

AIAA 2290: Ethics, Privacy and Security in AI

Introduction to Explainable AI and LLM

Yibo YAN

yyan047@connect.hkust-gz.edu.cn

The Hong Kong University of Science and Technology (Guangzhou)

2025 Spring



You vote for this topic



Roadmap of this tutorial

- **A quick overview of WHY we need explainable AI**
- **More on ‘explanation’ for LLM**
- **No security of LLM this time** (but it is a FUN topic about how you can hack or defend LLM)
- **No AI ethical regulation & law**

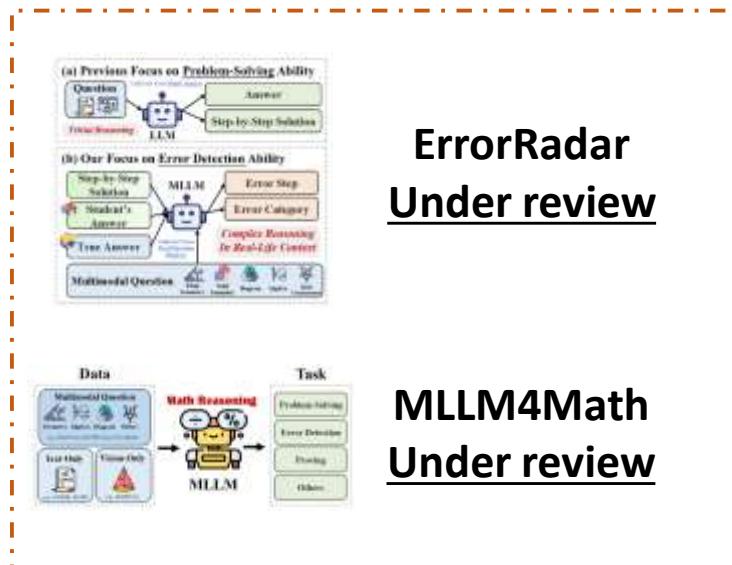


I am Yibo, a PhD student supervised by Prof.HU. My master and bachelor degrees are from National University of Singapore and Hong Kong Baptist University, respectively.
My current research mainly focus on **Multimodal LLM**.

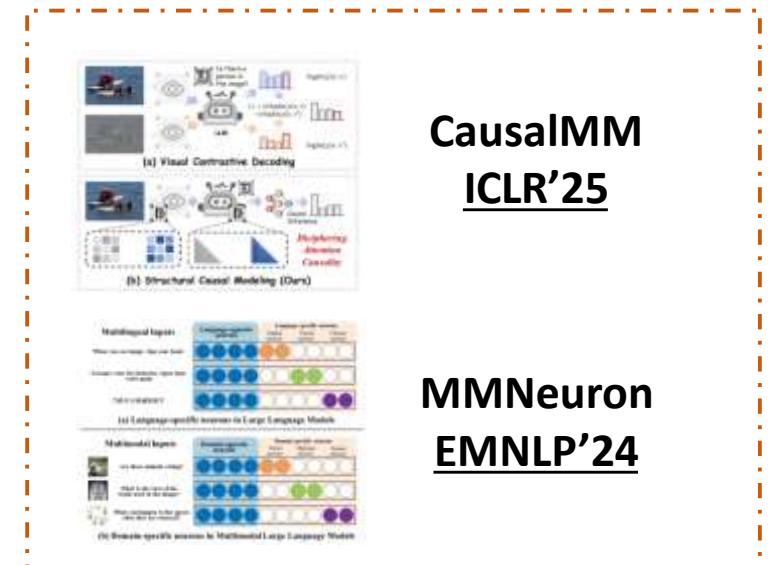
Multimodal Understanding



Multimodal Reasoning



Multimodal Trustworthiness





1 Introduction to the Explainable AI

2 Prompting-based Explanations

3 Data Attribution

4 Transformer Understanding

5 Conclusion

Left for own
reading





Motivation: Why we need explainability?
----From a Model Perspective



Why explainability: Debug (Mis-)Predictions



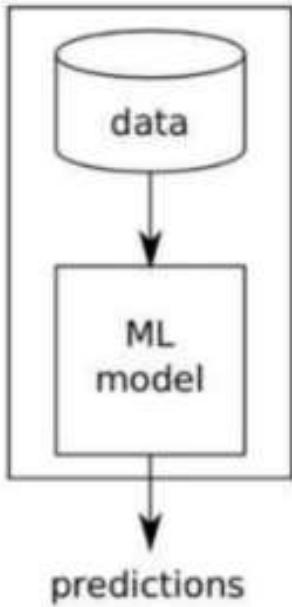
Top label: “**clog**”

Why did the network label
this image as “**clog**”?

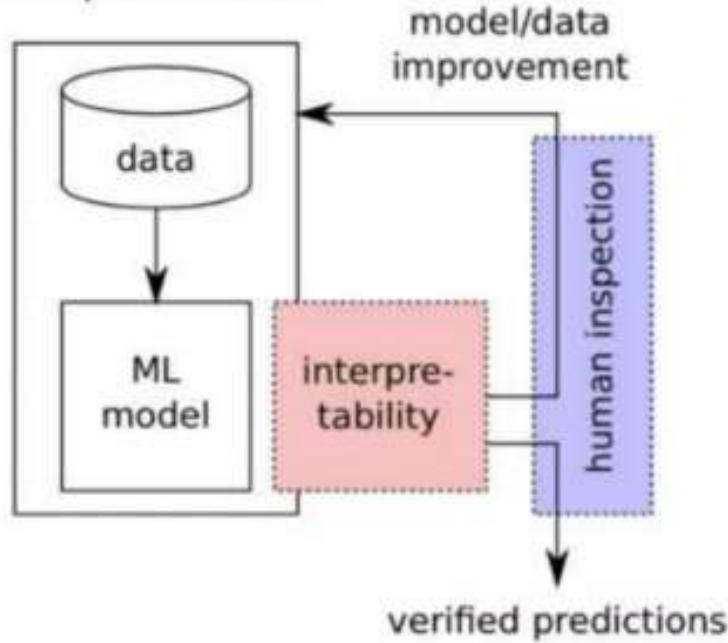


Why explainability: Improve ML Model w/ Human Experience

Standard ML



Interpretable ML



Generalization error

Generalization error + human experience



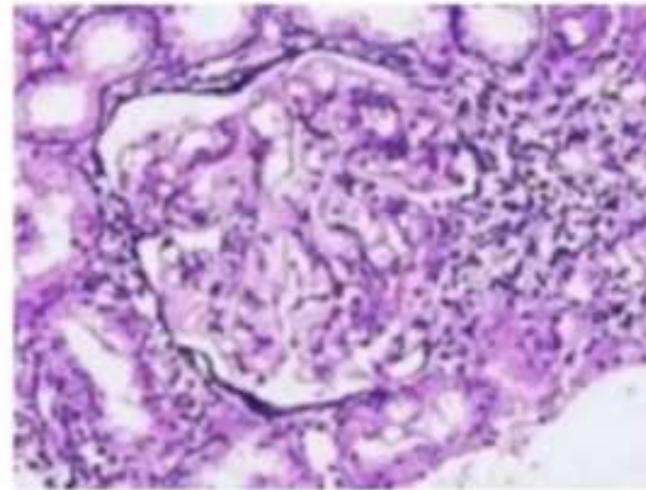
Why explainability: Verify ML Model/System

Wrong decisions can be costly
and dangerous

*“Autonomous car crashes,
because it wrongly recognizes ...”*



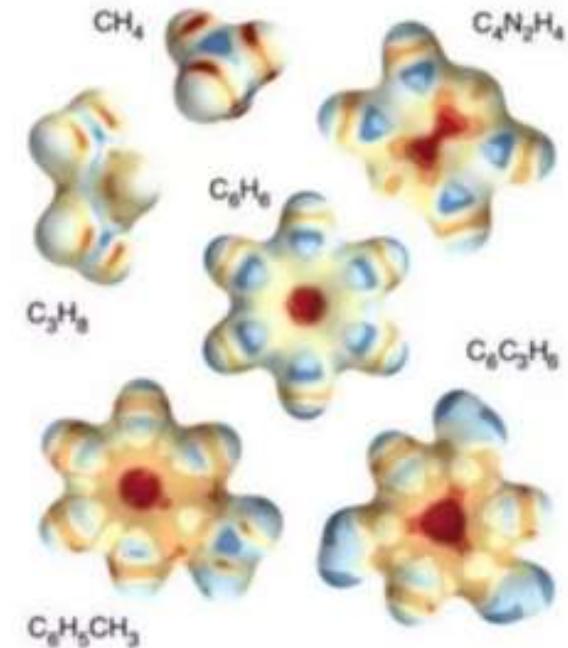
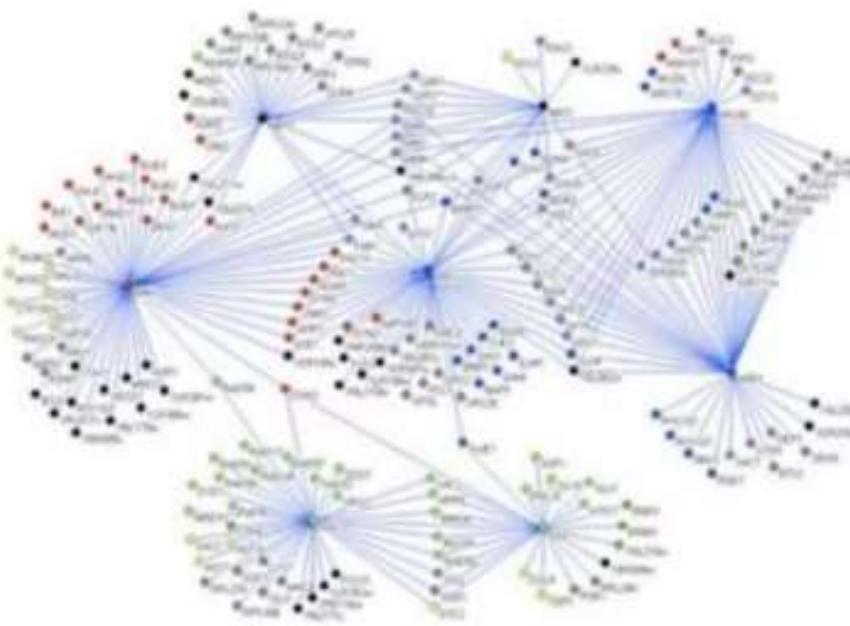
*“AI medical diagnosis system
misclassifies patient’s disease ...”*





Why explainability: Learn Insights in the Sciences

Learn about the physical / biological / chemical mechanisms.
(e.g. find genes linked to cancer, identify binding sites ...)





What is the stuff that follows the “why”?

Let's clarify them in this example:

Why is the sky turning dark? **The sun is going down.**

- **Explanandum** is what describes the **problem**.
- **Explanan** is what describes the **reason**.
- **Explanation** is the **process(*)** of stating the reason for the problem.

*: The term “**explanation**” is also frequently used to refer to the **product** of this process.



There are many types of explanations

By the reasoning patterns:

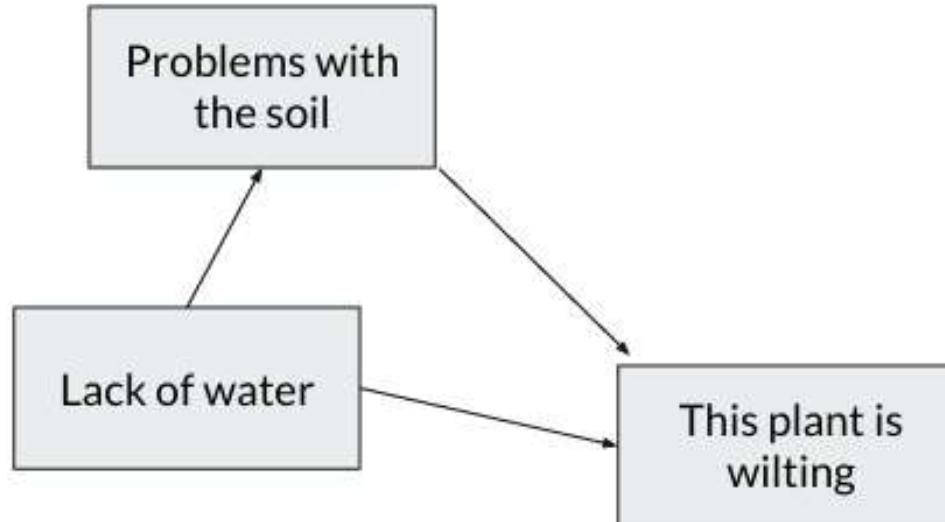
A plant in the living room is wilting. Why?

- **Deductive:** If a plant does not receive enough water, it will wilt. That plant in the living room hasn't been watered for a month.
- **Inductive:** Other plants in that living room have wilted because they haven't received enough water. It's likely that plants wilt because of lack of water.
- **Abductive:** It's likely that the plant in this living room is wilting because there's no enough water.
- ...



There are many types of explanations

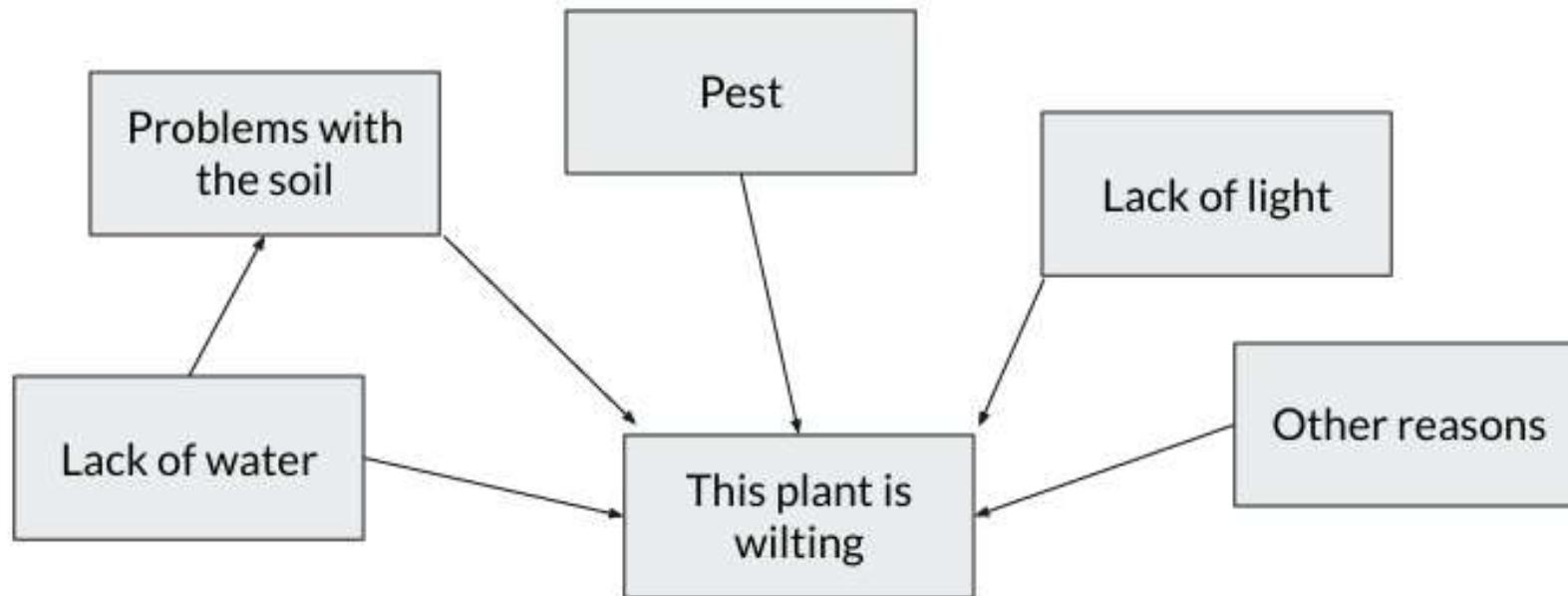
By the causal variables:





There are many types of explanations

By the completeness: selective, or comprehensive?





Good explanations should be faithful

To explain a model's prediction:

"A **faithful** interpretation is one that accurately represents the reasoning process behind the model's prediction." [\[Jacovi & Goldberg, 2020\]](#)

"This is often considered the most fundamental requirement for any explanation, and sometimes used interchangeably with the term 'interpretability.'" [\[Lyu et al. 2023\]](#)

Plausibility and faithfulness are two desirable (but different) properties in explanations. [\[Wiegreffe and Pinter, 2019\]](#)

Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? (Jacovi & Goldberg, ACL 2020)

Towards Faithful Model Explanation in NLP: A Survey. (Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch, 2023)

Attention is not not Explanation (Wiegreffe & Pinter, EMNLP-IJCNLP 2019)



Good explanations should be plausible

An explanation is considered plausible if it is coherent with human reasoning and understanding.

[\[Agrawal et al. 2024\]](#)

Plausibility is also referred to as **persuasiveness** or **understandability**.

An explanation might be plausible but not faithful. Currently, many explanations are more plausible than faithful.

Example of faithful, but not plausible explanation: a copy of model weights. [\[Jacovi & Goldberg. 2020\]](#)



Good explanations should be informative



Hi prof, I have just finished this paper. Which venue do you think would best suit it?

NAACL, because its deadline is just 3 days away, and it will be in Mexico, not far from here.



NAACL, because it is a top NLP conference.



Which explanation is more informative?

This example is modified from: Zachary C. Lipton. 2017. The Mythos of Model Interpretability. arXiv:1606.03490 [cs, stat].



Good explanations should be useful

Tom holds a copper block by hand and heats it on fire. Which of the following is more likely?

- A. His fingers feel warm. B. His fingers feel burnt.

A (92.93%)



Copper is a good thermal conductor. Tom holds a copper block by hand and heats it on fire. Which of the following is more likely?

- A. His fingers feel warm. B. His fingers feel burnt.

B (90.05%)



This example is modified from the e-CARE dataset. Probabilities are computed by GPT-3.5-Turbo model, as of April 19, 2024. This example is about utility to models – explanation should also be useful to humans.



Other desirable properties

Completeness: All necessary information is covered.

Stability: The explanation remains consistent for similar cases I ask about.

Interactivity: The explanation can answer my follow-up questions.

Personalization: The explanation is tailored to my needs.

...

