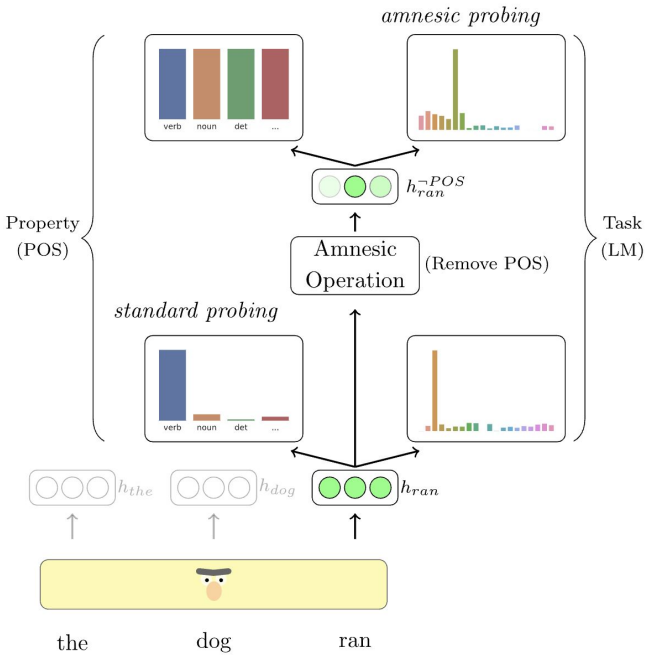
Probing Classifiers



- Goal: measure to what extent a representation **encodes** a property, not **how** that property **is used** to make predictions.
- Generally, supervised: define property of interest **up front**.
- **Not** a causal method.
  - **Correlation != Causation**
- Challenge: restricting the probe's complexity

[Belinkov 2022]

Probing Classifiers



- There are methods that perform causal interventions to measure **how** the property of interest **is used** to make predictions.

[Giulianelli et al. 2018, Elazar et al. 2021]

Outline

1. Neuron-level interpretability
    a. Sparse Autoencoders
2. Causality Background and Methods
    a. Causal Mediation
    b. Activation Patching/Causal Tracing and Path Patching
    c. High-level causal graphs
    d. Other methods
3. **What is mechanistic interpretability?**
4. Methods Leveraging Language Model Strengths
    a. Transformer Residual Stream and Linear Structure
    b. Vocabulary projection
    c. Decoding Natural Language Explanations from Representations
5. Conclusion + Q&A

What is Mechanistic Interpretability?

“reverse engineering the algorithms implemented by neural networks into human-understandable mechanisms, often by examining the weights and activations of neural networks to identify circuits [Cammarata et al., 2020, Elhage et al., 2021] that implement particular behaviors.”

What is Mechanistic Interpretability?

"reverse engineering the algorithms implemented by neural networks into human-understandable mechanisms, often by examining the weights and activations of neural networks to identify circuits [Cammarata et al., 2020, Elhage et al., 2021] that implement particular behaviors."

Desirable outcome of *all* interpretability research: human understanding

Focus of *most* interpretability research: understanding *specific* model behaviors

What is Mechanistic Interpretability?

"reverse engineering the algorithms implemented by neural networks into human-understandable mechanisms, often by examining the weights and activations of neural networks to identify circuits [Cammarata et al., 2020, Elhage et al., 2021] that implement particular behaviors."

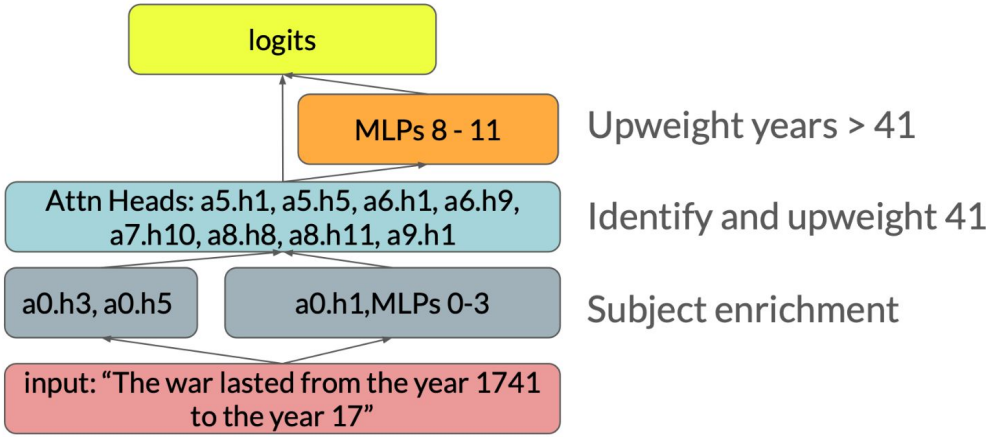
Format of the explanation

Finding a subset of a network that **traces through the entire network (from starting representation to prediction).**

Circuits

- Note: this has strong resemblances to sparse sub-network finding in the efficiency literature, but the methods employed to find them differ (also, no retraining of circuits is done)

Transformer circuits localize and characterize transformer LM behavior in a (small) set of components of the model.



Hanna et al., 2023

Slide credit: Michael Hanna

What is mechanistic interpretability?

- It is **inherently causal.**
  - NB! This is **not** how most people use the terminology today.
- It is **not the only** set of causal interpretability methods.
- **Traces through the entire network (from starting representation to prediction).**
- Evaluation:
  - 1) **Faithfulness**: the circuit or subnetwork should be able to *sufficiently* **replicate the full network** on the behavior of interest
  - 2) **Minimality**: obviously, smaller circuits/subnetworks are better

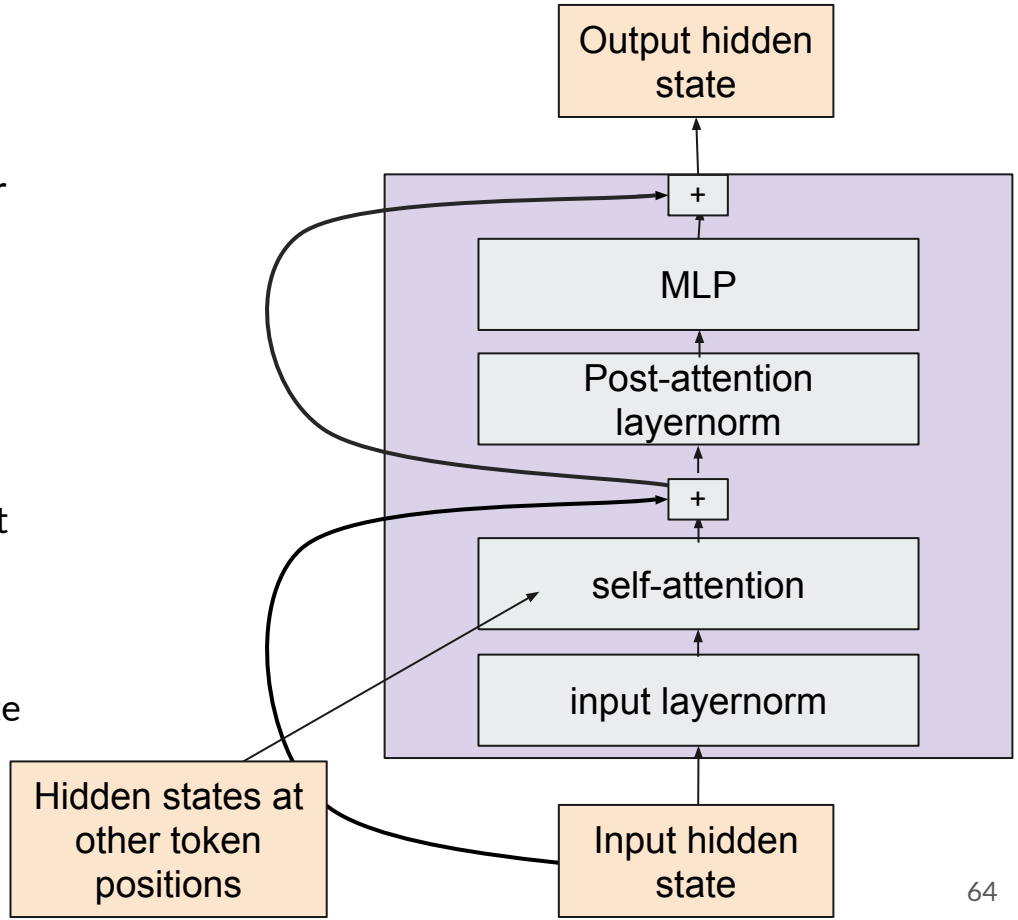
Outline

1. Neuron-level interpretability
    a. Sparse Autoencoders
2. Causality Background and Methods
    a. Causal Mediation
    b. Activation Patching/Causal Tracing and Path Patching
    c. High-level causal graphs
    d. Other methods
3. What is mechanistic interpretability?
4. **Methods Leveraging Language Model Strengths**
    a. Transformer Residual Stream and Linear Structure
    b. Vocabulary projection
    c. Decoding Natural Language Explanations from Representations
5. Conclusion + Q&A

[Elhage et al. 2021]

Transformer Residual Stream and Linear Structure

- Transformers have a surprising amount of linear structure due to residual connections
- Nonlinearities only occur in two places:
  - Applications of Softmax
    - when computing attention patterns
    - When converting logits to probits at final layer
  - In the MLP functions
- MLP and MHSA functions “read from” and “write to” residual stream to promote/demote certain tokens in output distribution.



[Elhage et al. 2021]

Transformer Residual Stream and Linear Structure

For input token embedding $\mathbf{x}_0 \in \mathbb{R}^d$, the output $\mathbf{x}_\ell \in \mathbb{R}^d$ of layer ℓ is defined as (for $\ell \in [1, L]$):

$$\mathbf{x}_\ell = \mathbf{x}_{\ell-1} + \mathrm{MHSA}_{\theta_\ell}\left(\mathrm{LN}(\mathbf{x}_{\ell-1})\right) + \mathrm{FFN}_{\theta_\ell}\left(\mathbf{x}_{\ell-1} + \mathrm{MHSA}_{\theta_\ell}\left(\mathrm{LN}(\mathbf{x}_{\ell-1})\right)\right)$$

- Output of previous layer = input to current layer
- Multi-head self-attention
- Layer norm (or some other input normalization scheme)
- Feed-forward network (MLP)
- Vector addition establishes residual connections

[Elhage et al. 2021]

[Elhage et al. 2021]

Transformer Residual Stream and Linear Structure

For input token embedding $\mathbf{x}_0 \in \mathbb{R}^d$, the output $\mathbf{x}_\ell \in \mathbb{R}^d$ of layer ℓ is defined as (for $\ell \in [1, L]$):

$$\mathbf{x}_\ell = \mathbf{x}_{\ell-1} + \mathrm{MHSA}_{\theta_\ell}\left(\mathrm{LN}(\mathbf{x}_{\ell-1})\right) + \mathrm{FFN}_{\theta_\ell}\left(\mathbf{x}_{\ell-1} + \mathrm{MHSA}_{\theta_\ell}\left(\mathrm{LN}(\mathbf{x}_{\ell-1})\right)\right)$$