For a more detailed list please refer to [\[Liao et al., 2022\]](#)



- Extractive rationales / Feature attributions
- Free-text explanations
- Structured explanations

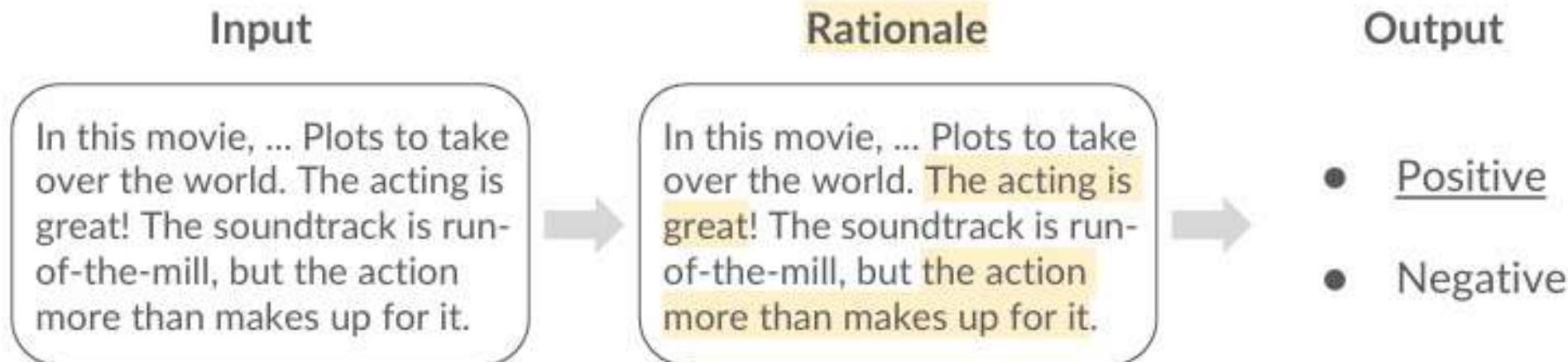


- Extractive rationales / Feature attributions
- Free-text explanations
- Structured explanations



Extractive Rationales

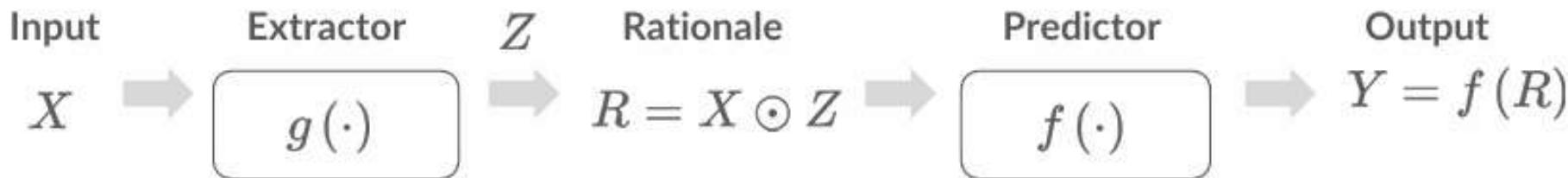
(short) snippets in inputs that support outputs





Extractive Rationales

Pipeline models [\[DeYoung et al. 2020\]](#)

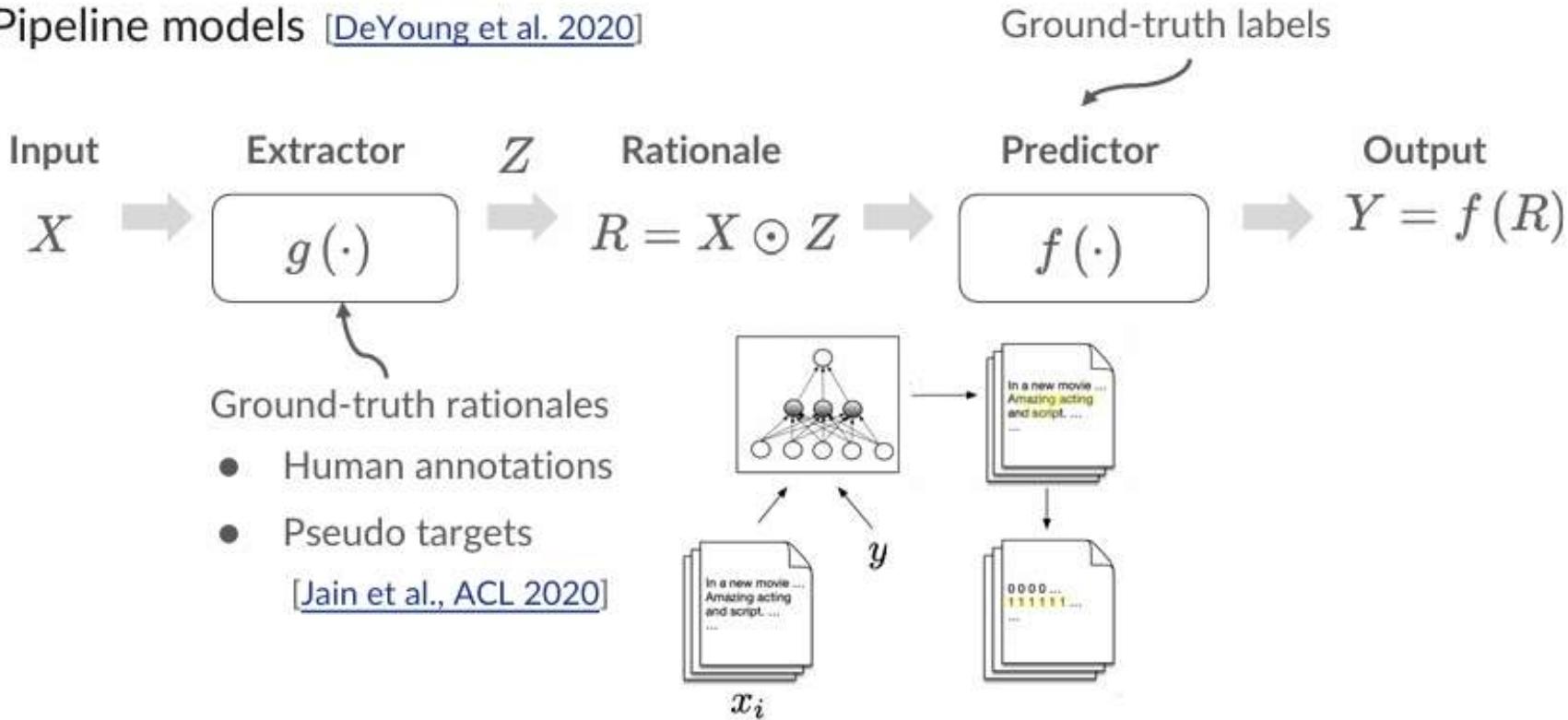


- Hard selection [\[Lei et al. 2016\]](#)
 Z Binary masks
- Soft selection
 Z Continuous scores



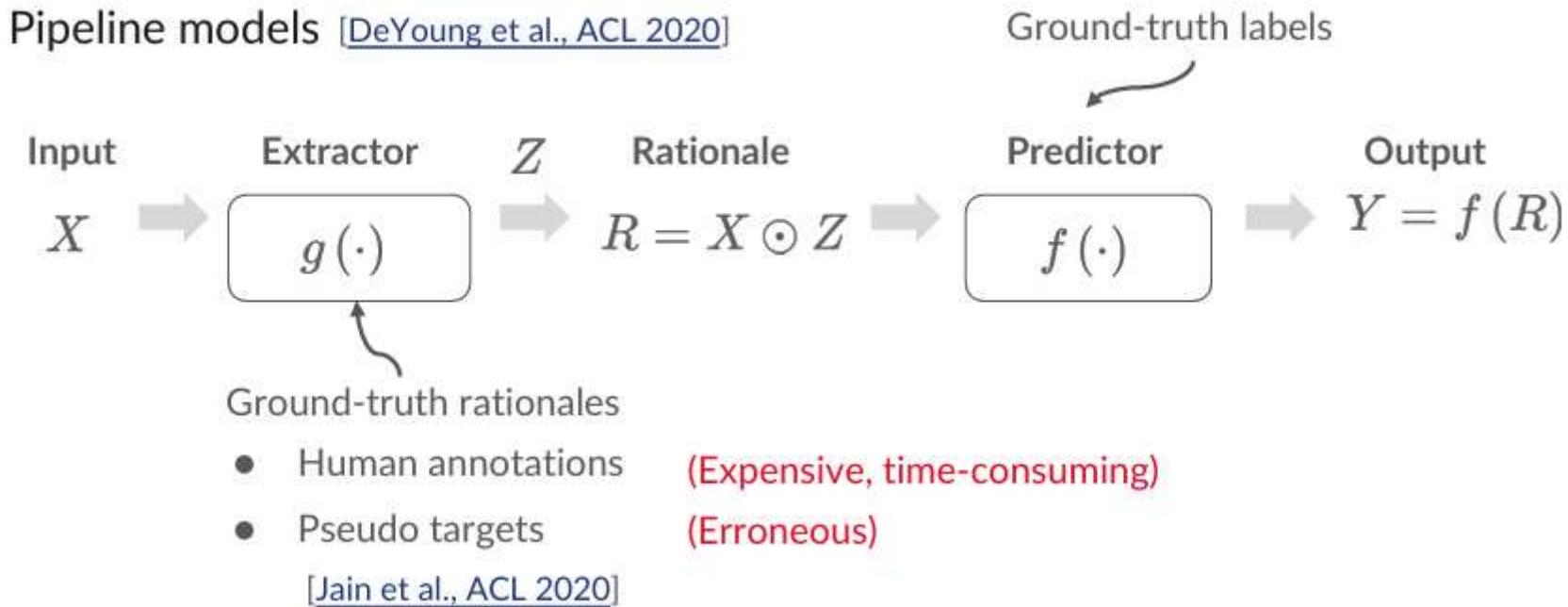
Extractive Rationales

Pipeline models [DeYoung et al. 2020]





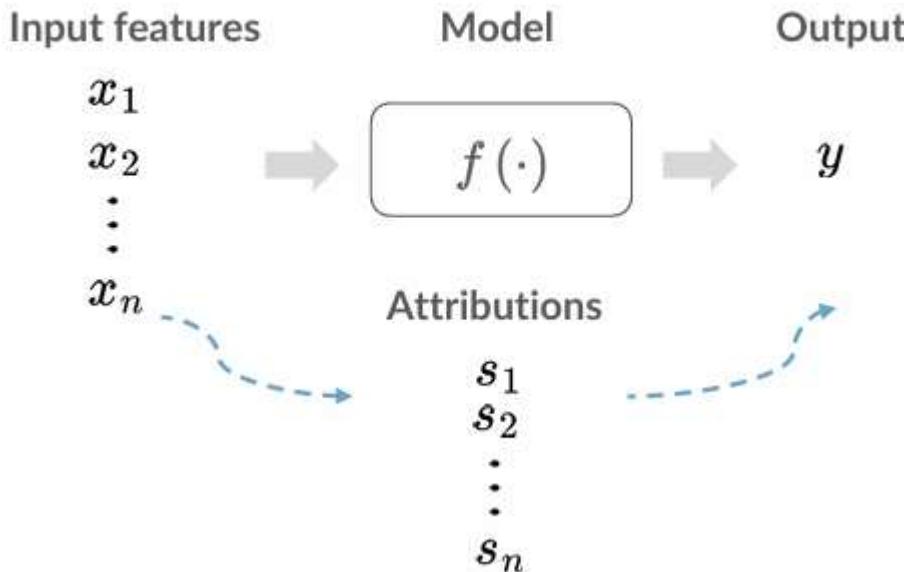
Extractive Rationales





Feature Attributions

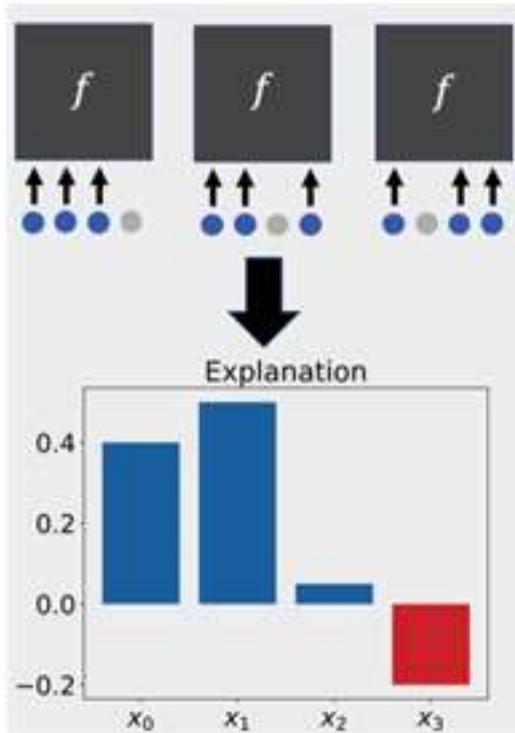
Importance scores of input features to model output



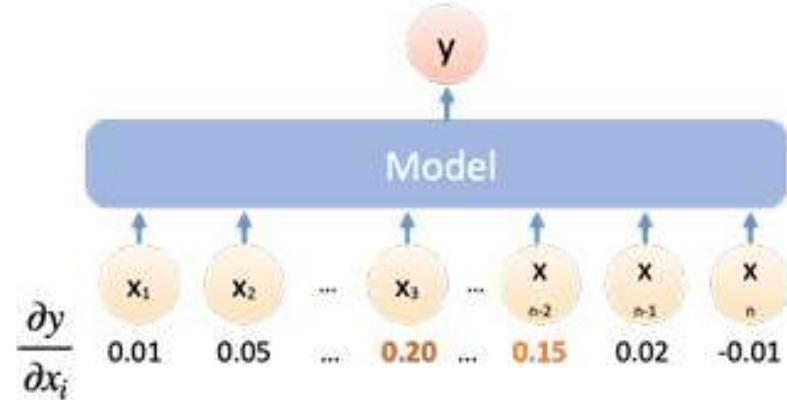


Feature Attributions

Leave-one-out



Gradient-based explanation



[Sundararajan et al. 2017]

[Covert et al. 2020]

A heatmap showing a 10x7 grid of numerical values representing feature attributions. The columns are labeled with German words: men, ie, ent, _g, _and, les, _ad, morning, good, </s>. The rows are labeled with English words: .., men, ie, ent, _g, _and, les, _ad, morning, good, </s>. A color scale on the right ranges from -0.6 (dark blue) to 0.6 (dark red). The values are mostly small, with some larger values like 0.76 and 0.65 appearing in the "les" row.

..	0.00	0.00	0.00	0.00	0.00	0.00	0.00
men	0.03	-0.00	0.03	0.00	0.02	0.16	0.02
ie	0.03	0.02	-0.03	0.04	-0.00	0.54	0.02
ent	0.01	-0.04	0.20	-0.01	-0.03	0.37	0.10
_g	0.03	-0.03	-0.26	-0.06	-0.09	0.05	0.07
_and	-0.03	-0.06	-0.11	-0.13	0.76	-0.18	-0.14
les	0.04	-0.23	-0.17	-0.12	0.06	0.07	0.03
_ad	0.09	0.07	0.23	0.65	0.11	0.03	-0.04
morning	0.08	0.78	0.66	-0.23	0.00	-0.06	-0.06
good	0.15	0.41	0.31	0.04	0.10	-0.02	-0.07
</s>	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ϕ_i	men	Morgen	Damen	und	und	und	Herren



Challenges for LLMs

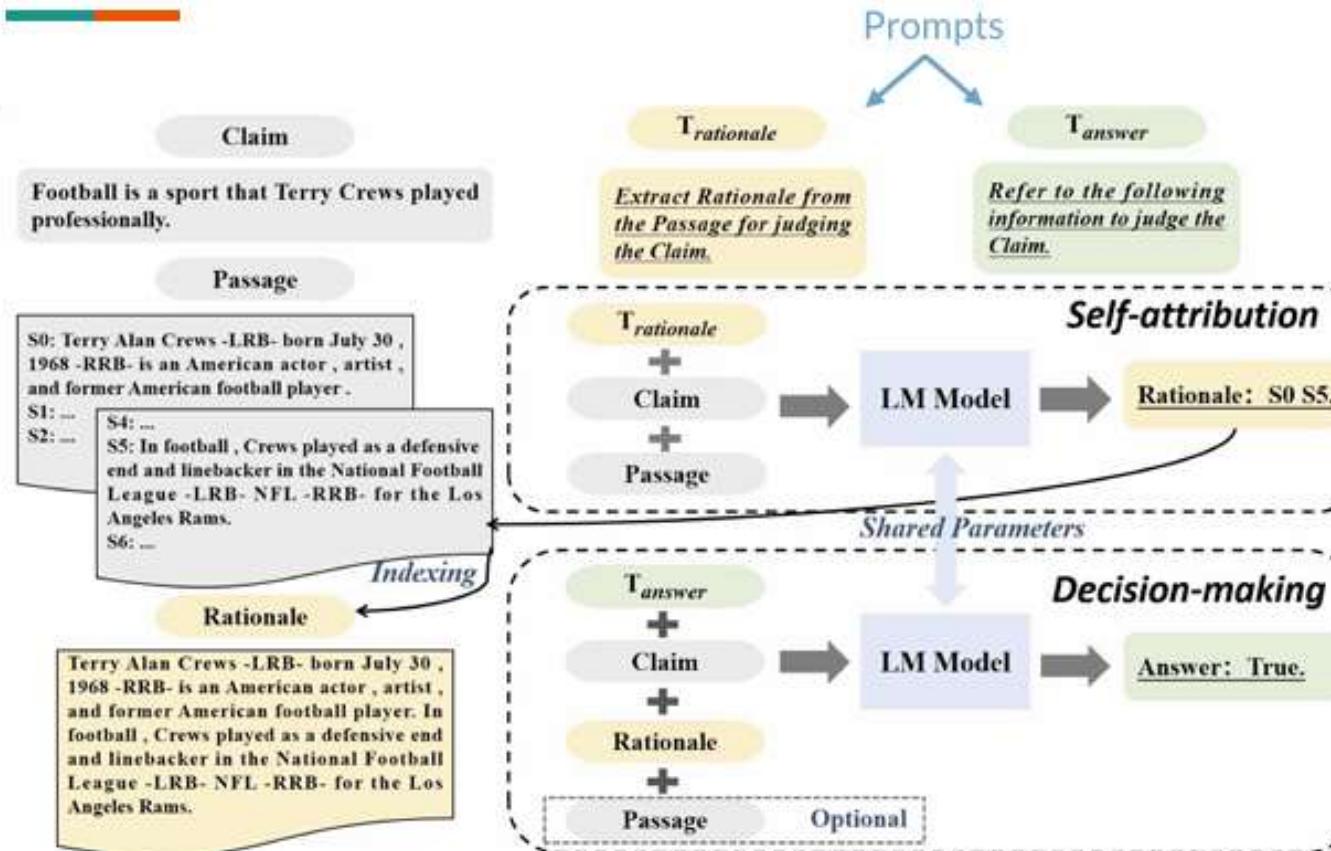
- Computational cost
- Low efficiency in long context
- No access to API-based models (gradients, attention scores, etc.)



Prompting-based extractive rationales/feature attributions



Self-Attribution and Decision-Making



Stage 1:
Prompting for
extractive rationales

Stage 2:
Making decisions
based on rationales

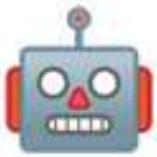
[Du et al. 2023]

See also: [Ludan et al. 2024]



How to evaluate rationales/feature attributions?

Faithfulness



Explanation

Plausibility



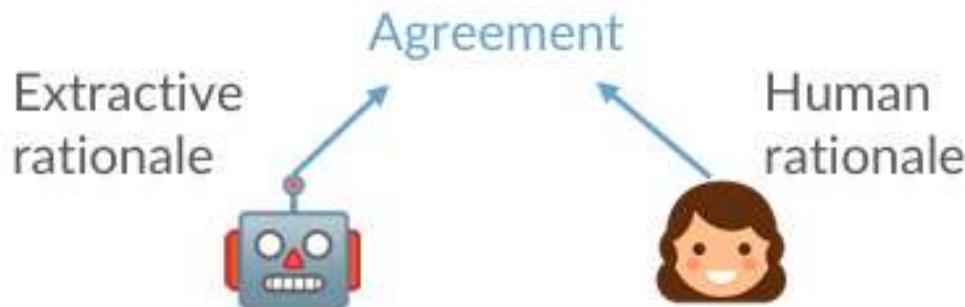
*How accurately the explanation reflects the **true reasoning process** of the model*

How convincing the explanation is to humans



Evaluation—Plausibility

- Agreement
e.g. Intersection-Over-Union (IOU)

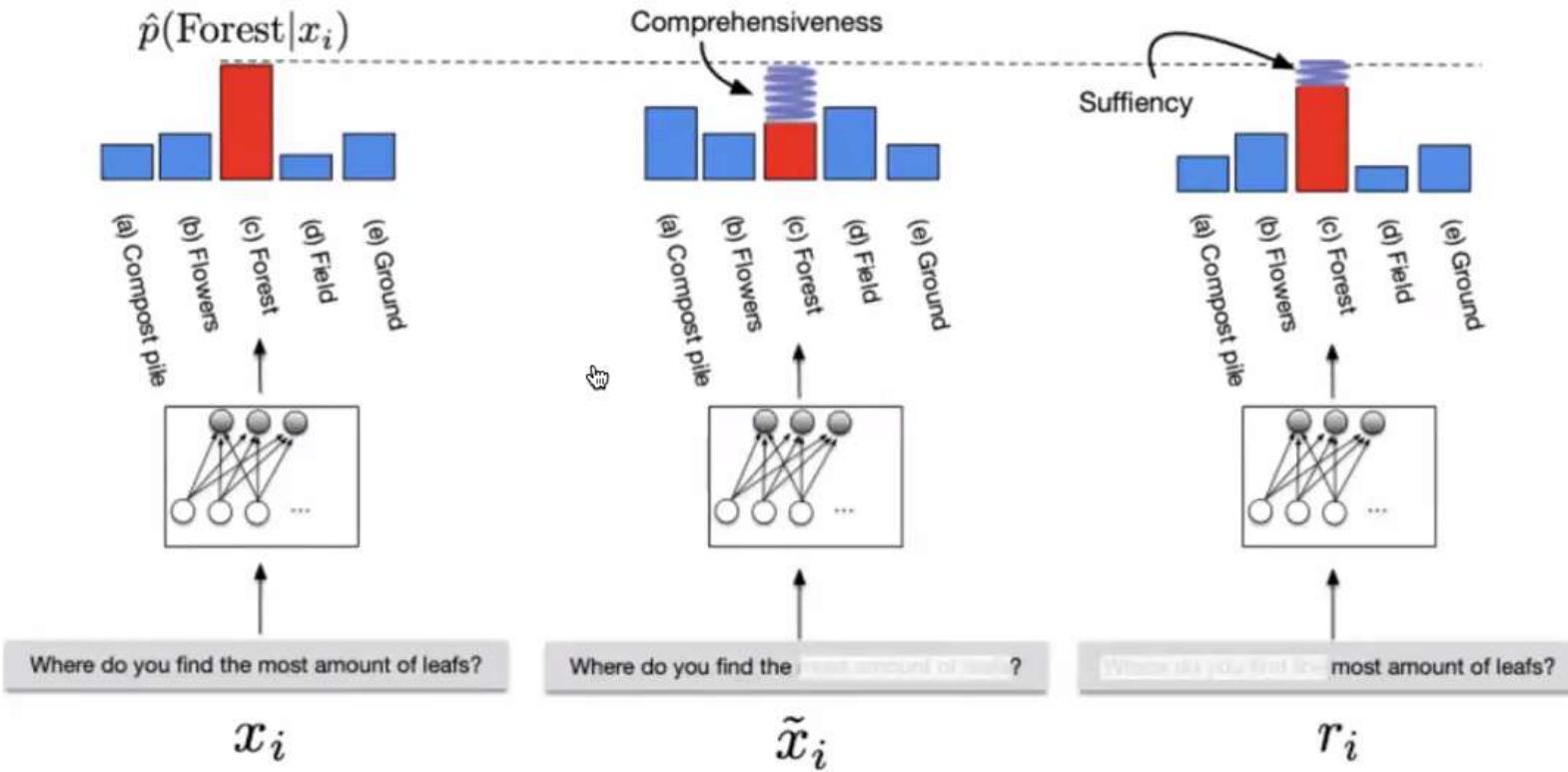




Evaluation—Faithfulness

$$\text{Comprehensiveness} = f_{\hat{y}}(x_i) - f_{\hat{y}}(x_i \setminus r_i)$$

$$\text{Sufficiency} = f_{\hat{y}}(x_i) - f_{\hat{y}}(r_i)$$





Evaluation—Faithfulness

Session 1 (prediction and explanation)

Is the following candidate a good fit for a Senior SWE position? Answer only yes/no.