Output of previous layer = input to current layer

Multi-head self-attention

Layer norm (or some other input normalization scheme)

Feed-forward network (MLP)

Vector addition establishes residual connections

$$\mathbf{x}_L = \mathbf{x}_0 + \sum_{\ell=0}^{L-1}\left[\mathrm{MHSA}_{\theta_{\ell+1}}\left(\mathrm{LN}(\mathbf{x}_\ell)\right) + \mathrm{FFN}_{\theta_{\ell+1}}\left(\mathbf{x}_\ell + \mathrm{MHSA}_{\theta_{\ell+1}}\left(\mathrm{LN}(\mathbf{x}_\ell)\right)\right)\right]$$

Transformer Residual Stream and Linear Structure

- Self-attention output further linearly decomposable into individual attention heads:

$$W_O^{(\ell)} \in \mathbb{R}^{d \times d}$$

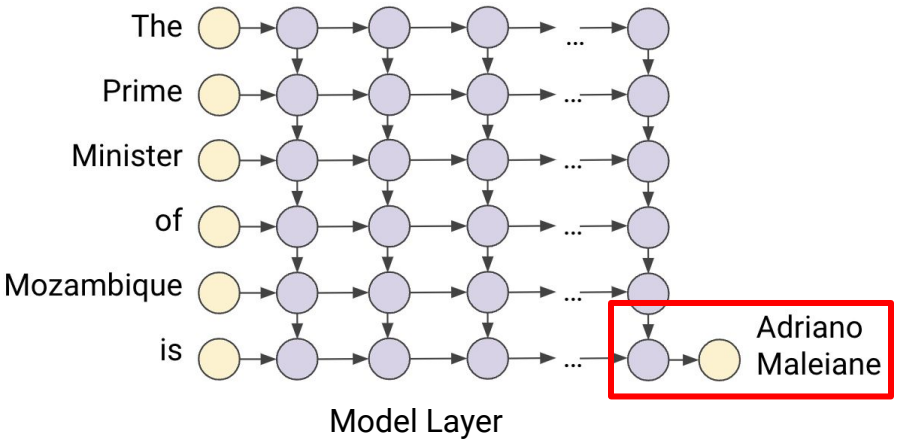
Output weight matrix for MHSA function at layer ℓ

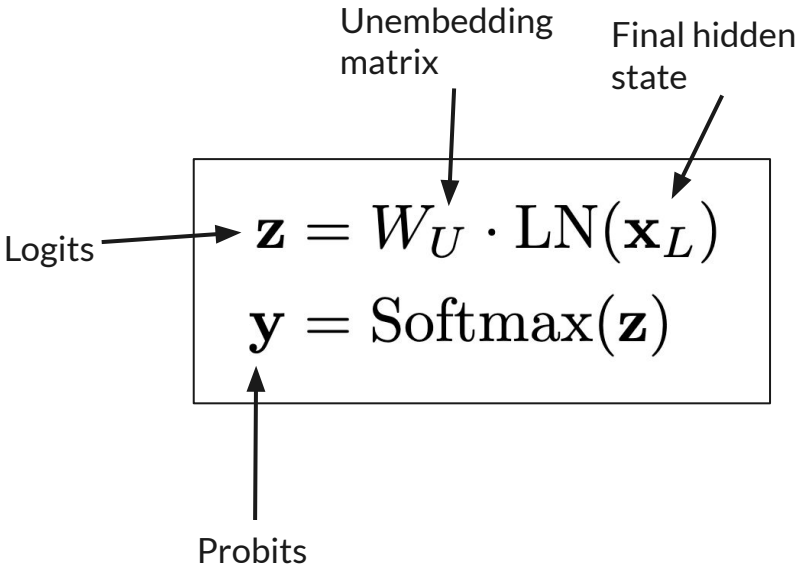
$$\mathbf{MHSA}_{\theta_\ell}\left(\mathbf{LN}(\mathbf{x}_{\ell-1})\right) = \sum_{h=1}^{H} W_{O,h}^{(\ell)} \cdot \mathbf{Att}_h^{(\ell)}\left(\mathbf{LN}(\mathbf{x}_{\ell-1})\right)$$

Output of individual attention head at layer ℓ

Specific columns of W_O that weigh the contribution of head h

Transformer Residual Stream and Linear Structure





Implications: Direct Additive Contributions

- Each hidden state output by a {attention head, MHSA function, FFN function, or full Transformer block} has a **direct additive contribution to the final hidden state of the model**
- And, by distributivity of vector addition and vector-matrix multiplication, thus has a **direct additive contribution** to the **final logits**.

$$\mathbf{x}_L = \mathbf{x}_0 + \sum_{\ell=0}^{L-1} \left[\mathrm{MHSA}_{\theta_{\ell+1}}\left(\mathrm{LN}(\mathbf{x}_\ell)\right) + \mathrm{FFN}_{\theta_{\ell+1}}\left(\mathbf{x}_\ell + \mathrm{MHSA}_{\theta_{\ell+1}}\left(\mathrm{LN}(\mathbf{x}_\ell)\right)\right) \right]$$

$$\mathbf{z} = W_U \cdot \mathrm{LN}\left(\mathbf{x}_0 + \sum_{\ell=0}^{L-1} \left[\mathrm{MHSA}_{\theta_{\ell+1}}\left(\mathrm{LN}(\mathbf{x}_\ell)\right) + \mathrm{MLP}_{\theta_{\ell+1}}\left(\mathbf{x}_\ell + \mathrm{MHSA}_{\theta_{\ell+1}}\left(\mathrm{LN}(\mathbf{x}_\ell)\right)\right) \right] \right)$$

Direct vs. Indirect Effects

- It's important to note that each hidden state has both a **direct linear** and **indirect nonlinear** contribution to the final hidden state.
- The additive decomposition only applies to **direct** contributions.

$$\mathbf{z} = W_U \cdot \mathrm{LN}\left(\mathbf{x}_0 + \sum_{\ell=0}^{L-1}\left[\mathrm{MHSA}_{\theta_{\ell+1}}\left(\mathrm{LN}(\mathbf{x}_\ell)\right) + \mathrm{MLP}_{\theta_{\ell+1}}\left(\mathbf{x}_\ell + \mathrm{MHSA}_{\theta_{\ell+1}}\left(\mathrm{LN}(\mathbf{x}_\ell)\right)\right)\right]\right)$$

This MHSA function has a direct additive contribution to the logits via this term

But it also has an indirect, nonlinear contribution as input to the MLP function

Direct vs. Indirect Effects

- It's important to note that each hidden state has both a **direct linear** and **indirect nonlinear** contribution to the final hidden state.
- The additive decomposition only applies to **direct** contributions.

$$\mathbf{z} = W_U \cdot \mathrm{LN}\left(\mathbf{x}_0 + \sum_{\ell=0}^{L-1}\left[\mathrm{MHSA}_{\theta_{\ell+1}}\left(\mathrm{LN}(\mathbf{x}_\ell)\right) + \mathrm{MLP}_{\theta_{\ell+1}}\left(\mathbf{x}_\ell + \mathrm{MHSA}_{\theta_{\ell+1}}\left(\mathrm{LN}(\mathbf{x}_\ell)\right)\right)\right]\right)$$

Similarly, the MLP has a direct linear effect through the residual stream, but also becomes input to the (nonlinear) MHSA function at the next layer.

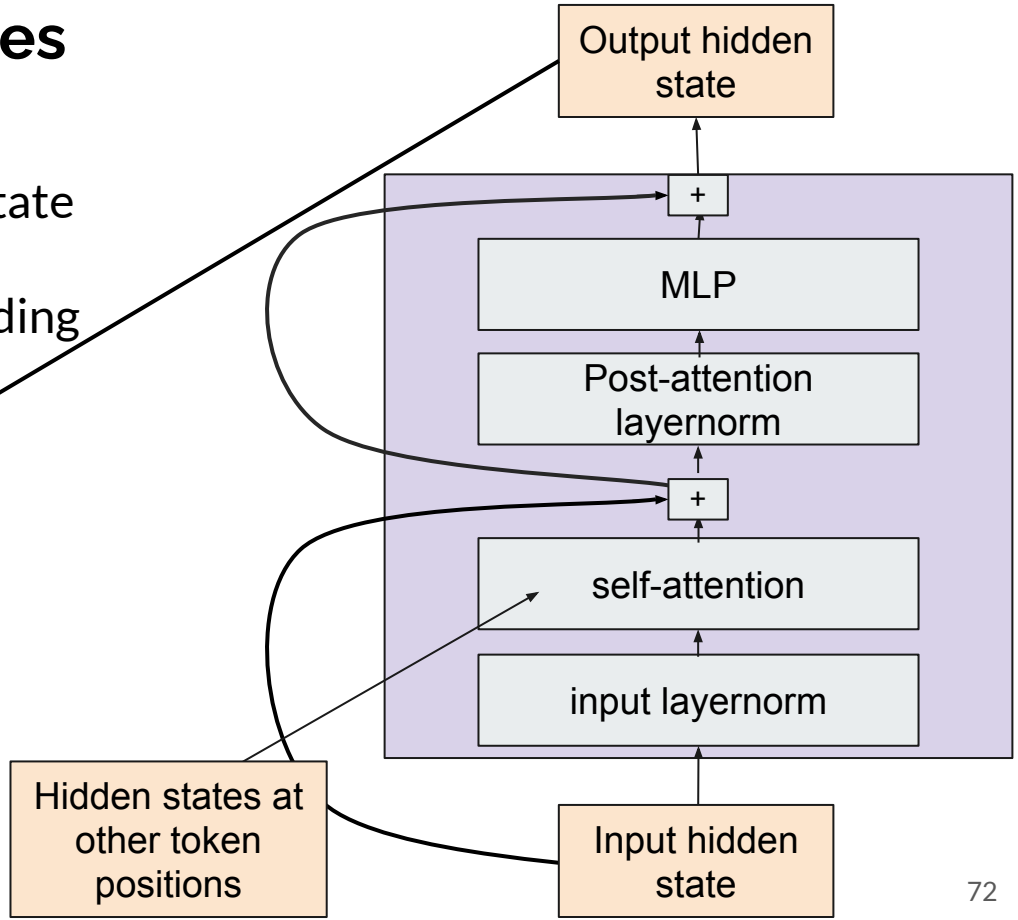
[Nostalgebraist 2020, Geva et al. 2022]

Vocabulary Projection on Transformer Hidden States

- Propose to project each hidden state to the space of probabilities over vocab tokens using the unembedding matrix

$$\mathbf{y} = \text{Softmax}(W_U \cdot \text{LN}(\mathbf{x}_L))$$

Final hidden state - **replace with any d-dimensional hidden state from the network.**