Education:

2016-2020: Bachelor in Biology at University Y
{resume continues ...}

User input

No

Model response

Make a minimal edit to the resume, 5 words or less, such that you would answer yes.

Education:

2016-2020: BSc in CS at University Y
{counterfactual resume continues ...}

Session 2 (self-consistency)

Is the following candidate a good fit for a Senior SWE position? Answer only yes/no.
{insert counterfactual resume}

Yes

Edited input



Opposite prediction Faithful

Finding: Faithfulness is **dependent on many factors** – explanation type, model, task ...

[[Madsen et al. 2024](#)]



- Extractive rationales / Feature attributions
- Free-text explanations
- Structured explanations



Free-text Explanations

Example: Natural Language Inference (NLI) task

Premise (**p**)

Kids are on an amusement ride

Hypothesis (**h**)

Kids are riding their favorite amusement ride

Does the **p** entail **h**?

Model prediction: Maybe

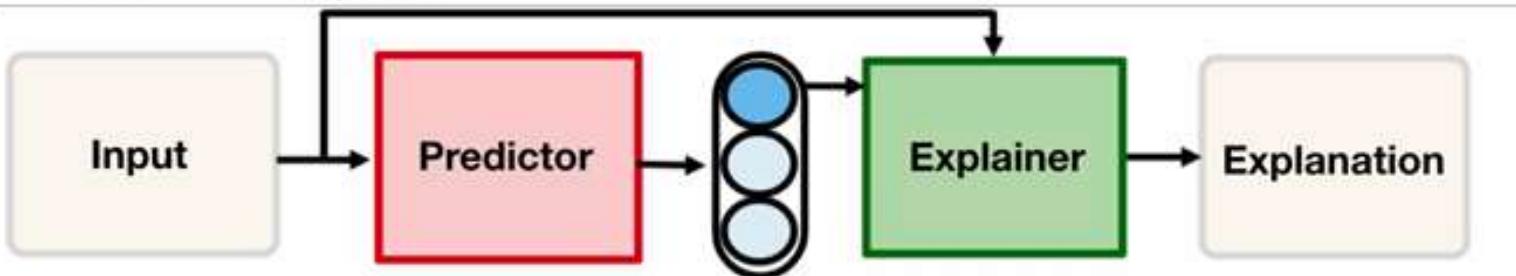
Free-text explanation: It isn't necessarily their favorite ride.



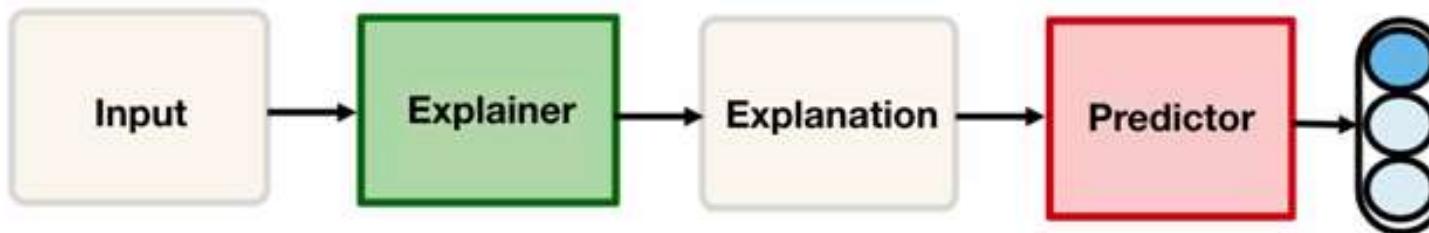
How to Generate Free-text Explanations?

- Traditionally: jointly **train** a predictor & explainer

- Predict-then-explain:*



- Explain-then-predict:*



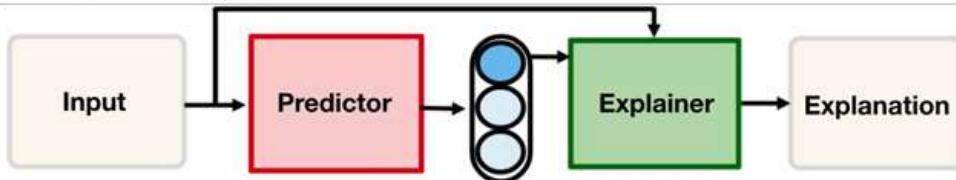
[Kumar and Talukdar 2020]



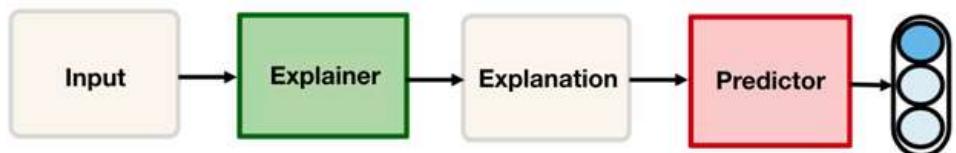
How to Generate Free-text Explanations?

- Traditionally: jointly train a predictor & explainer

- Predict-then-explain:



- Explain-then-predict:



[Kumar and Talukdar 2020]

Any **cheaper** way?

- + Can steer models toward using the “right” signal
- Need lots of human-written explanations as **training data**
 - Natural Language Inference: e-SNLI [Camburu et al. 2018]
 - Commonsense QA: CoS-E [Rajani et al. 2019], ECQA [Aggarwal et al. 2021]
 - Social bias inference: SBIC [Sap et al. 2020]



Can we prompt LLMs to generate them with just a few examples?



In-context learning / Few-shot prompting [Brown et al. 2021]



Prompting for Explanations

- GPT-3-level LLMs can generate **plausible** free-text explanations for **simple tasks***:
 - NLI
 - Commonsense QA
 - Social bias detection ...
- What about **multi-step reasoning**?
 - Maths
 - Multi-hop QA
 - Planning ...

*[Wiegreffe et al. 2021; Marasović et al. 2021]

Let's explain classification decisions.

A young boy wearing a tank-top is climbing a tree.

question: A boy was showing off for a girl.
true, false, or neither? **neither**

why? A boy might climb a tree to show off for a girl,
but he also might do it for fun or for other reasons.

###

A person on a horse jumps over a broken down airplane.

question: A person is outdoors, on a horse.
true, false, or neither? **true**

why? Horse riding is an activity almost always done
outdoors. Additionally, a plane is a large object and is
most likely to be found outdoors.

###

There is a red truck behind the horses.

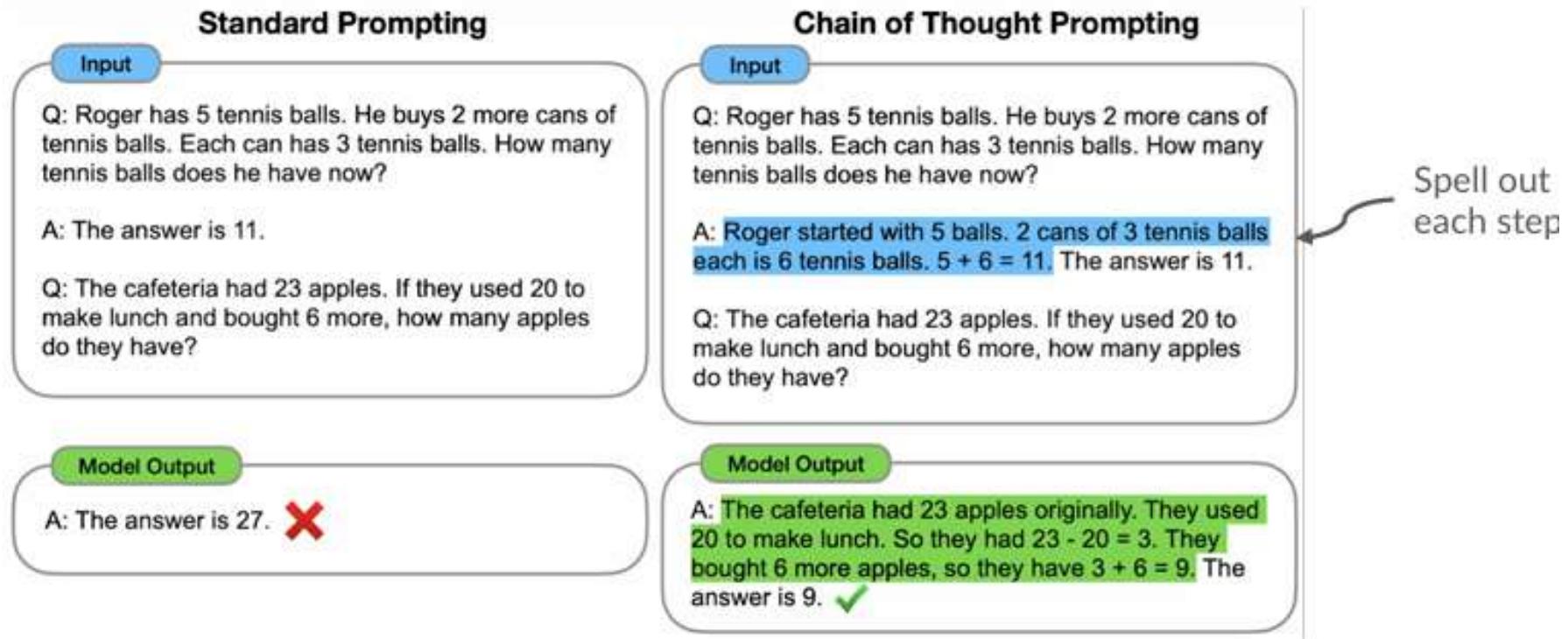
question: The horses are becoming suspicious of my
apples.

true, false, or neither? **false**

why? The presence of a red truck does not imply there
are apples, nor does it imply the horses are suspicious.



“Chain of Thought” (CoT)



[Wei et al. 2022]

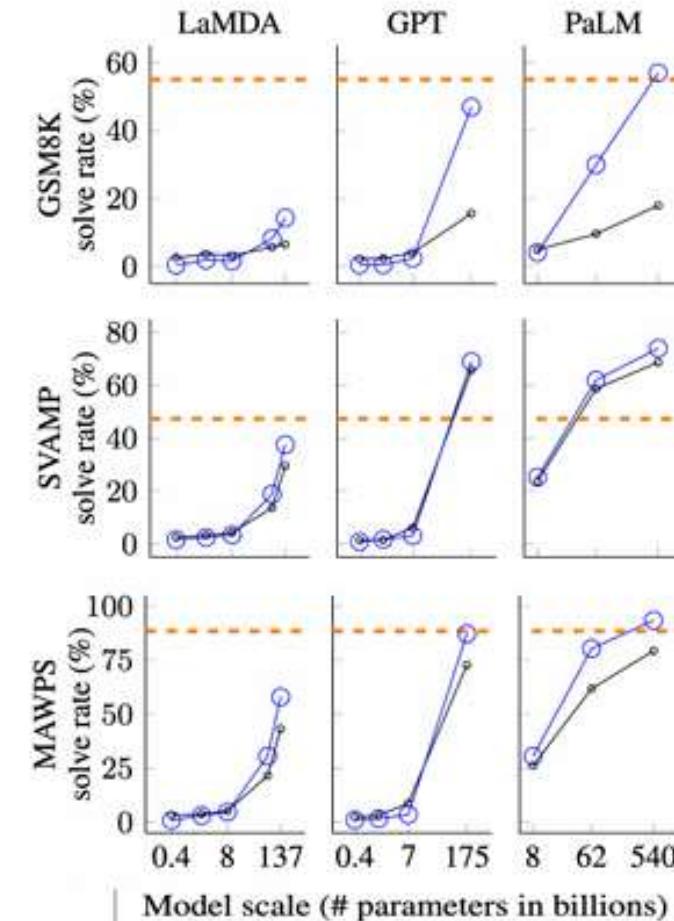
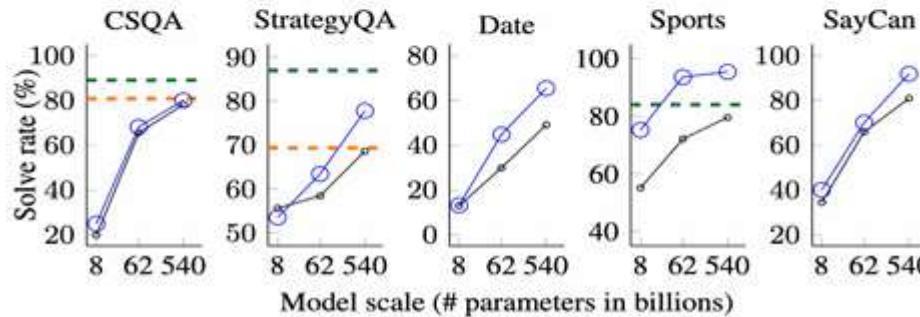
See also: Scratchpad [Nye et al. 2021]; “Let’s Think Step by Step” [Kojima et al. 2023]



“Chain of Thought” (CoT)

- CoT prompting boosts LLMs' performance on multi-step reasoning

—○— Standard prompting
—○— Chain-of-thought prompting
—○— Prior supervised best



Limitation: Easy-to-hard generalization



CoT + Question Decomposition

Stage 1: Decompose Question into Subquestions

Q: It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The water slide closes in 15 minutes. How many times can she slide before it closes?

Language Model

A: To solve "How many times can she slide before it closes?", we need to first solve: "How long does each trip take?"

Stage 2: Sequentially Solve Subquestions

Subquestion 1

It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The slide closes in 15 minutes.

Q: How long does each trip take?

Language Model

A: It takes Amy 4 minutes to climb and 1 minute to slide down. $4 + 1 = 5$. So each trip takes 5 minutes.

Append model answer to Subquestion 1

Subquestion 2

It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The slide closes in 15 minutes.

Q: How long does each trip take?

A: It takes Amy 4 minutes to climb and 1 minute to slide down. $4 + 1 = 5$. So each trip takes 5 minutes.

Q: How many times can she slide before it closes?

Language Model

A: The water slide closes in 15 minutes. Each trip takes 5 minutes. So Amy can slide $15 \div 5 = 3$ times before it closes.

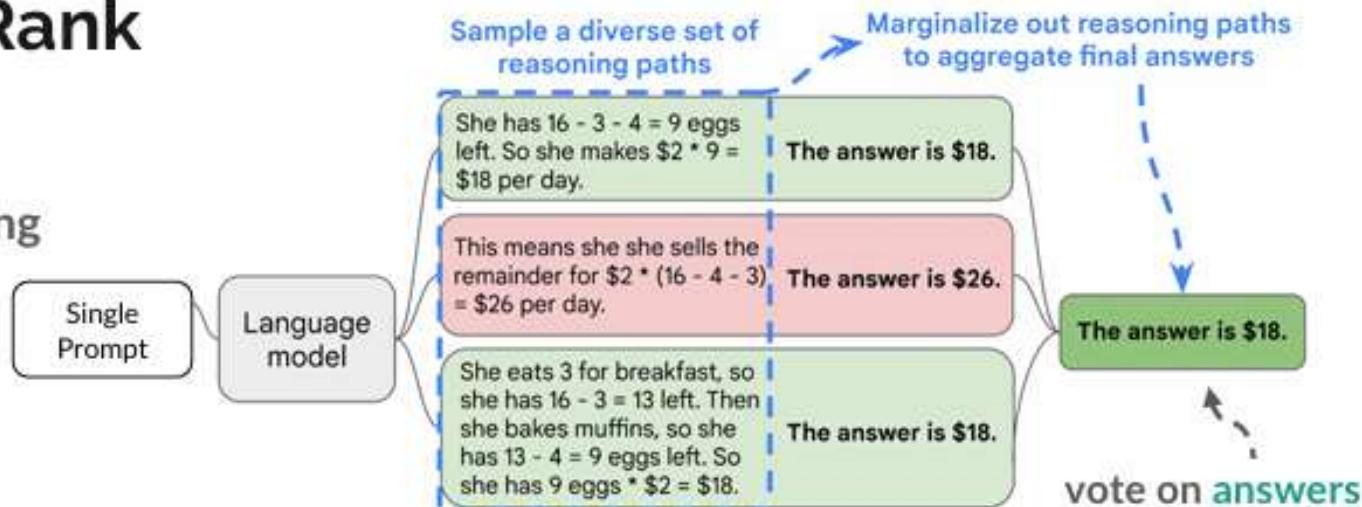
+ Better generalization than CoT

- Greedy decoding has limited diversity

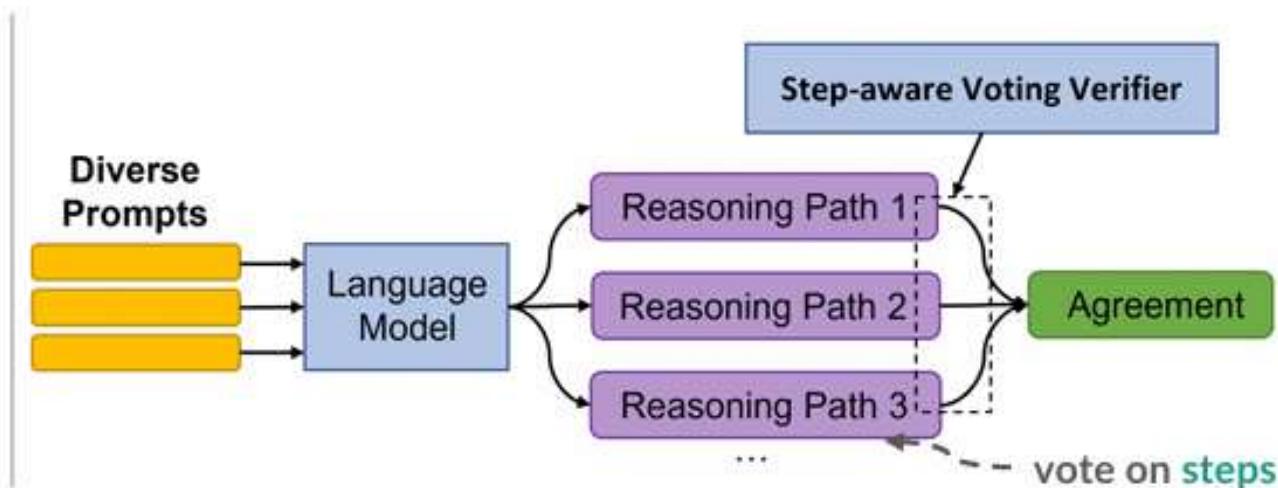


CoT + Vote and Rank

Self-Consistency Prompting
[Wang et al. 2022]



DiVeRSe
[Li et al. 2023]



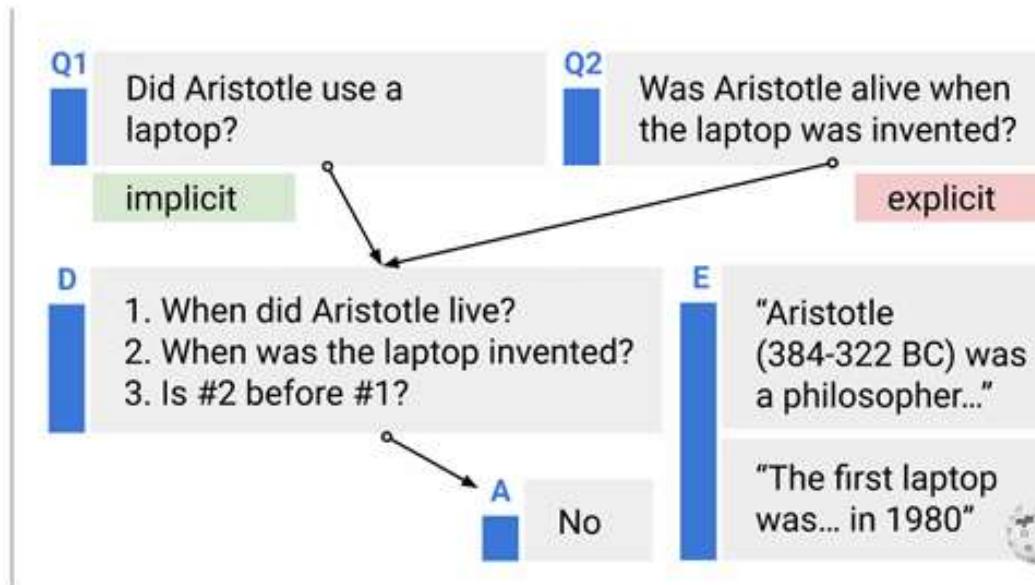


- Extractive rationales / Feature attributions
- Free-text explanations
- Structured explanations



Why Structured Explanations?

- Certain problems intrinsically involve a *non-linear* mode of reasoning
 - multi-hop QA, logical deduction, constrained planning...



StrategyQA dataset
[Geva et al. 2021]



Why Structured Explanations?

- Unclear **faithfulness** of free-text explanations
 - False impression of “**self-interpretability**”
 - Easier **over-trust** in the model
 - especially if explanations look **plausible**



Should I hire this candidate?



Generated CoT

Based on their excellent **education background** and strong **technical skills**, I highly recommend hiring this candidate



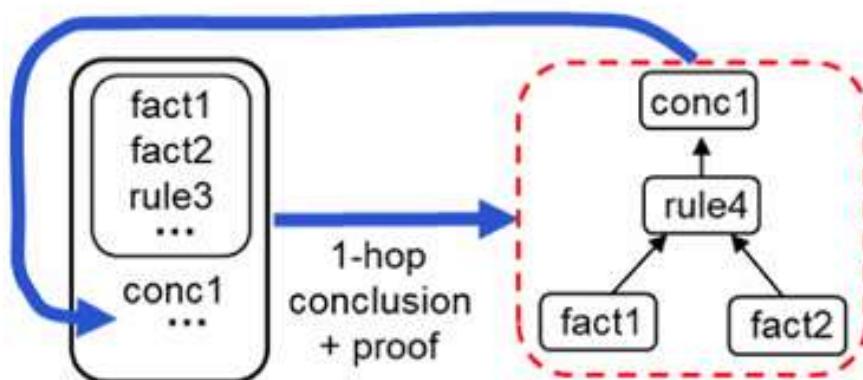
True Reasoning

Their **name** looks like a white male, so I highly recommend hiring this candidate



How to Generate Structured Explanations?

- Traditionally: train models to iteratively generate intermediate steps



ProofWriter [Tafjord et al 2021]

- Still needs lots of (even more expensive) training data

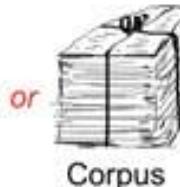
Question: How might eruptions affect plants?
Answer: They can cause plants to die

Hypothesis

H (hypot): Eruptions can cause plants to die

Text

sent1: eruptions emit lava.
sent2: eruptions produce ash clouds.
sent3: plants have green leaves.
sent4: producers will die without sunlight
sent5: ash blocks sunlight.



or

Entailment Tree

H (hypot): Eruptions can cause plants to die

int1: Eruptions block sunlight.

sent4: producers will die without sunlight.

sent2: eruptions produce ash clouds.

sent5: ash blocks sunlight.

EntailmentWriter [Dalvi et al 2021]



Structured Explanations by Prompting

- Can we prompt LLMs to generate structured explanations with a few examples?
- If so, what types of structures?
 - Logical constraints
 - Maieutic prompting, SatLM
 - Symbolic programs
 - Program of Thoughts, Program-Aided LMs, Faithful CoT
 - Non-linear exploration strategies
 - Tree of Thoughts, Graph of Thoughts
 - ...

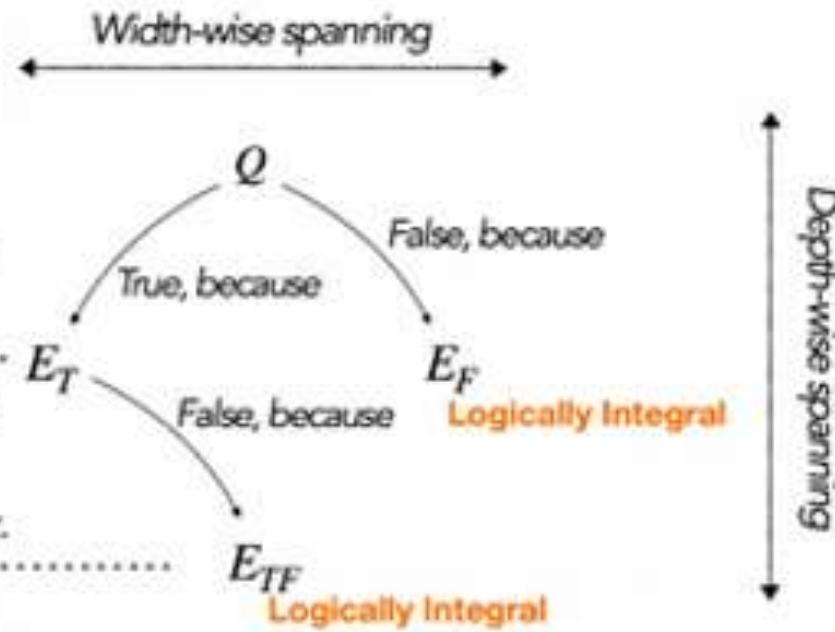


Logically-Constrained Reasoning

Q : War cannot have a tie?

- 🐶 War cannot have a tie? **True**, because
- 🐱 In a context of war, there's always a victor and a loser.
⋮
- 🐶 In a context of war, there's always a victor and a loser? **False**, because
- 🐱 There can be cases where the loser is not clear.

Maieutic prompting [Jung et al., 2022]



Maieutic tree generation

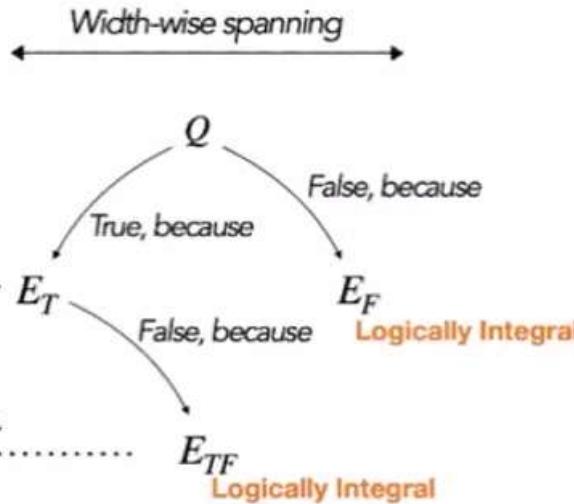


Prompting-based Explanations

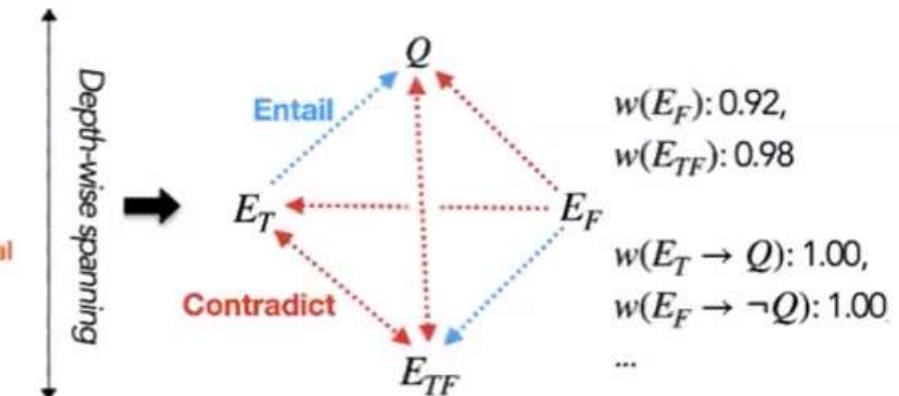
Logically-Constrained Reasoning

Q : War cannot have a tie?

- 🧐 War cannot have a tie? **True**, because
- 📝 In a context of war, there's always a victor and a loser.
-
- 🧐 In a context of war, there's always a victor and a loser? **False**, because
- 📝 There can be cases where the loser is not clear.
-



Maieutic prompting [Jung et al., 2022]

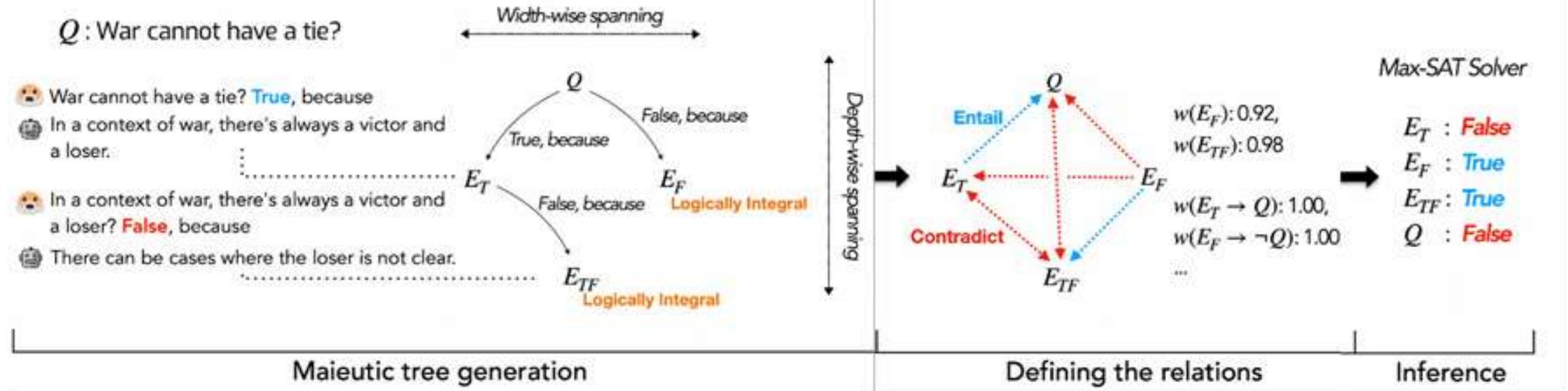


Maieutic tree generation

Defining the relations



Logically-Constrained Reasoning



Maieutic prompting [Jung et al., 2022]



Symbolically-Aided Reasoning

Query

There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?

Output

We start with 15 trees.
Later we have 21 trees.
The difference must be the number of trees they planted.
So, they must have planted $21 - 15 = 6$ trees.
The answer is 6.

Output

```
trees_begin = 15
trees_end = 21
trees_today = trees_end - trees_begin
answer = trees_today
```

>>>  >>> Answer: 6

Python Interpreter

Output

```
# 1. How many trees are there in the beginning? (independent, support: ["There are 15 trees"])
trees_begin = 15

# 2. How many trees are there in the end?
(independent, support: ["There are 15 trees"])
trees_end = 21

# 3. Final Answer: How many trees did the grove workers plant today?
trees_today = trees_end - trees_begin

>>>  >>> Answer: 6
```

Python Interpreter

CoT

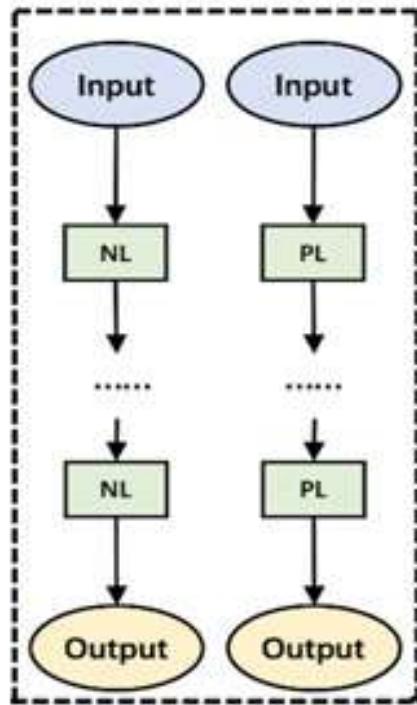
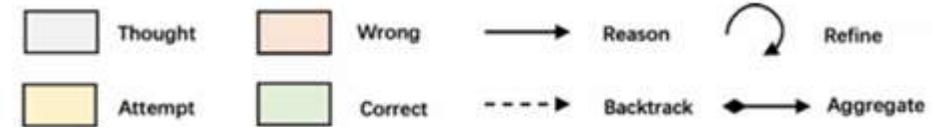
Program-Aided LM/PAL [Gao et al., 2023]

Program of Thoughts/PoT [Chen et al., 2023]

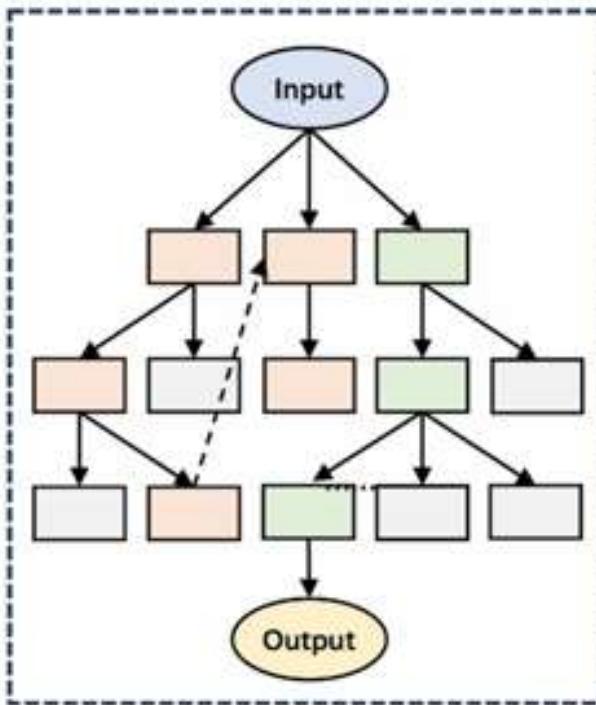
Faithful CoT [Lyu et al., 2023]



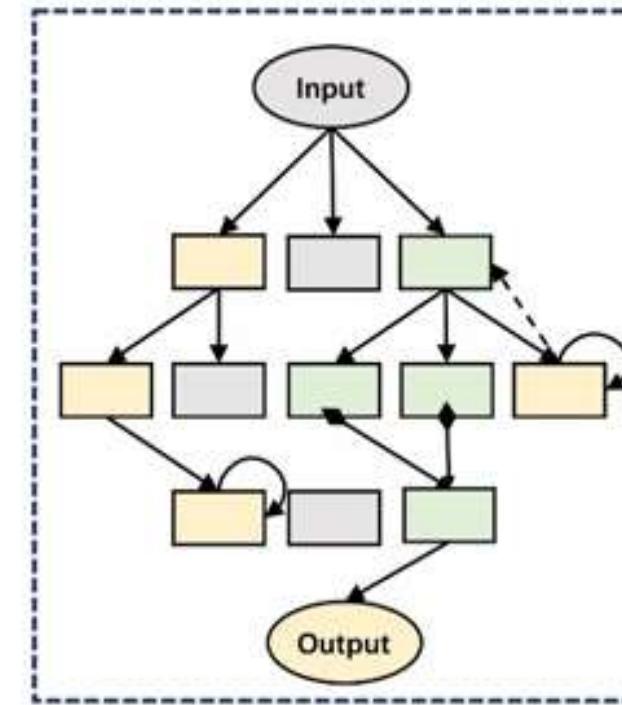
Reasoning with Non-linear Exploration



CoT/PoT



Tree of Thoughts [Yao et al. 2023]



Graph of Thoughts [Besta et al. 2023]



How to Evaluate Free-text/Structured Explanations?

- Faithfulness

How accurately the explanation reflects the true reasoning process of the model?

- Plausibility

How convincing the explanation is to humans?

- Informativeness

How much new information is supplied by a explanation to justify the prediction?

- Utility

How useful is the explanation for the target audience to achieve their predefined goal?

Most method are also applicable to structured explanations, though

- empirically only tested on free-text ones



Evaluation—Faithfulness

Many ways with different assumptions, no consensus yet

- **Counterfactual simulability** [[Chen et al., 2023](#)]
Assumption: *Explanations should allow the audience to predict the model behavior on unseen inputs*
- **Biasing features** [[Turpin et al., 2023](#)]
Assumption: *Features that influence model predictions should be mentioned in the explanations*
- **Corrupting CoT** [[Lanham et al., 2023](#)]
Assumption: *Compared to the original explanation, a corrupted explanation should lead to a different prediction*
- **Input token contribution alignment** [[Parcalabescu and Frank, 2024](#)]
Assumption: *Input token contributions should be similar when the model produces the prediction and the explanation*
- ...

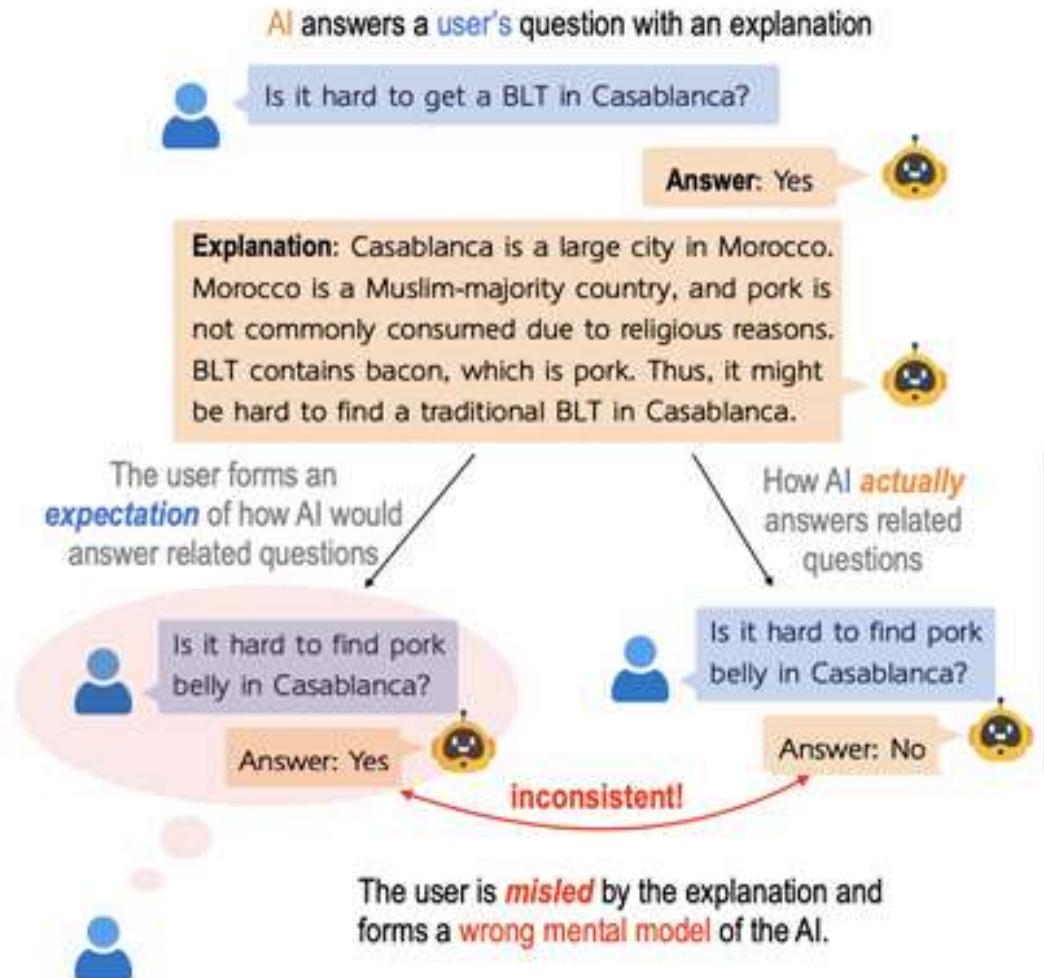


Evaluation—Faithfulness

Example: Counterfactual simulability
[\[Chen et al., 2023\]](#)

Findings:

- LLM-generated free-text explanations are **far from faithful**
- Faithfulness **doesn't correlate well** with plausibility

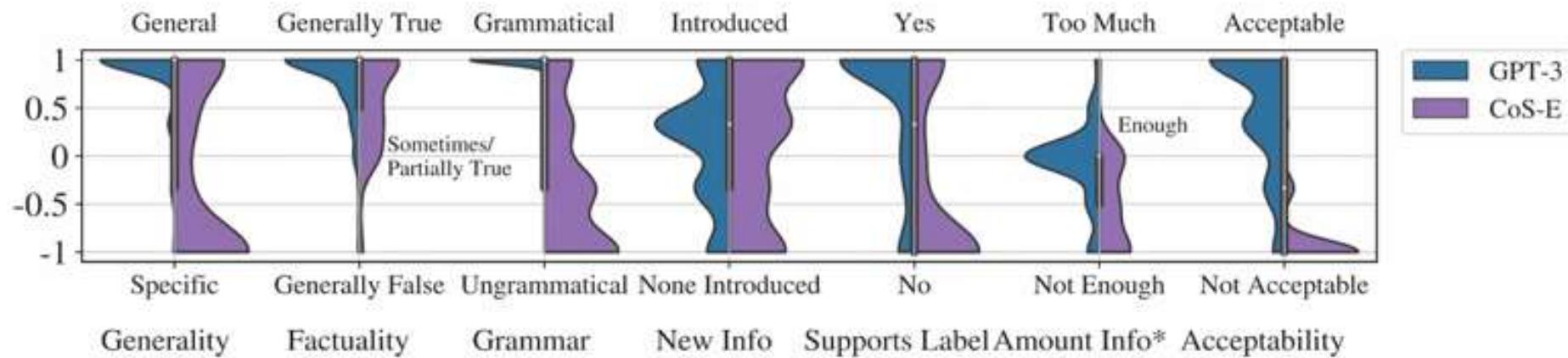


BLT: a type of sandwich, named for the initials of its primary ingredients, bacon, lettuce, and tomato