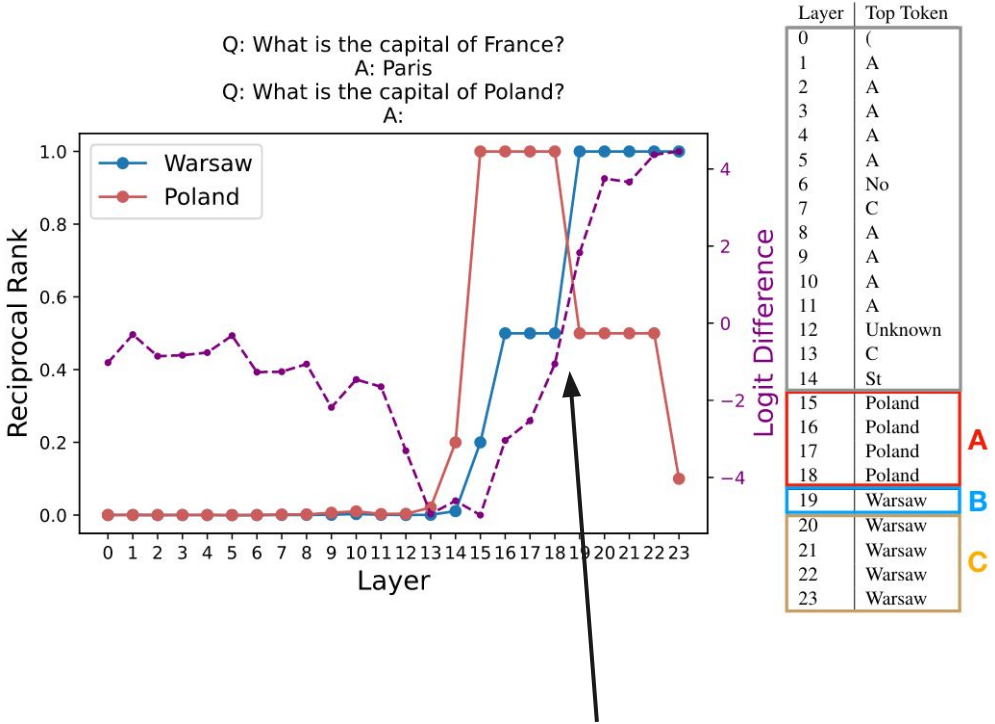
Vocabulary Projection on Transformer Hidden States

		Concept	Sub-update top-scoring tokens
GPT2	$\mathbf{v}^{3}_{1018}$	Measurement semantic	kg, percent, spread, total, yards, pounds, hours
	$\mathbf{v}^{8}_{1900}$	WH-relativizers syntactic	which, whose, Which, whom, where, who, wherein
	$\mathbf{v}^{11}_{2601}$	Food and drinks semantic	drinks, coffee, tea, soda, burgers, bar, sushi
WIKILM	$\mathbf{v}^{1}_{1}$	Pronouns syntactic	Her, She, Their, her, she, They, their, they, His
	$\mathbf{v}^{6}_{3025}$	Adverbs syntactic	largely, rapidly, effectively, previously, normally
	$\mathbf{v}^{13}_{3516}$	Groups of people semantic	policymakers, geneticists, ancestries, Ohioans

Table 1: Example value vectors in GPT2 and WIKILM promoting human-interpretable concepts.

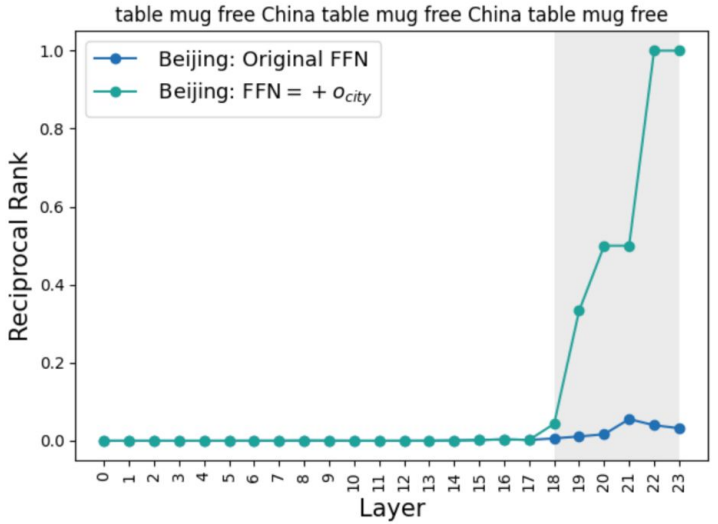
[Merullo et al 2024]