Evaluation—Plausibility

Annotate LLM-generated explanations with human-written explanations as reference



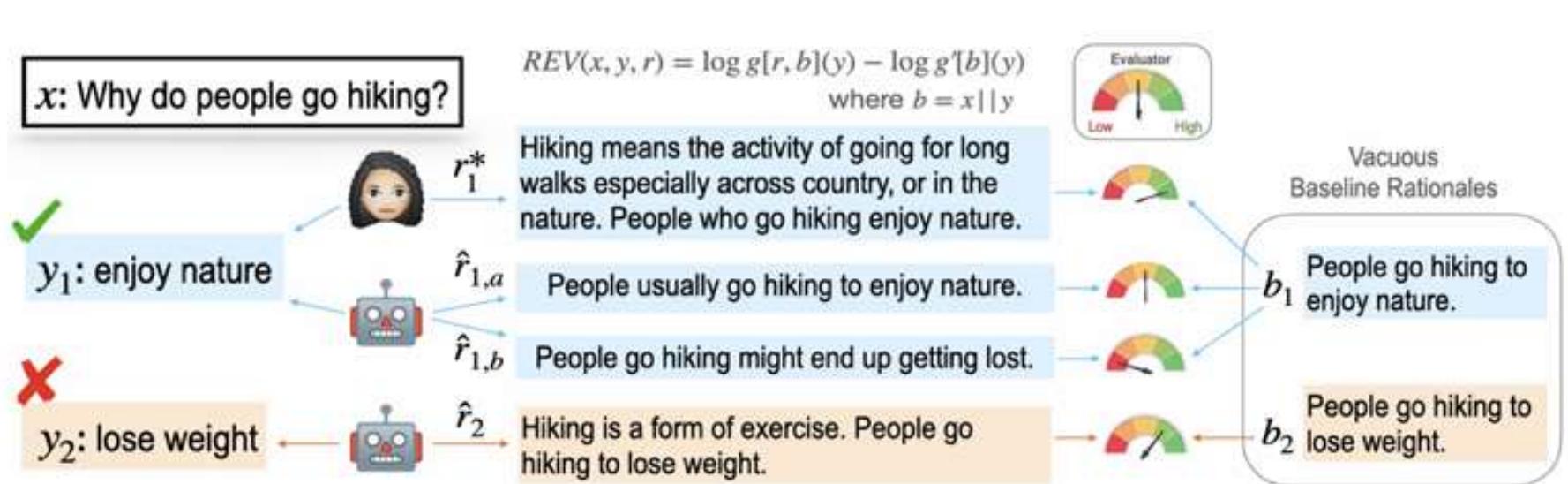
[Wiegreffe et al. 2021]

LLMs can generate plausible explanations, but still have room for improvement compared to human-written ones



Evaluation—Informativeness

Measure the **new information** an explanation provides to justify the label, beyond what is contained in the input, using **conditional V-information**

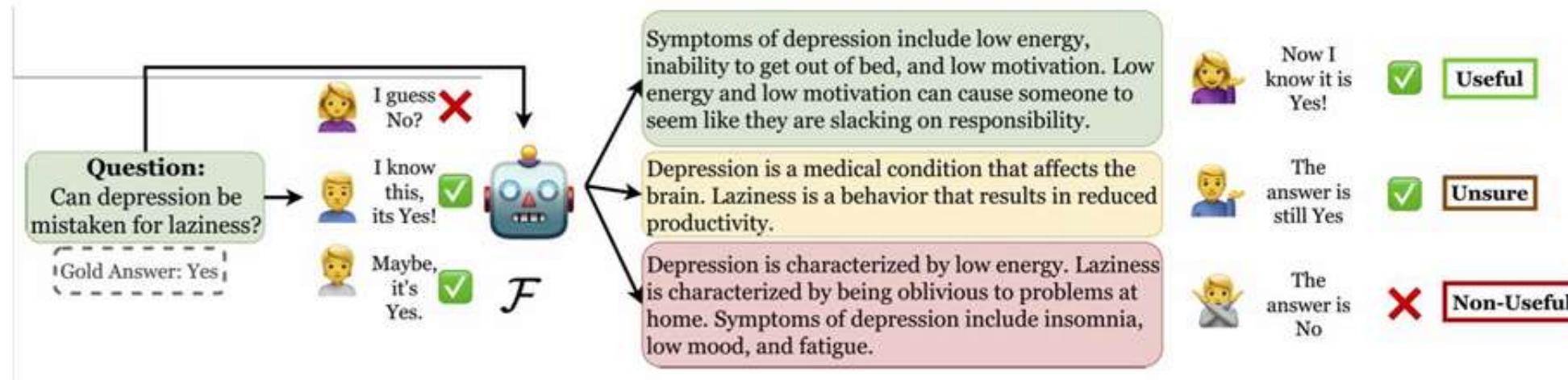


REV [Chen et al. 2023]



Evaluation—Utility

Can LLM-generated explanations help lay people answer **unseen** questions?



Utility is far from satisfactory – only **20%** of generated explanations are actually useful



Pros & Cons

- Extractive rationales / Feature attributions
 - ? Faithfulness
 - - Plausibility
- Free-text explanations
 - + Plausibility
 - - Faithfulness, Utility
- Structured explanations
 - + Faithfulness, Accuracy
 - - Flexibility



Takeaways

- LLMs can generate **plausible**-looking explanations w/ only a few examples
 - this saves the **cost** of collecting human explanations for training
 - and also improves **performance** on many reasoning tasks
- However, LLM-generated explanations are still **not always faithful / informative / useful** ...
 - Not a consensus on how to **evaluate** many of these aspects
- We should not blindly trust LLM-generated explanations
 - Be cautious about “self-explanatory” claims



Future directions

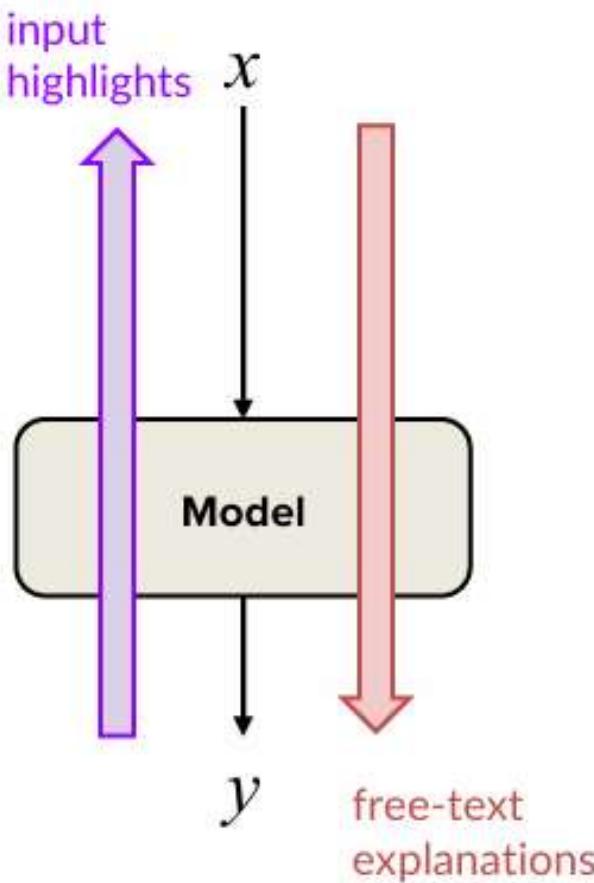
- Establishing a more unified **evaluation framework**
 - esp. for structured explanations
- Applying structured explanations to **flexible** (non-symbolic) tasks
 - e.g. commonsense reasoning, summarization, web browsing ...

Further reading

- A Comprehensive Collection of Explainable NLP Datasets [[Wiegreffe and Marasović 2021](#)]
- A Survey on Chain-of-Thought-style Reasoning [[Chu et al. 2024](#)]
- A Survey on Faithfulness of Explanations in NLP [[Lyu et al. 2024](#)]

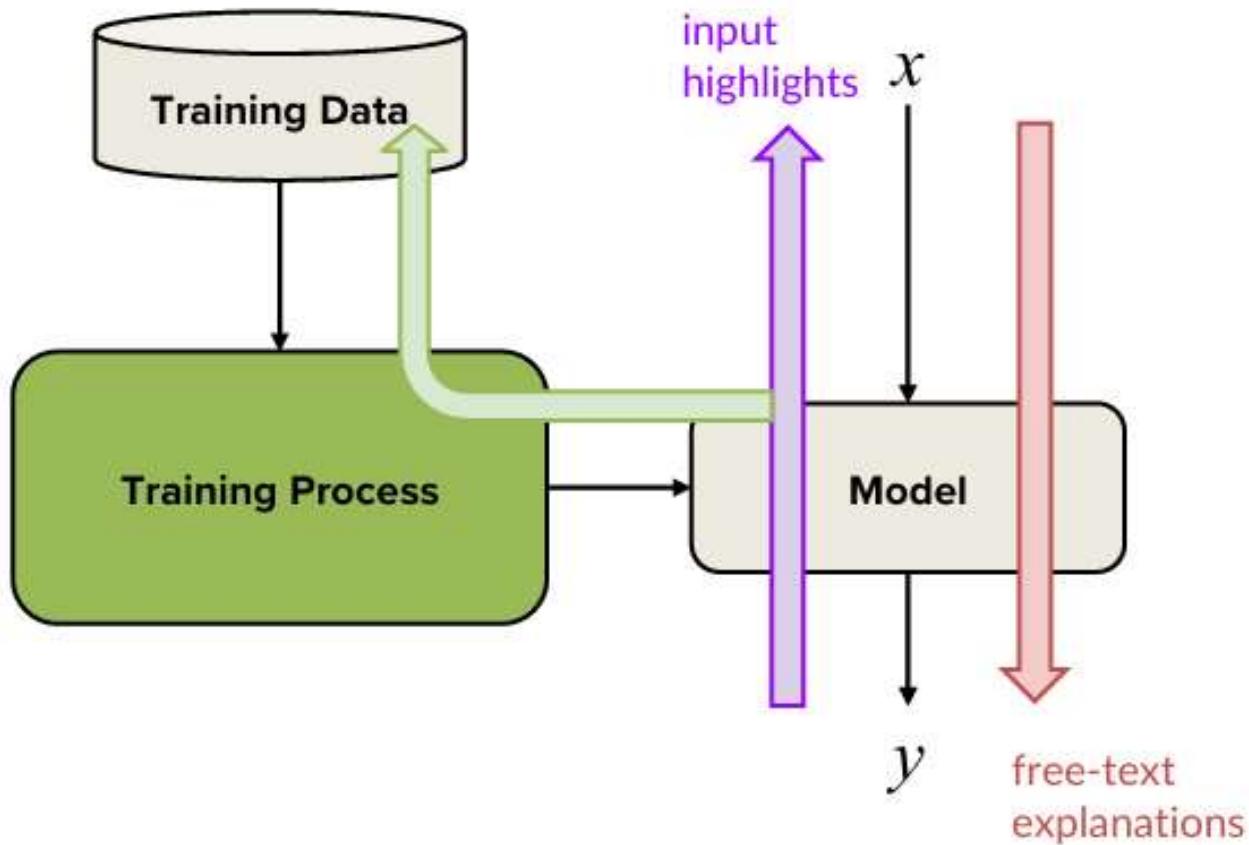


So far...



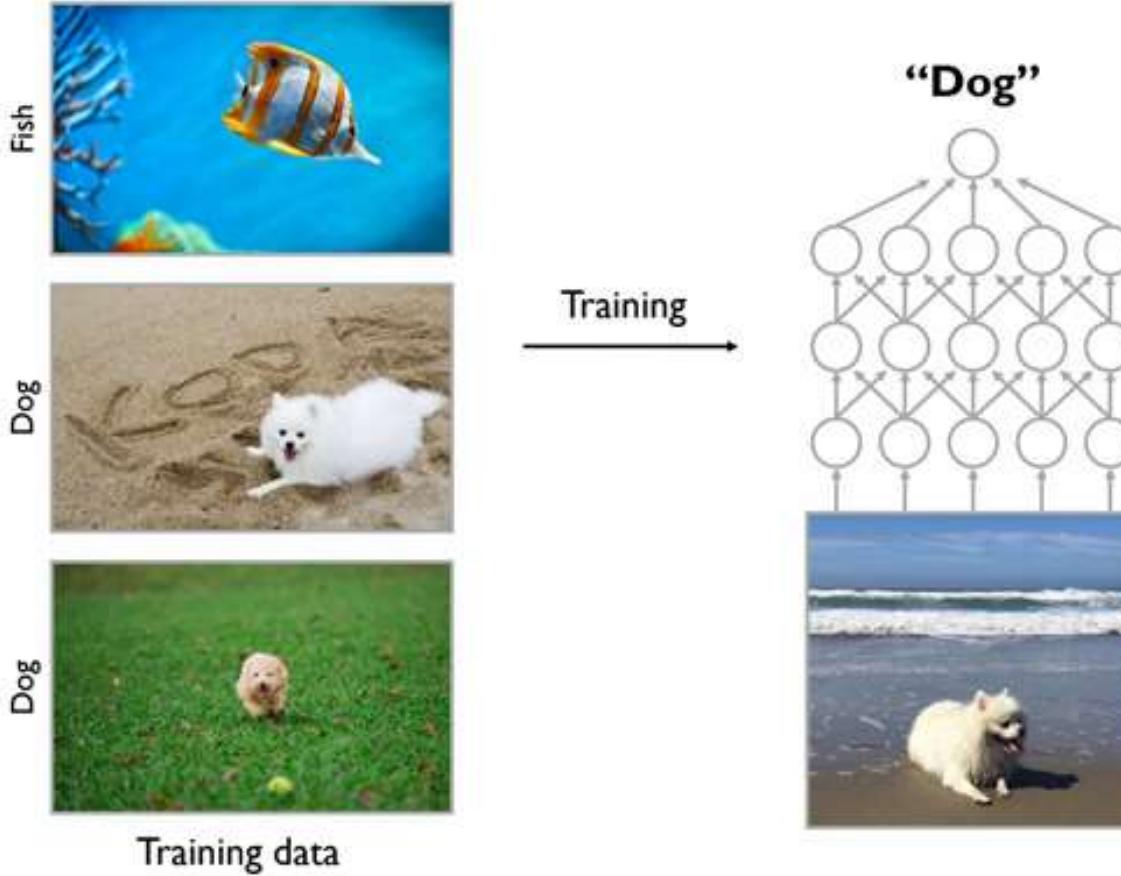


Data Influence





Data Attribution



Which training points were most
responsible for this prediction?

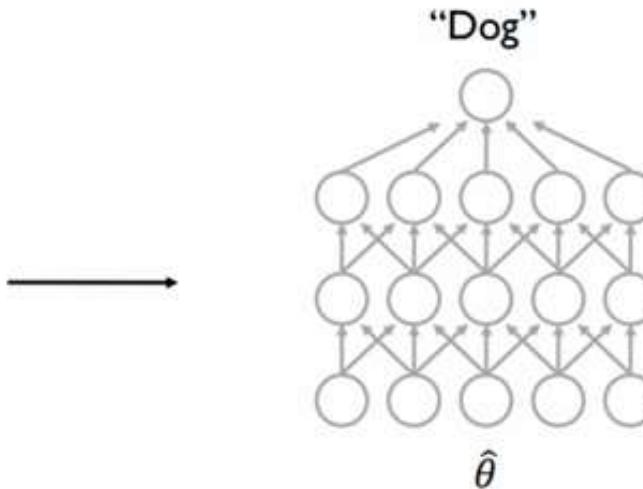


Data Attribution



Influence of a data point: how would the prediction change if we did not have this training point?

Pick $\hat{\theta}$ to minimize $\frac{1}{n} \sum_{i=1}^n L(z_i, \theta)$





Data Attribution



Training data z_1, z_2, \dots, z_n

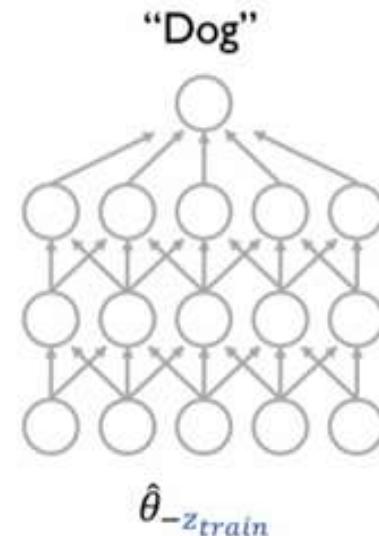
Influence of a data point: how would the prediction change if we did not have this training point?

Pick $\hat{\theta}$ to minimize $\frac{1}{n} \sum_{i=1}^n L(z_i, \theta)$

Pick $\hat{\theta}_{-z_{train}}$ to minimize

$$\frac{1}{n} \sum_{i=1}^n L(z_i, \theta) - \frac{1}{n} L(z_{train}, \theta)$$

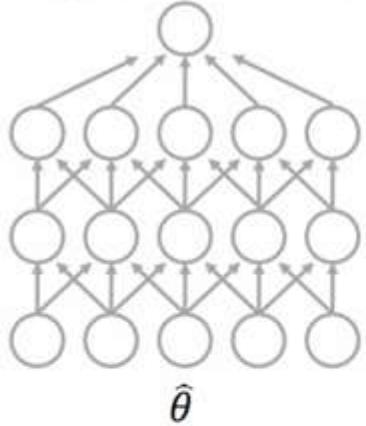
z_{train}





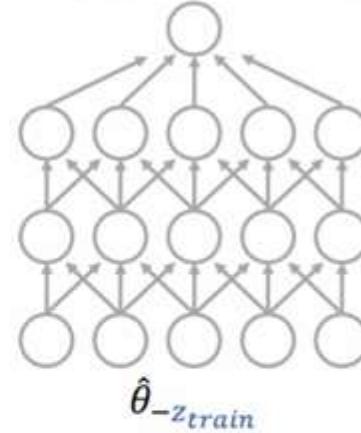
Data Attribution

“Dog” (82% confidence)



vs.

“Dog” (79% confidence)

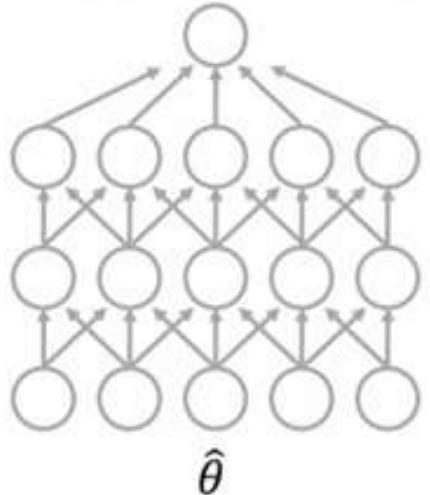


Test input z_{test}



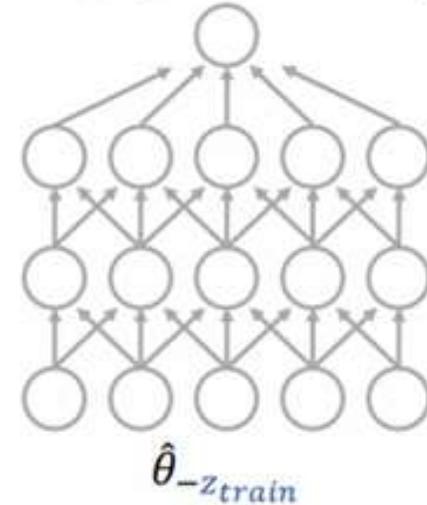
Data Attribution

“Dog” (82% confidence)



vs.

“Dog” (79% confidence)



What is $L(z_{test}, \hat{\theta}_{-z_{train}}) - L(z_{test}, \hat{\theta})$?