Vocabulary Projection on Transformer Hidden States

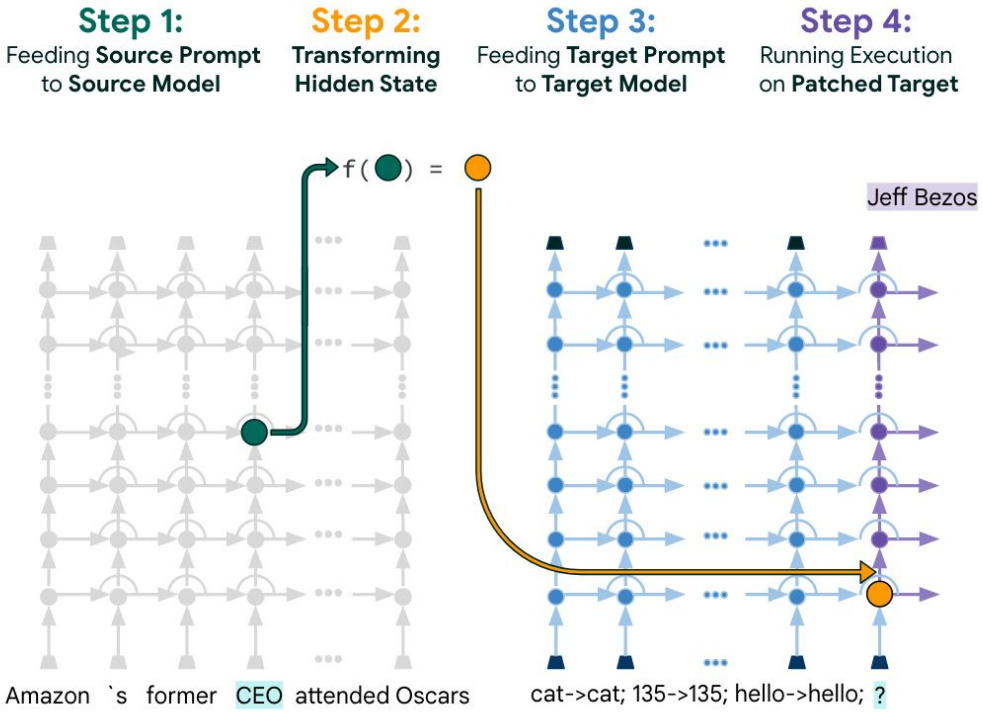


Layer	Top Token
0	(
1	A
2	A
3	A
4	A
5	A
6	No
7	C
8	A
9	A
10	A
11	A
12	Unknown
13	C
14	St
15	Poland
16	Poland
17	Poland
18	Poland
19	Warsaw
20	Warsaw
21	Warsaw
22	Warsaw
23	Warsaw

Validated with causal intervention:



Patchscopes



[Gandeharioun et al. 2024]

[Belrose et al. 2023]

Learning Linear Transformation Matrices

- Propose to project each hidden state to the space of probabilities over vocab tokens using ~~the unembedding matrix~~ a learned weight matrix for each layer

$$\mathbf{y} = \text{Softmax}(W_U \cdot \text{LN}(\mathbf{x}_L))$$

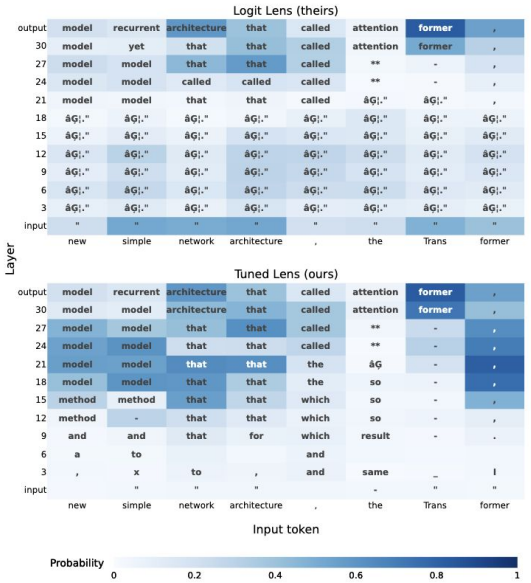


Figure 1. Comparison of our method, the *tuned lens* (bottom), with the "logit lens" (top) for GPT-Neo-2.7B prompted with an except from the abstract of Vaswani et al. (2017). Each cell shows the top-1 token predicted by the model at the given layer and token index. The logit lens fails to elicit interpretable predictions before layer 21, but our method succeeds.

Notes

- Can be thought of as “early exiting” the Transformer block at inference time
- From a causal perspective:
  - (Attempts to) measure *direct* effects
    - How faithful this is depends on the exact application of normalization
    - It is not a causal mediation
- Can’t uncover ways in which hidden states are promoting tokens in other linear (or non-linearly decodable) subspaces
  - I.e., negative results are uninformative
  - Mostly only useful at later layers
- Top-k tokens being coherent: does this just mean that the unembedding matrix is well-formed?

Outline

1. Neuron-level interpretability
    a. Sparse Autoencoders
2. Causality Background and Methods
    a. Causal Mediation
    b. Activation Patching/Causal Tracing and Path Patching
    c. High-level causal graphs
    d. Other methods
3. What is mechanistic interpretability?
4. **Methods Leveraging Language Model Strengths**
    a. Transformer Residual Stream and Linear Structure
    b. Vocabulary projection
    c. Decoding Natural Language Explanations from Representations
5. Conclusion + Q&A

Focus of This Part

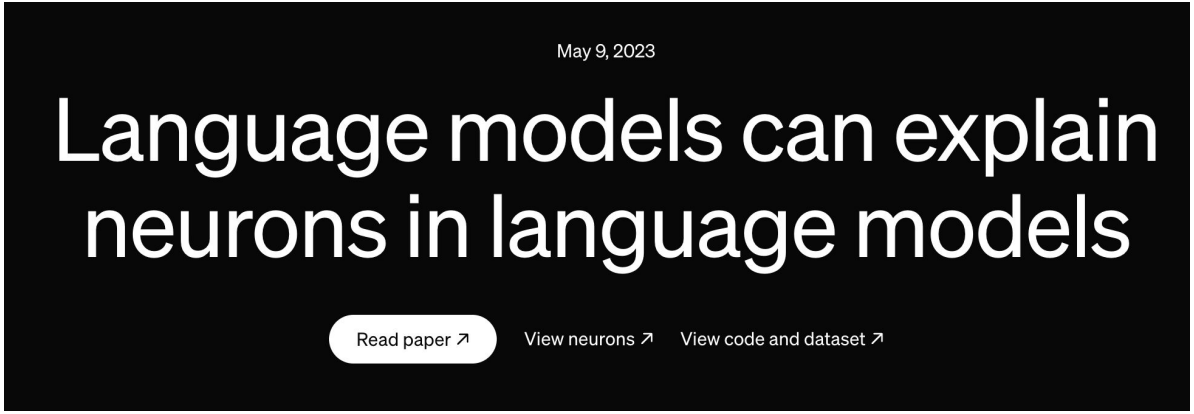
Decoding **natural Language explanations** from neurons

(using LLMs)

Prominent paradigm of using LLMs for automating the process of explaining neurons

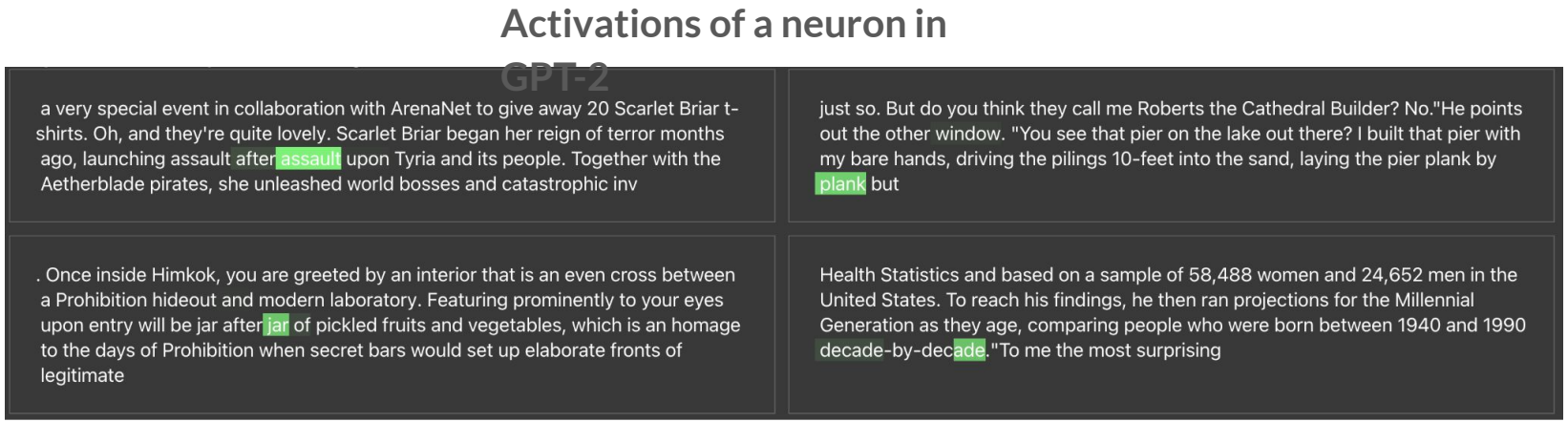
- Step1: Propose hypothesis explanations
- Step2: Verify explanations

Using GPT-4 to Explain Neurons of GPT-2



[OpenAI, 2023]

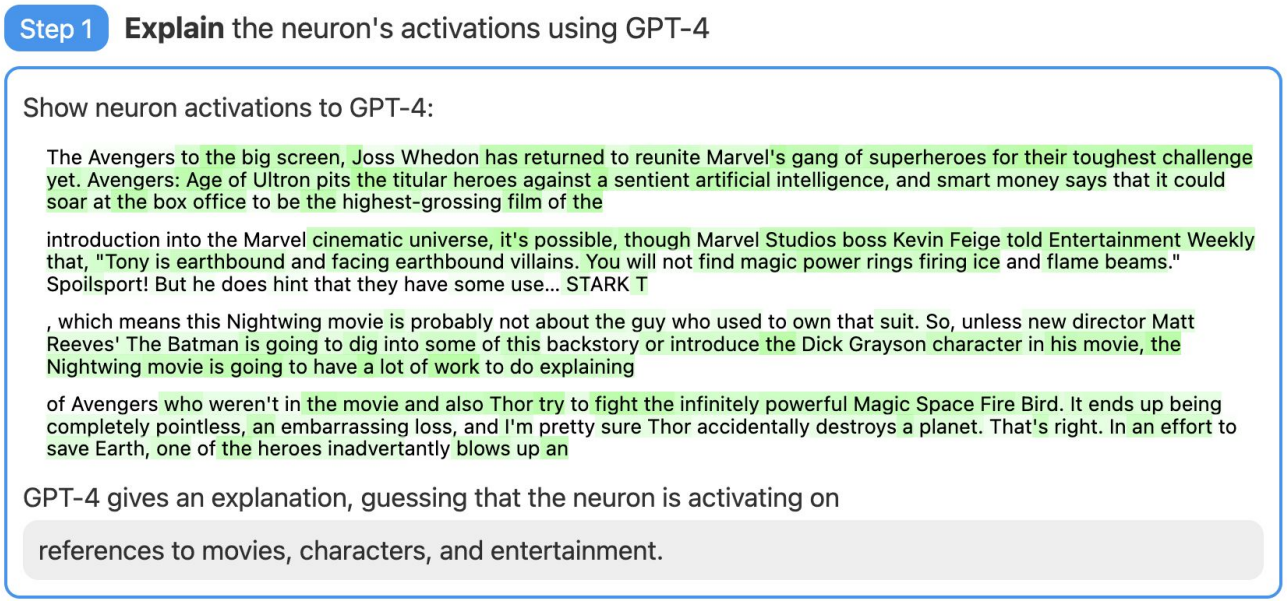
Using GPT-4 to Explain Neurons of GPT-2



Explanation: X by / after X

Using GPT-4 to Explain Neurons of GPT-2

Propose hypothesis: few-shot prompting



[OpenAI, 2023]