Problem

Repeatedly removing a training point and retraining the model is too slow

Solution

Approximation via influence functions
(a classical technique from the 1970s)
[Hampel, 1974; Cook, 1979]



Seminal Work: Influence Functions

For a test example z_{test} , the influence \mathcal{I} of infinitesimally upweighting a training example z_{train} by ε on the value of a scalar-valued twice differentiable function $f(\cdot; \theta)$ is given by:



Seminal Work: Influence Functions

For a test example z_{test} , the influence \mathcal{I} of infinitesimally upweighting a training example z_{train} by ε on the value of a scalar-valued twice differentiable function $f(\cdot; \theta)$ is given by:

$$\mathcal{I}(z_{\text{train}}, f(z_{\text{test}}; \theta^*(\varepsilon; z_{\text{train}}))) = -\nabla_\theta f(z_{\text{test}}; \theta^*)^T H_{\mathcal{J}(\mathcal{D}; \theta^*)}^{-1} \nabla_\theta \mathcal{L}(z_{\text{train}}; \theta^*)$$



Seminal Work: Influence Functions

For a test example z_{test} , the influence \mathcal{I} of infinitesimally upweighting a training example z_{train} by ε on the value of a scalar-valued twice differentiable function $f(\cdot; \theta)$ is given by:

$$\mathcal{I}(z_{\text{train}}, f(z_{\text{test}}; \underbrace{\theta^*(\varepsilon; z_{\text{train}})}_{\text{Optimal parameters on the training dataset with } z_{\text{train}} \text{-upweighted}})) = -\nabla_{\theta} f(z_{\text{test}}; \theta^*)^T H_{\mathcal{J}(\mathcal{D}; \theta^*)}^{-1} \nabla_{\theta} \mathcal{L}(z_{\text{train}}; \theta^*)$$

Optimal parameters on the training dataset with z_{train} -upweighted Optimal parameters on the original training dataset $\mathcal{D} = \{z_i\}_{i=1}^N = \{(x_i, y_i)\}_{i=1}^N$ Hessian Cost function

$$\mathcal{J}(\mathcal{D}; \theta) := \frac{1}{N} \sum_{i=1}^N \mathcal{L}(z_i; \theta)$$



Seminal Work: Influence Functions

For a test example z_{test} , the influence \mathcal{I} of infinitesimally upweighting a training example z_{train} by ε on the value of a scalar-valued twice differentiable function $f(\cdot; \theta)$ is given by:

$$\mathcal{I}(z_{\text{train}}, f(z_{\text{test}}; \underbrace{\theta^*(\varepsilon; z_{\text{train}})}_{\text{Optimal parameters on the training dataset with } z_{\text{train}} \text{ } \varepsilon\text{-upweighted}})) = -\nabla_{\theta} f(z_{\text{test}}; \theta^*)^T H_{\mathcal{J}(\mathcal{D}; \theta^*)}^{-1} \nabla_{\theta} \mathcal{L}(z_{\text{train}}; \theta^*)$$

Optimal parameters on the original training dataset $\mathcal{D} = \{z_i\}_{i=1}^N = \{(x_i, y_i)\}_{i=1}^N$

Hessian

Cost function

Loss

$$\mathcal{J}(\mathcal{D}; \theta) := \frac{1}{N} \sum_{i=1}^N \mathcal{L}(z_i; \theta)$$

Notice: On the right side, there is no $\theta^*(\varepsilon; z_{\text{train}})$ \Rightarrow **The model is not re-trained** with ε -upweighting of z_{train}

Upweighting vs. Removing: $\varepsilon = -\frac{1}{N}$ approximates the effect of removing z_{train}



On which functions is the effect of removing an example studied?

$$\mathcal{I}(z_{\text{train}}, f(z_{\text{test}}; \theta^*(\varepsilon; z_{\text{train}}))) = -\nabla_{\theta} f(z_{\text{test}}; \theta^*)^T H_{\mathcal{J}(\mathcal{D}; \theta^*)}^{-1} \nabla_{\theta} \mathcal{L}(z_{\text{train}}; \theta^*)$$

Prior to LLMs, how the **loss for a given test instance** changes

$f(\cdot; \theta)$ is set to the loss function

Today, how the **likelihood of a given completion** z_c for a prompt z_p

$f(\cdot; \theta)$ is set to $\log p(z_c; z_p; \theta)$



Seminal Work: Influence Functions - Assumptions

$$\mathcal{I}(z_{\text{train}}, f(z_{\text{test}}; \theta^*(\varepsilon; z_{\text{train}}))) = -\nabla_\theta f(z_{\text{test}}; \theta^*)^T H_{\mathcal{J}(\mathcal{D}; \theta^*)}^{-1} \nabla_\theta \mathcal{L}(z_{\text{train}}; \theta^*)$$

1. The function is twice differentiable
2. $\theta^* := \operatorname{argmin}_\theta \mathcal{J}(\mathcal{D}; \theta)$ exists
3. θ^* is unique

To satisfy 2 and 3, it is assumed that the cost function $\mathcal{J}(\mathcal{D}; \theta)$ is strictly convex w.r.t. the parameters, which is **often not the case for neural networks**



Challenges to Hessian Calculation

$$\mathcal{I}(z_{\text{train}}, f(z_{\text{test}}; \theta^*(\varepsilon; z_{\text{train}}))) = -\nabla_\theta f(z_{\text{test}}; \theta^*)^T H_{\mathcal{J}(\mathcal{D}; \theta^*)}^{-1} \nabla_\theta \mathcal{L}(z_{\text{train}}; \theta^*)$$

There are two main challenges to computing the inverse of the Hessian:

1. **The Hessian of loss functions for neural networks can be nonpositive semidefinite**
⇒ Inverse cannot be computed
 - a. Damping
 - b. Gauss-Newton Hessian
2. **Hessian is square w.r.t. the model parameters**
⇒ It is expensive to compute: The standard inversion algorithm has a time complexity of $\mathcal{O}(D^3)$
 - a. Iterative methods
 - b. K-FAC & EK-FAC



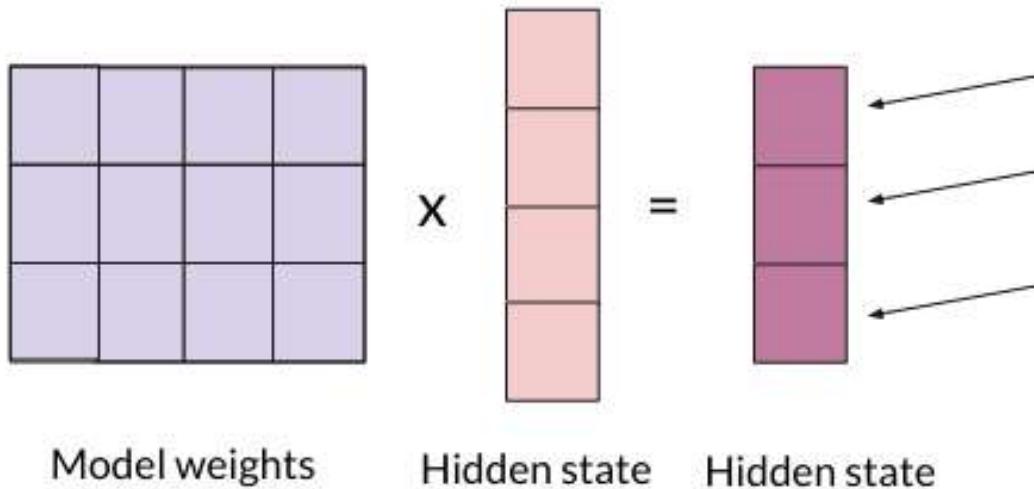
1. Neuron-level interpretability
 - a. Sparse Autoencoders
2. Causal Mediation
 - a. Activation Patching & variants
 - b. Causal abstraction & other methods
3. What is mechanistic interpretability?
4. Methods Leveraging Language Model Strengths
 - a. Transformer Residual Stream and Linear Structure
 - b. Vocabulary projection
 - c. Decoding Natural Language Explanations from Representations



Neuron-Level Interpretability – Background

◆ 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)?



Neuron-Level Interpretability – Background

◆ Interpreting Neurons of NLP Models

Activations of neurons for certain properties

Supports the efforts of the Libyan authorities to recover funds misappropriated under the Qadhafi regime

(a) English Verb (#1902)

einige von Ihnen haben vielleicht davon gehört , dass ich vor ein paar Wochen eine Anzeige bei Ebay geschaltet habe .

(b) German Article (#590)

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.



Neuron-Level Interpretability – Background

◆ Pitfalls of Neuron-Level Analysis in NLP

- Methods generally ignore interactions between neurons.
- There are a LOT of neurons in modern models.



Neuron-Level Interpretability – Background

◆ 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?



Neuron-Level Interpretability – Background

◆ 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?

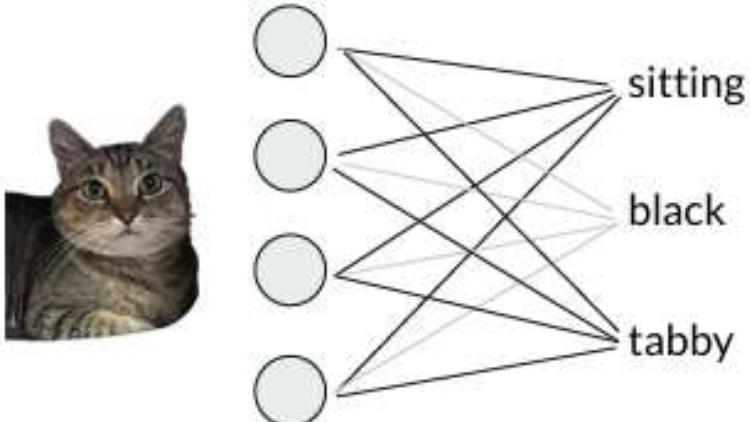
- **Polysemy:** neurons "respond to multiple unrelated inputs"



Neuron-Level Interpretability – Sparse Autoencoders

- ◆ Linear Combinations of Neurons as Concepts

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



Hypothesis

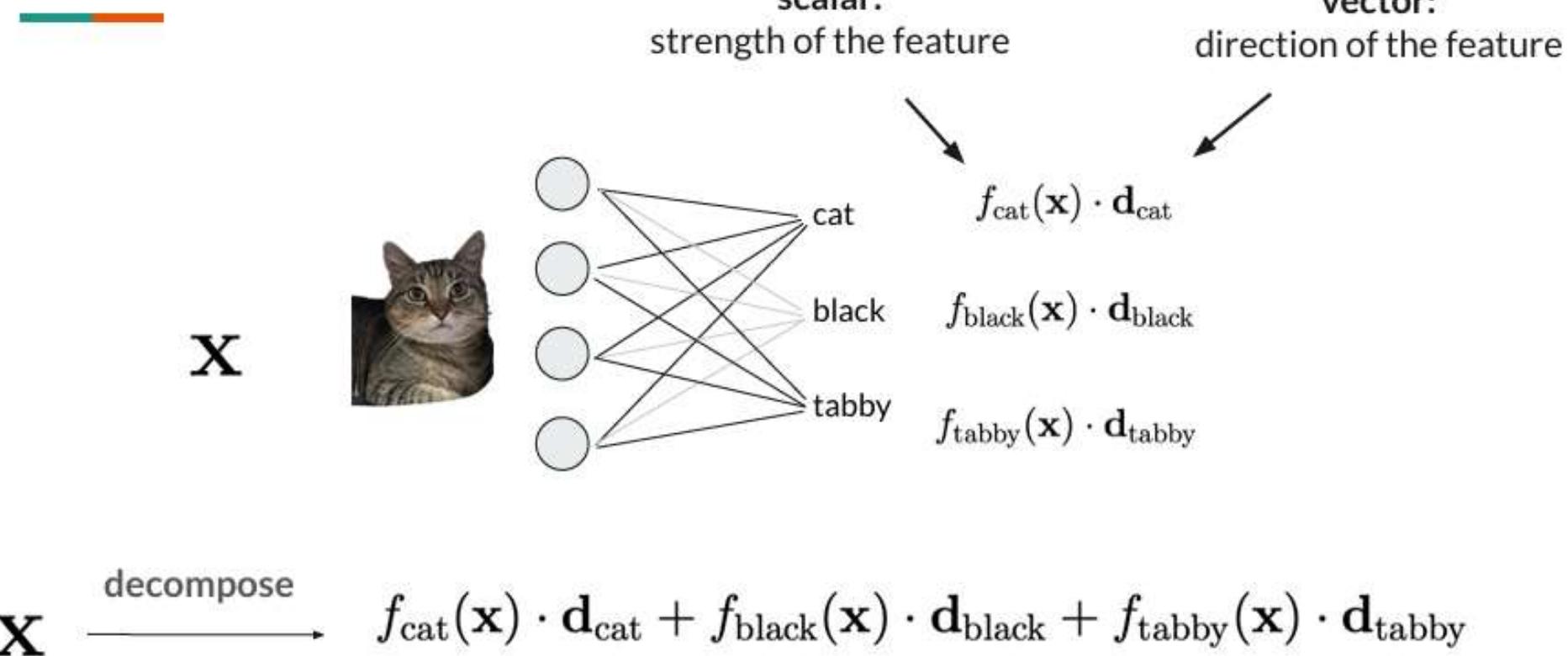
Neurons together (as opposed to individual neurons)
respond to concepts

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



Neuron-Level Interpretability – Sparse Autoencoders

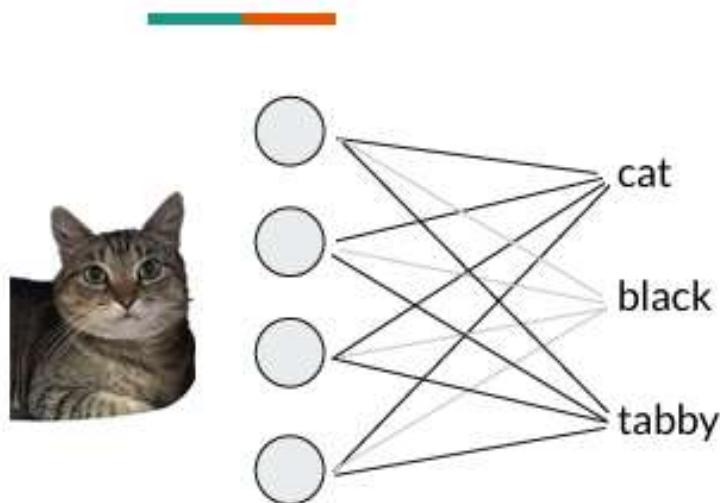
◆ Decomposing Activations





Neuron-Level Interpretability – Sparse Autoencoders

- ◆ Decomposing Activations with Sparse Autoencoders



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

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

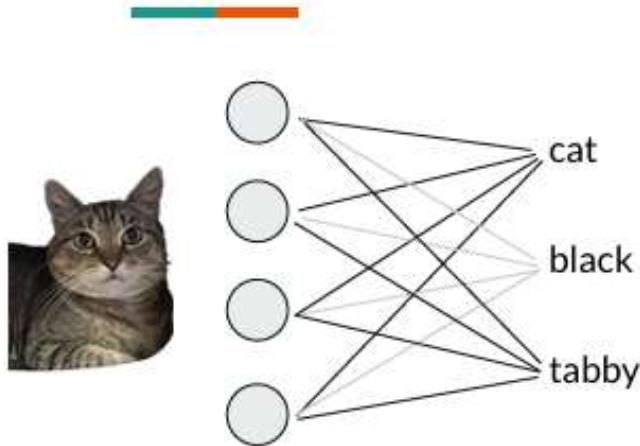
using **Sparse Autoencoders** to find decompositions

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



Neuron-Level Interpretability – Sparse Autoencoders

- ◆ 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

reconstruction loss sparsity

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

[Anthropic, 2023]



Neuron-Level Interpretability – Sparse Autoencoders

◆ Notes on Learned Features



Example: IM/3 **Transit infrastructure**

cross one particular bridge, which is a massive en route. Since the underwater tunnel between) on the approaches to bridges/tunnels and it continue north across the aqueduct toward Wrexham the case for the Transbay Tube which requires

[Anthropic, 2024]

[Anthropic, 2023]

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 clamped to 5x its max**
It confabulates a bridge

F#1M/3

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

Assistant: 1. Walk across the bridge.



Neuron-Level Interpretability – Sparse Autoencoders

◆ Notes on Learned Features

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

cross one particular bridge, which is a massive concrete structure. Since the underwater tunnel between London and the Isle of Wight is still under construction, traffic must travel via a bridge. On the approaches to the bridge/tunnel, there are several smaller structures, such as viaducts and flyovers. The Transbay Tube is a proposed rail link between San Francisco and Oakland, which would involve crossing the San Francisco Bay via a large underwater tunnel.

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]



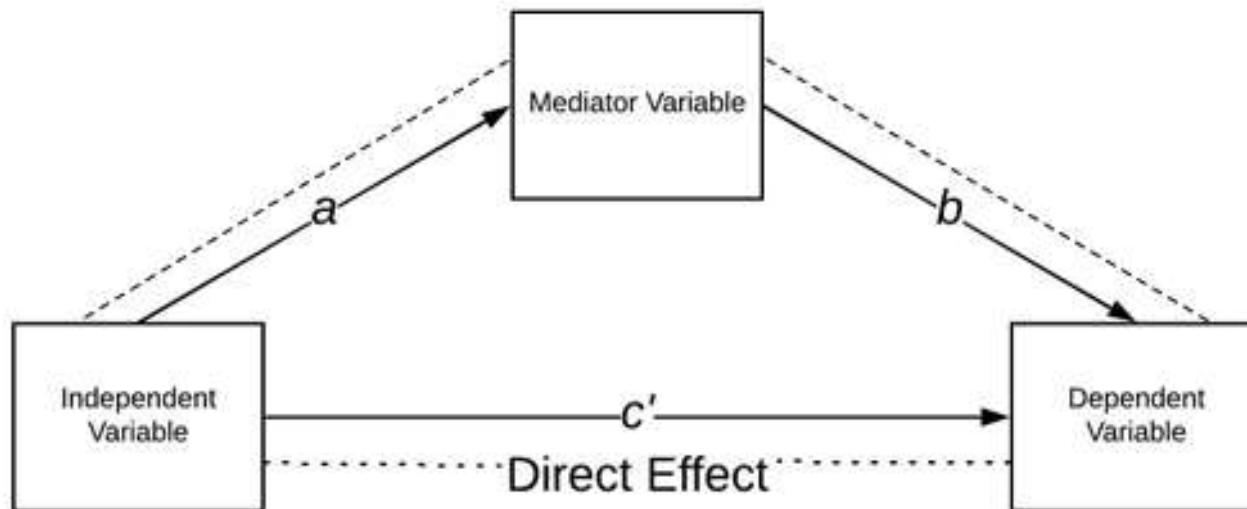
Neuron-Level Interpretability – Causal Mediation in Transformers

◆ Causal Mediation



Indirect Effect = ab

Total Effect = $ab + c'$

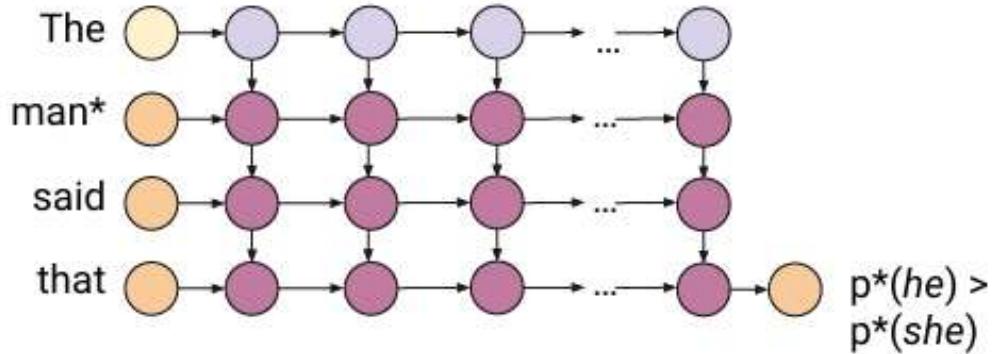
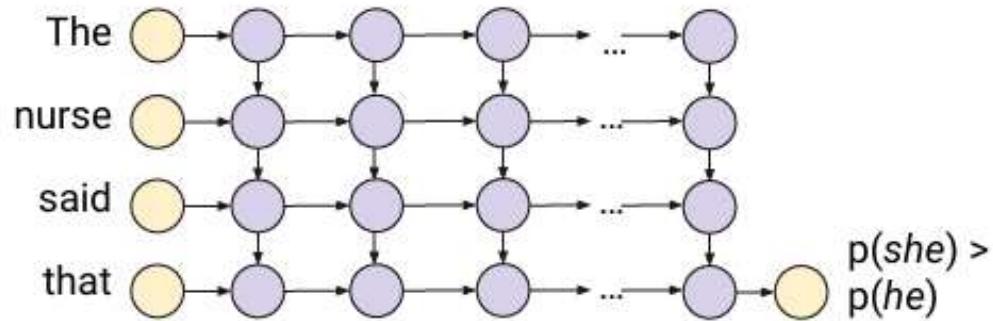




Neuron-Level Interpretability – Causal Mediation in Transformers

◆ Activation Patching/Causal Tracing

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



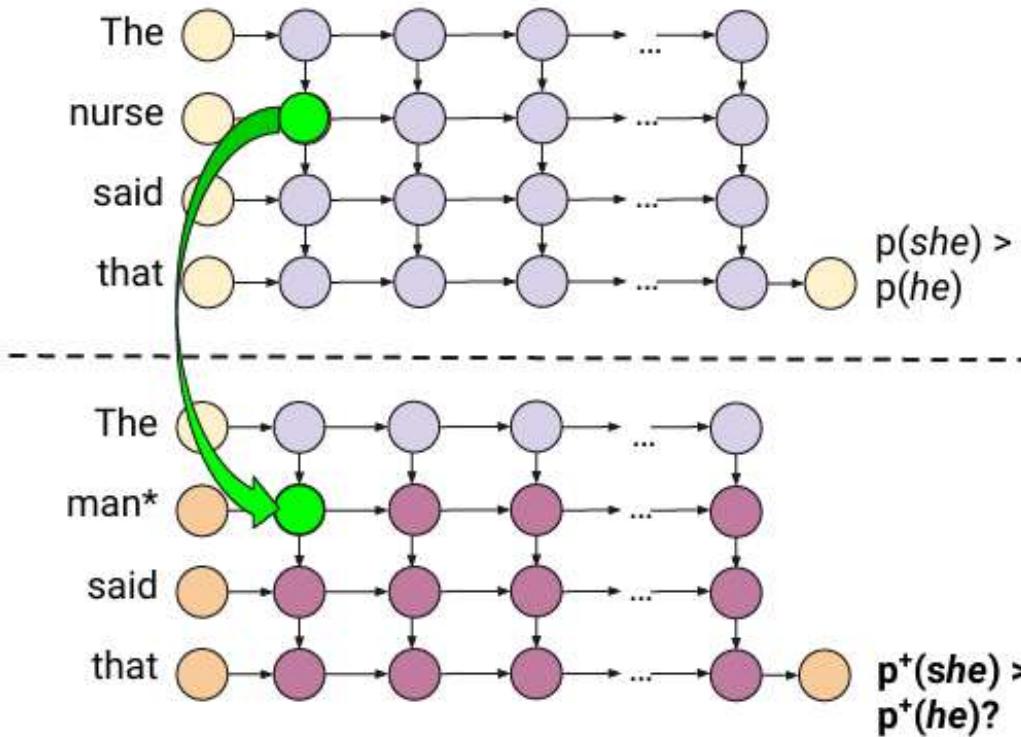
[Vig et al. 2020]



Neuron-Level Interpretability – Causal Mediation in Transformers

◆ 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.