Using GPT-4 to Explain Neurons of GPT-2

Propose hypothesis: few-shot prompting

Few-Shot Activation-Explanation Pairs + Input Activations

Explanations

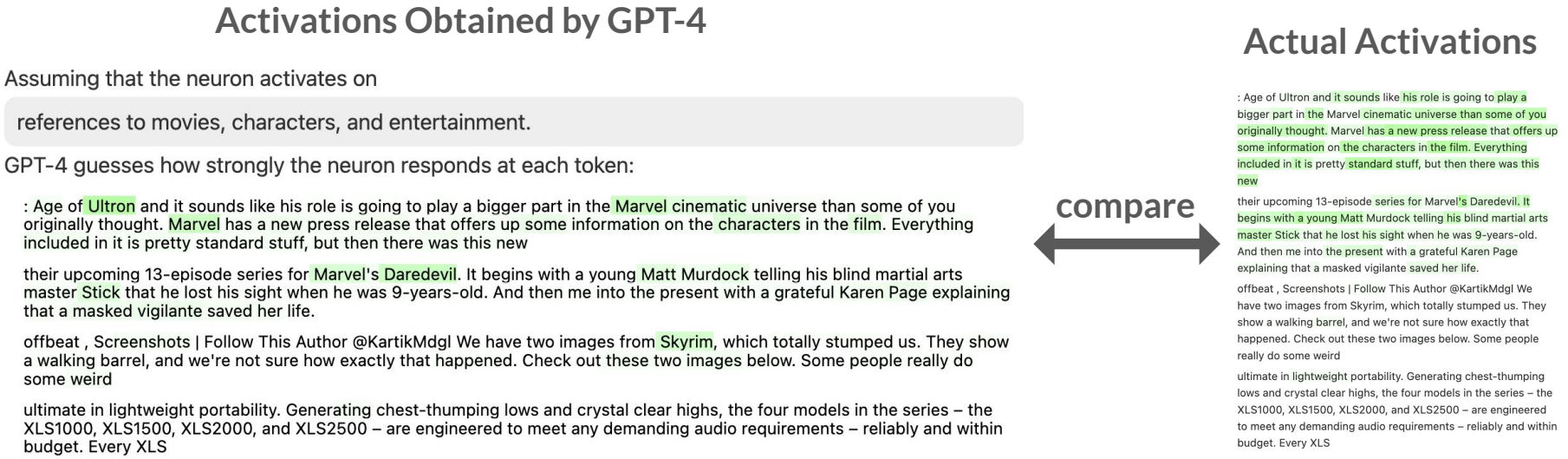
```
<start>
 together       3
ness            7
town            1
<end>
<start>
[prompt truncated …]
<end>

Explanation of neuron 1 behavior: the main thing this neuron does is find phrases related to
community
```

[OpenAI, 2023]

Using GPT-4 to Explain Neurons of GPT-2

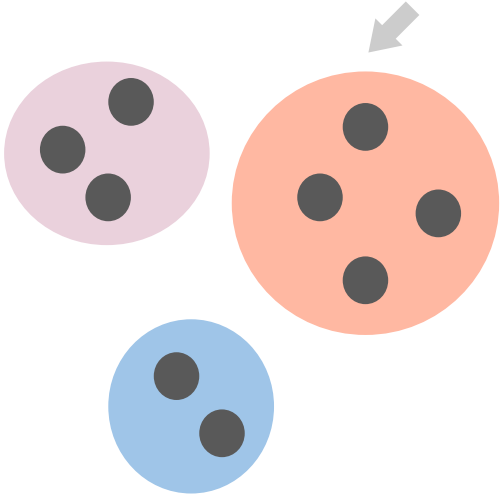
Verify hypothesis:
simulate activations based on explanations; compare simulated and actual activations



[OpenAI, 2023]

Similar Work: Annotating Concept of Word Clusters

clustering words based on contextualized representations



prompt LLMs to annotate the concept of clusters

```
Assistant is a large language model
trained by OpenAI
Instructions:
Give a short and concise label that best
describes the following list of words:
[“word 1”, “word 2”, ..., “word N”]
```

According to human evaluation, ChatGPT annotations are generally acceptable and moderately preferred over human annotations

[Mousi et al., EMNLP 2023]

Evaluating the NL Explanations of Neurons

Whether explanations accurately align with neuron activations?

- Type-1 Error (recall): falsely predicts that the neuron will activate on a concept
- Type-2 Error (precision): falsely predicts that the neuron will not activate on a concept

Explanation	True Positives	Type I Errors	Type II Errors
days of the week	I have a music class every Wednesday evening	Thursday is usually reserved for grocery	Philadelphia is where the Declaration of Independence
years, specifically four-digit years	Castro took power in Cuba in 1959 .	rated during re - entry in 2003 .	We need to rev amp the website to attract more

Not well aligned: Around 0.6 F1 score across 300 of the top-scoring explanations found by GPT-4

[Huang et al., BlackboxNLP 2023]

Takeaways

LLMs can help annotate/summarize concepts from collections of text snippets

But LLM-produced neuron-level explanations are not accurate enough

Outline

1. Neuron-level interpretability
    a. Sparse Autoencoders
2. Causality Background and Methods
    a. Causal Mediation
    b. Activation Patching/Causal Tracing and Path Patching
    c. High-level causal graphs
    d. Other methods
3. What is mechanistic interpretability?
4. Methods Leveraging Language Model Strengths
    a. Transformer Residual Stream and Linear Structure
    b. Vocabulary projection
    c. Decoding Natural Language Explanations from Representations
5. **Conclusion + Q&A**

Recap: Looking into transformers

- Pros
  - We are actually studying the weights in the model. Intuitively, it is more likely to be faithful.
  - Many methods are extensible to other types of models, modalities, etc.
- Cons
  - High-dimensional spaces remain challenging to make sense of and there could existing fundamental limitations so that it is impossible to reverse-engineer the model or for humans to make sense of these models
  - Illusion of understanding
  - Negative results can be uninformative
  - Lack of standardized evaluation & benchmarks
- Open questions
  - What granularity or type of model internals to target?
  - How to unify work from various methods/communities?