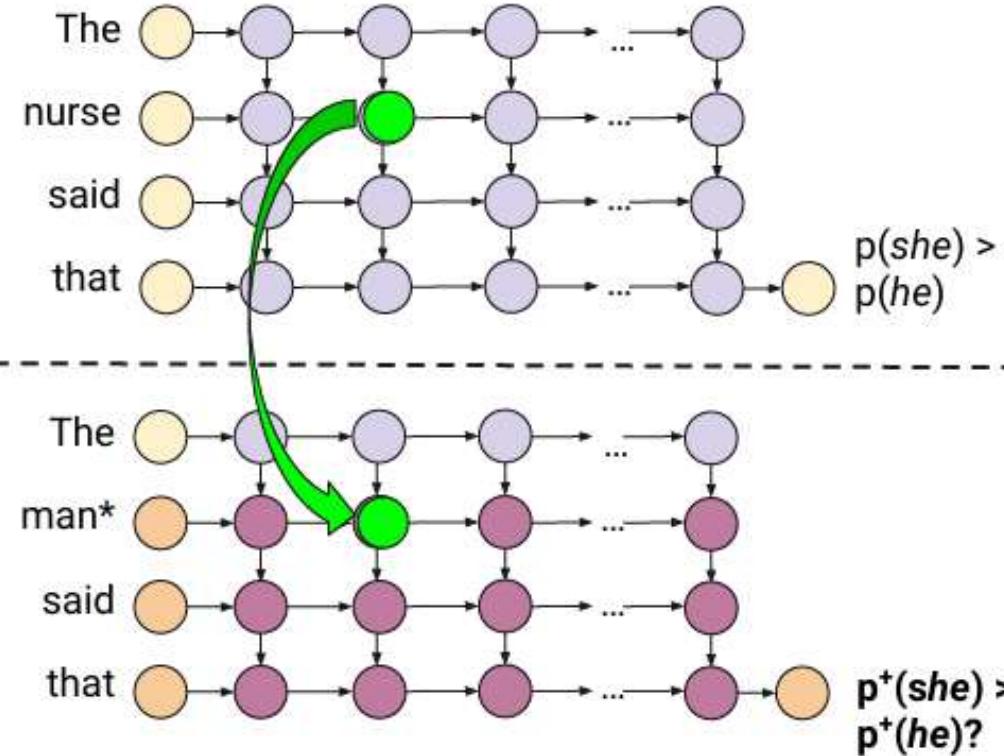




Neuron-Level Interpretability – Causal Mediation in Transformers

◆ 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.

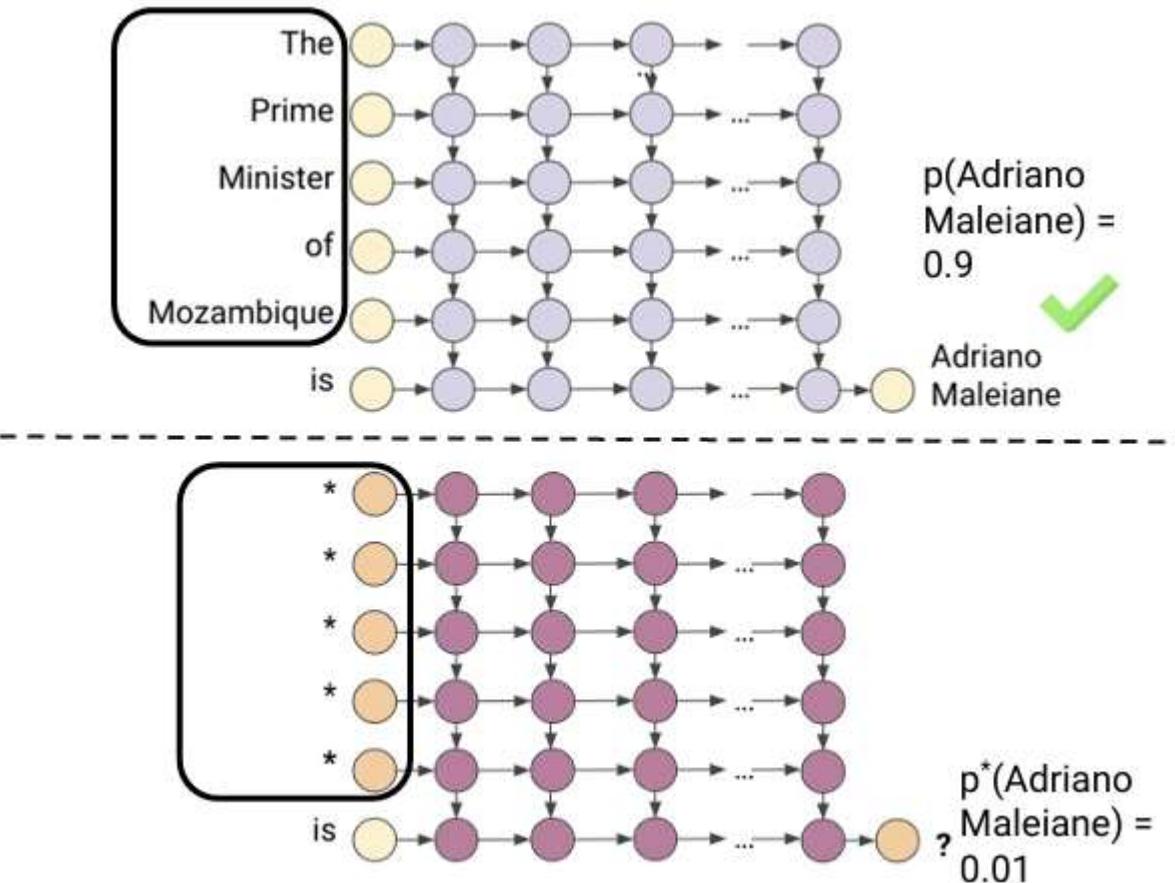




Neuron-Level Interpretability – Causal Mediation in Transformers

◆ Method

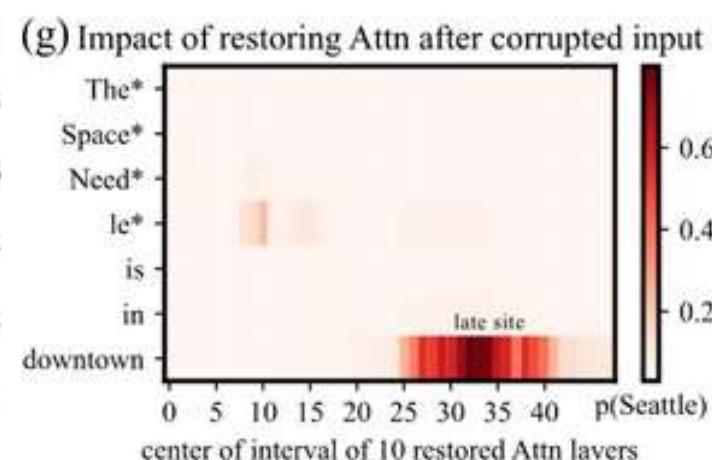
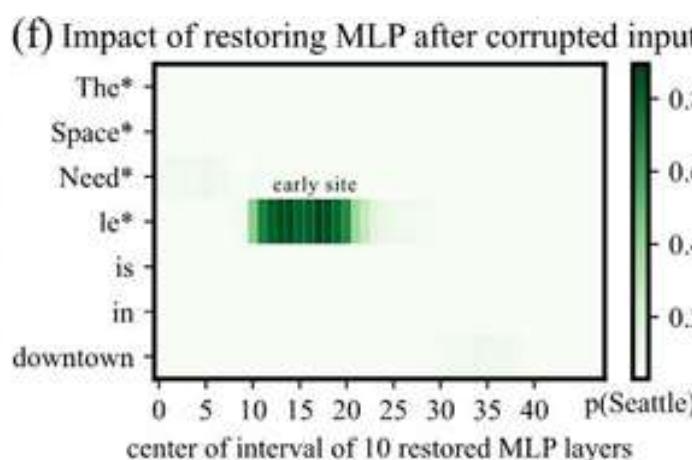
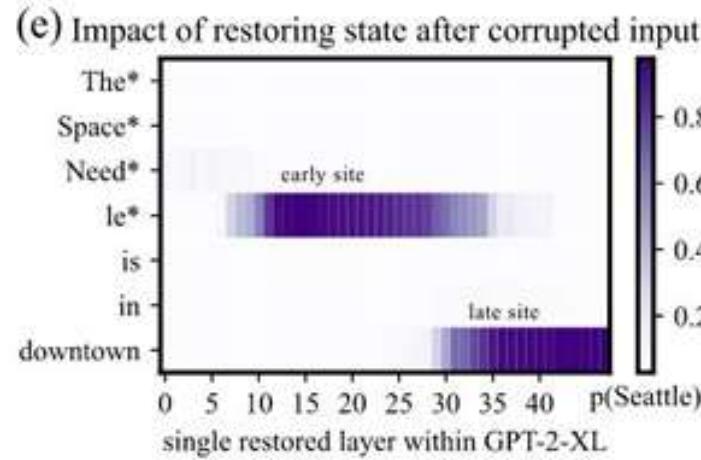
$$\text{Textual Effect} = p(y) - p^*(y) = 0.89$$





Neuron-Level Interpretability – Causal Mediation in Transformers

◆ Results

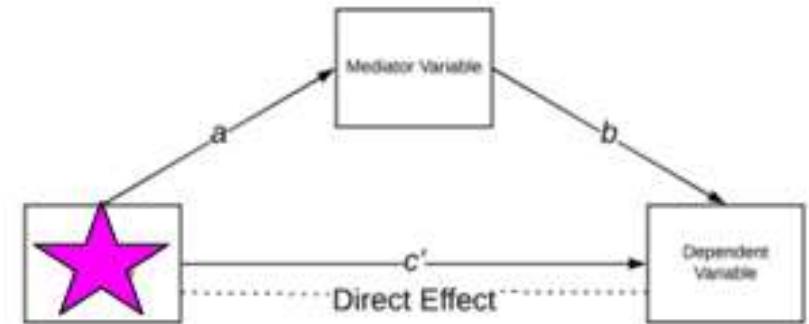




Neuron-Level Interpretability – Causal Mediation in Transformers

◆ 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](#)]

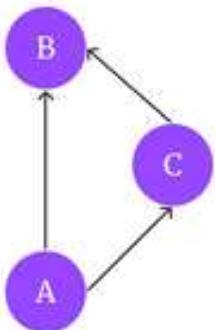




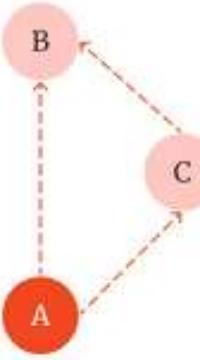
Neuron-Level Interpretability – Causal Mediation in Transformers

◆ Path Patching

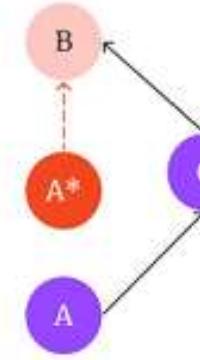
- Controls more carefully which effect you can measure



(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.

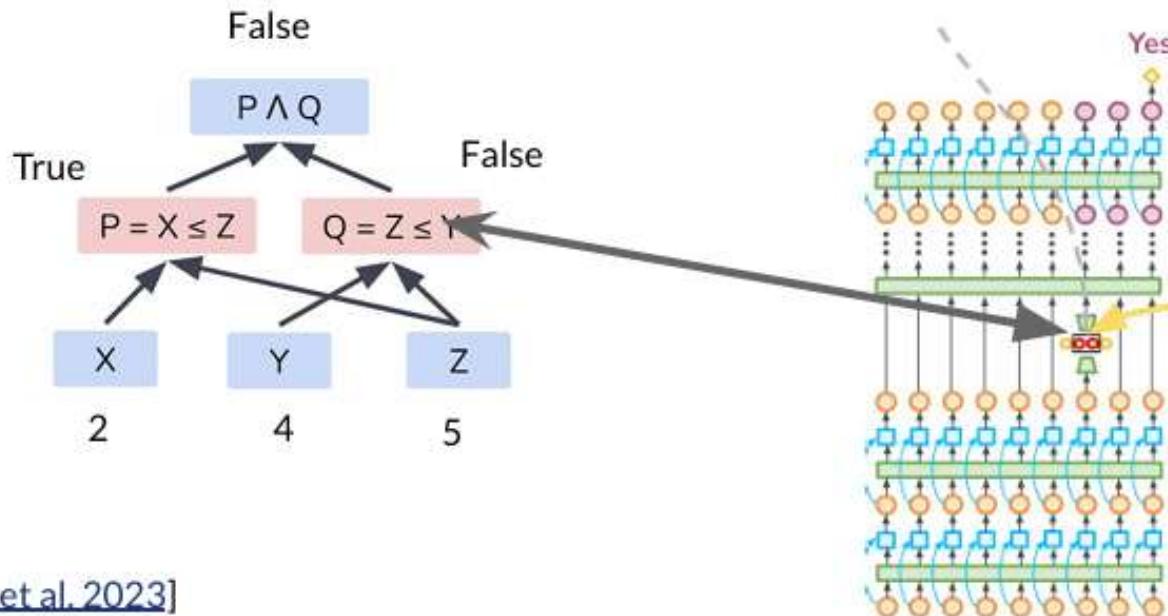


Neuron-Level Interpretability – Causal Mediation in Transformers

◆ 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 $[X, Y]$





Neuron-Level Interpretability – Causal Mediation in Transformers

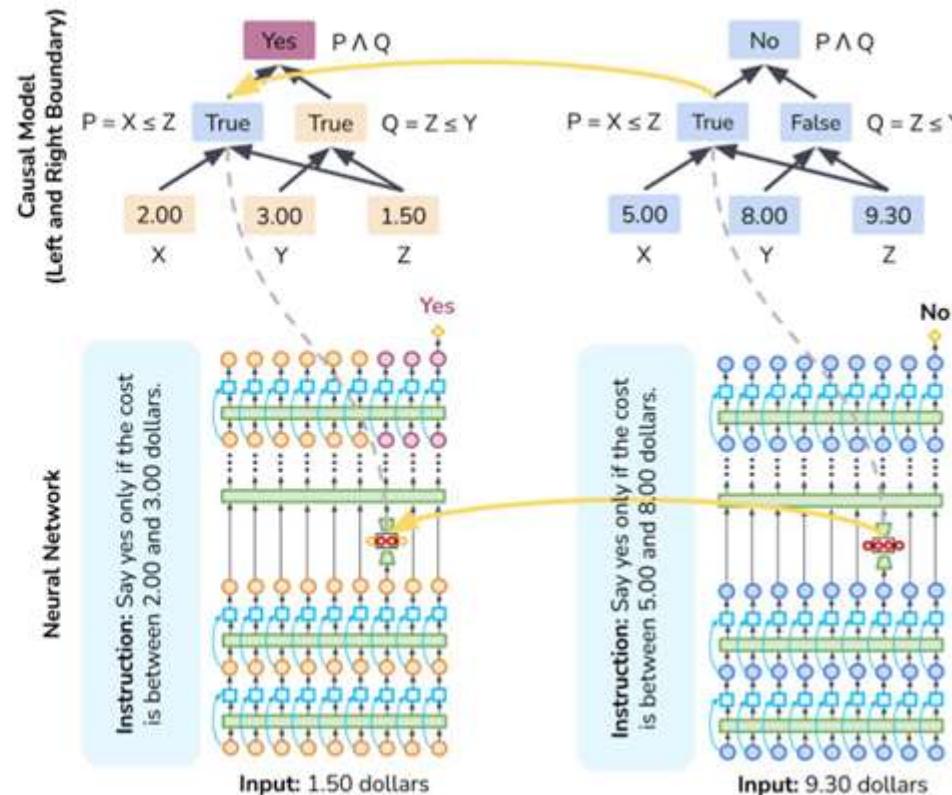
- ◆ Intervention analyses is essential to causal abstraction



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

Coming up with a high-level causal graph is highly non-trivial in practice!

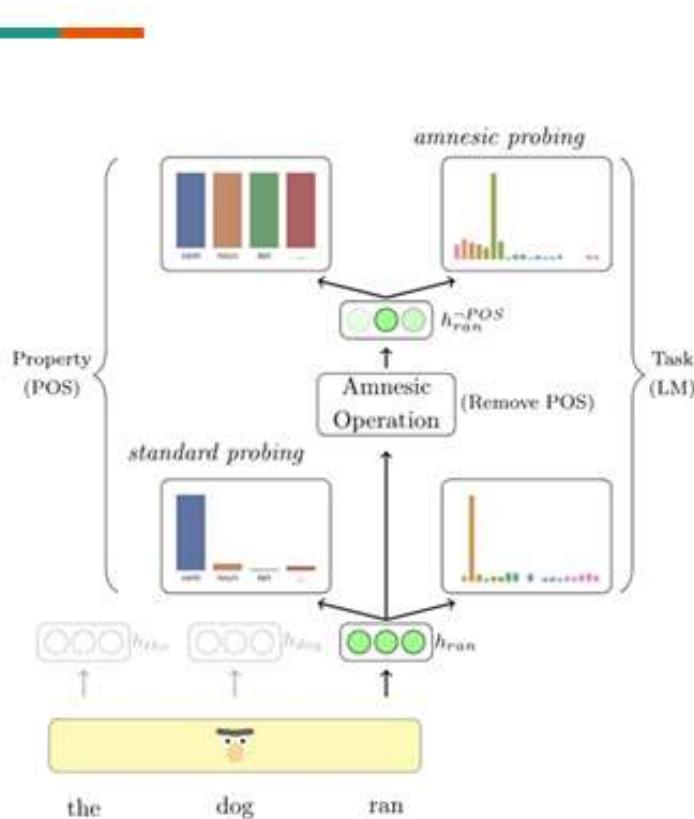
[Geiger et al 2023, Wu et al. 2023]





Neuron-Level Interpretability – Causal Mediation in Transformers

◆ Causal Probing



- Traditional probing classifiers are not causal.
- 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](#)]



Neuron-Level Interpretability – Causal Mediation in Transformers

◆ Linear Subspace Projections + Concept Erasure

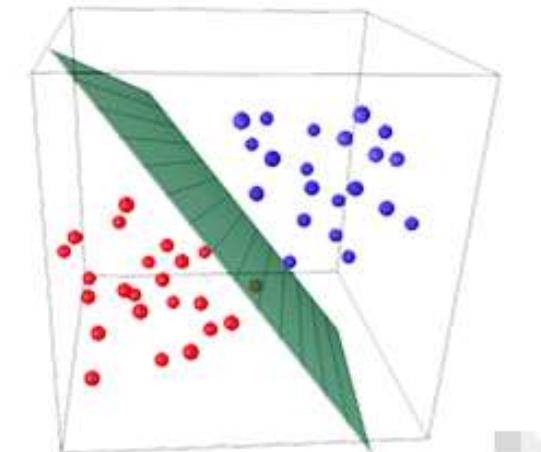
- Models encode many interpretable concepts linearly.

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](#)





Mechanistic Interpretability

◆ 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



Mechanistic Interpretability

◆ 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).

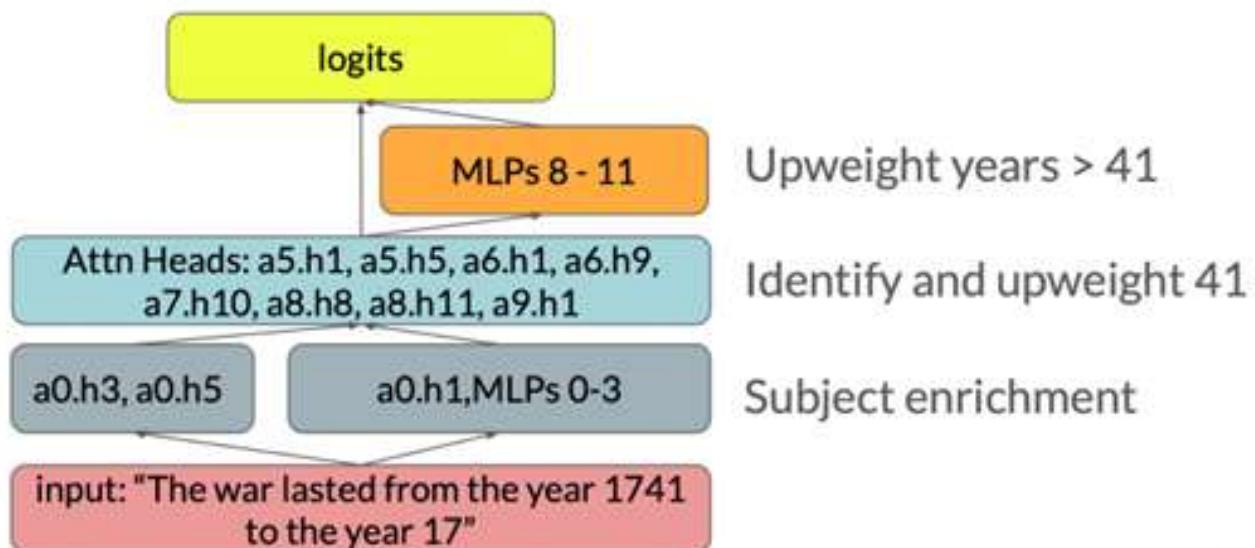


Mechanistic Interpretability

◆ 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



Mechanistic Interpretability

◆ What is Mechanistic Interpretability?

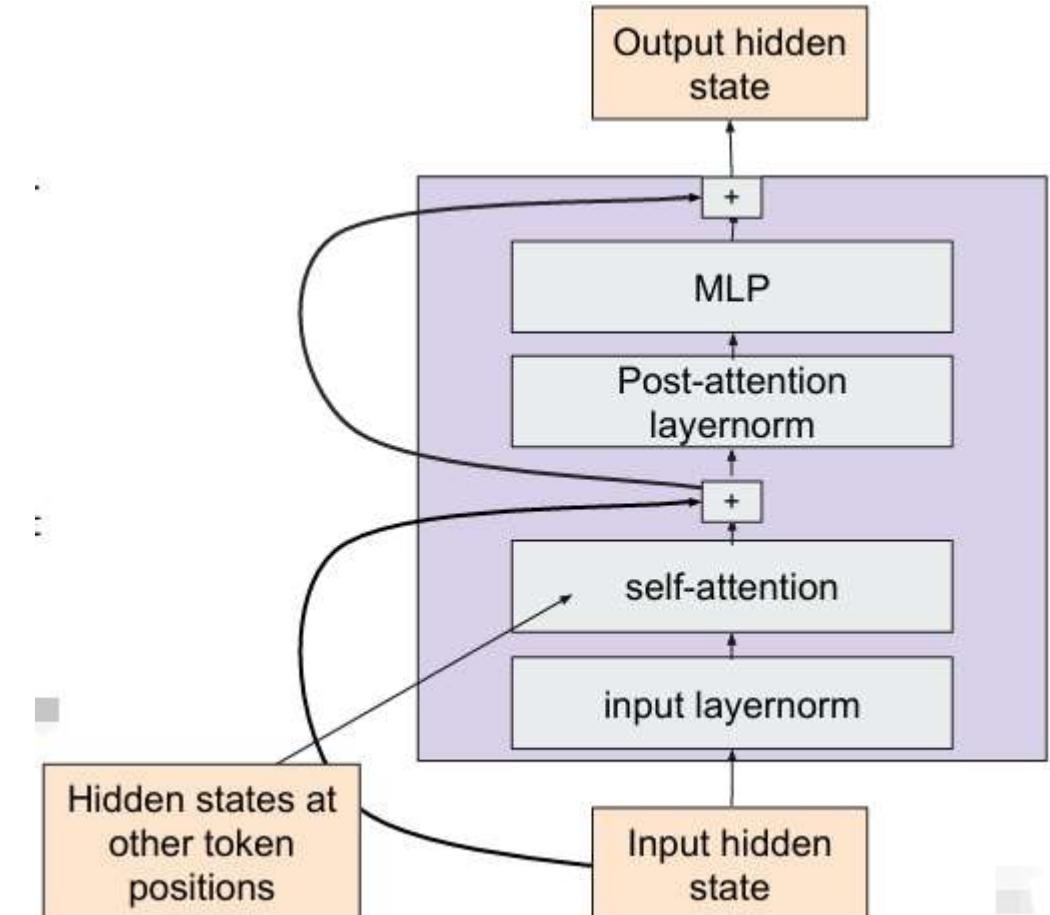
- It is **inherently causal**.
 - NB! This is **not** how most people use the terminology today.
- It is **not the only** set of causal interpretability methods.
- Traces through the entire network (**from starting representation to prediction**).
- Evaluation:
 - 1) **Faithfulness**: the circuit or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



Methods Leveraging Language Model Strengths - Linear Structure in Transformers

◆ 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.





Methods Leveraging Language Model Strengths - Linear Structure in Transformers

◆ 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 $\ell \in [1, L]$):

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

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



Methods Leveraging Language Model Strengths - Linear Structure in Transformers

◆ 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 $\ell \in [1, L]$):

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

Output of
previous layer
= input to
current layer

Multi-head
self-attention

Layer norm
(or some
other input
normalization
scheme)

Feed-forward
network
(MLP)

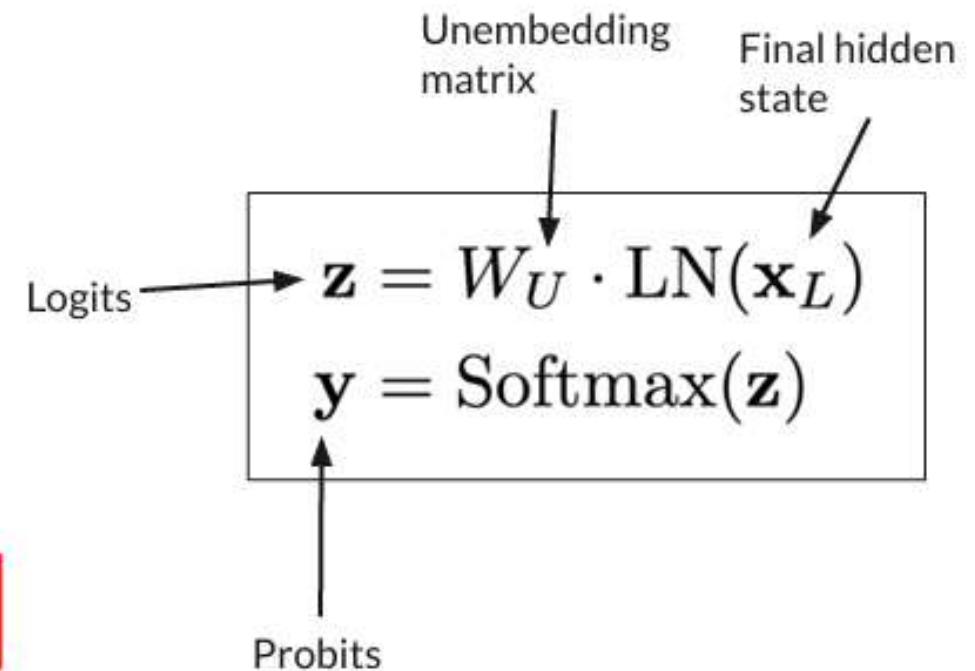
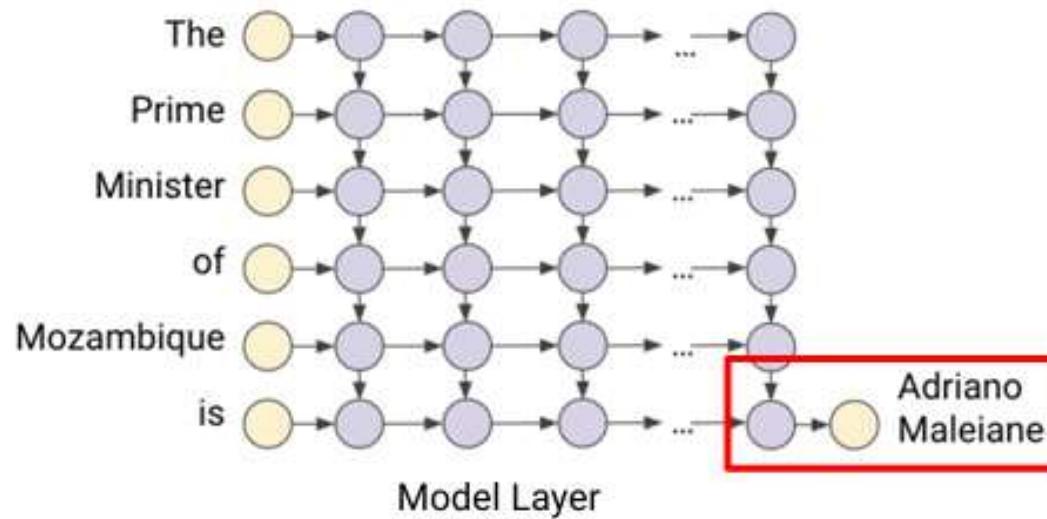
Vector addition
establishes residual
connections

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



Methods Leveraging Language Model Strengths - Linear Structure in Transformers

◆ Transformer Residual Stream and Linear Structure





Methods Leveraging Language Model Strengths - Linear Structure in Transformers

◆ 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.**

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



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



Methods Leveraging Language Model Strengths - Linear Structure in Transformers

◆ 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.

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

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



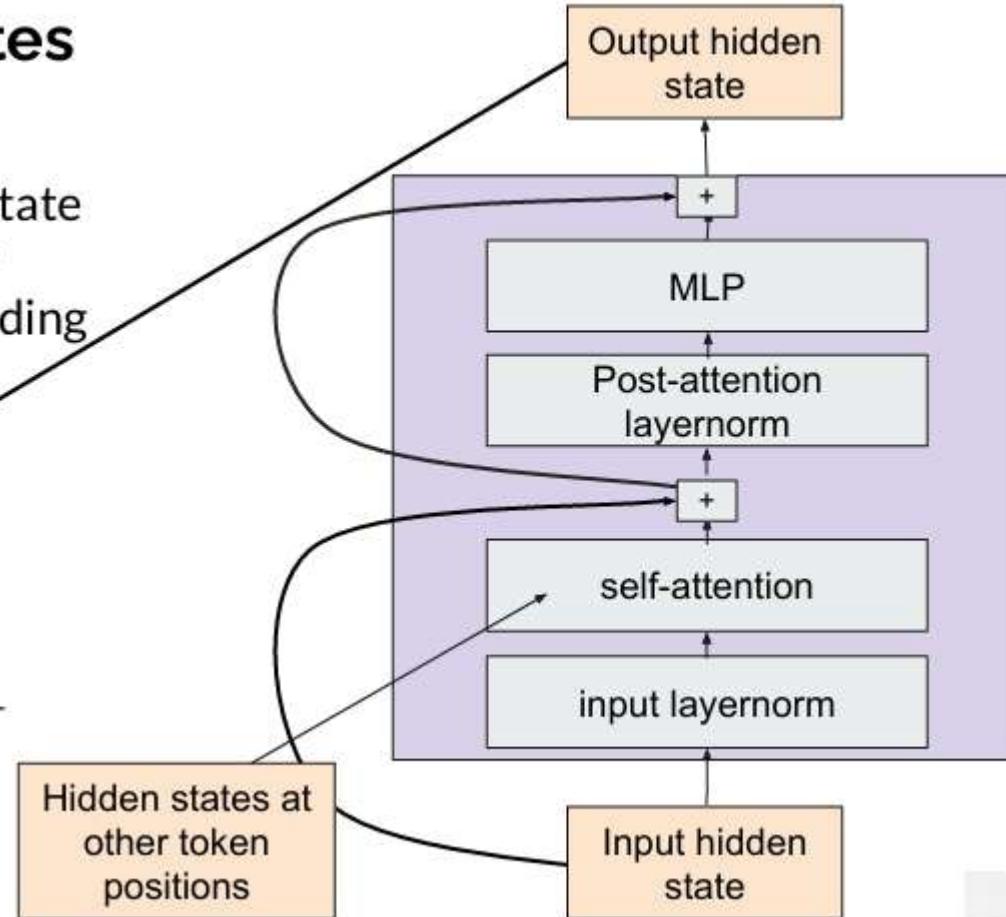
Methods Leveraging Language Model Strengths - Vocabulary Projection

◆ 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.





Methods Leveraging Language Model Strengths - Vocabulary Projection

◆ Vocabulary Projection on Transformer Hidden States

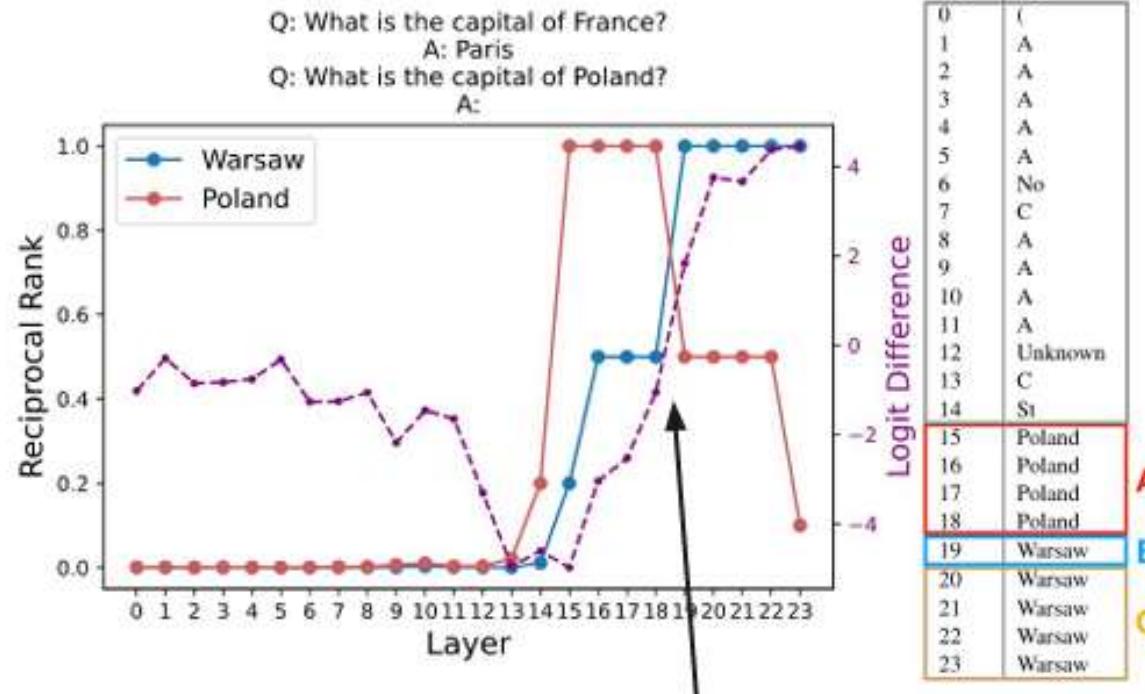
	Concept	Sub-update top-scoring tokens
GPT2	v_{1018}^3 Measurement semantic	kg, percent, spread, total, yards, pounds, hours
	v_{1900}^8 WH-relativizers syntactic	which, whose, Which, whom, where, who, wherein
	v_{2601}^{11} Food and drinks semantic	drinks, coffee, tea, soda, burgers, bar, sushi
WIKILM	v_1^1 Pronouns syntactic	Her, She, Their, her, she, They, their, they, His
	v_{3025}^6 Adverbs syntactic	largely, rapidly, effectively, previously, normally
	v_{3516}^{13} Groups of people semantic	policymakers, geneticists, ancestries, Ohioans

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

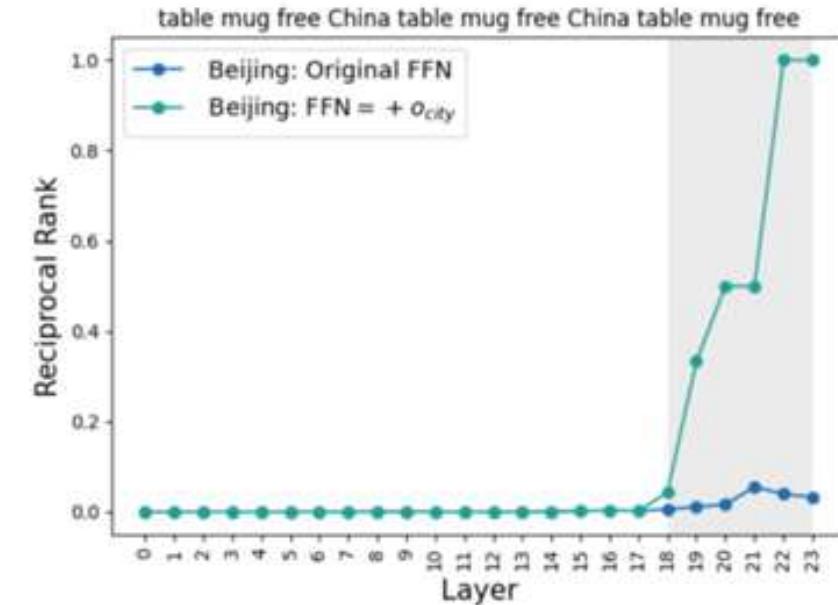


Methods Leveraging Language Model Strengths - Vocabulary Projection

◆ Vocabulary Projection on Transformer Hidden States



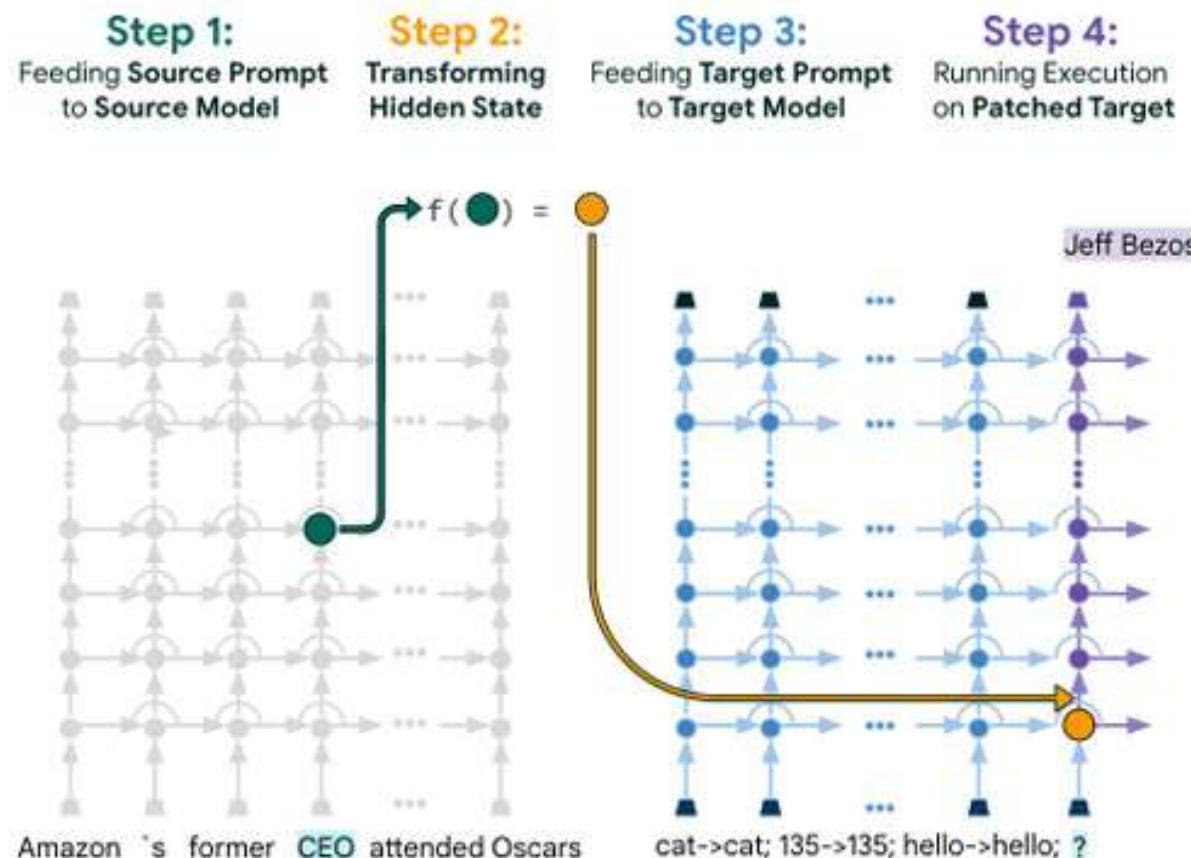
Validated with causal intervention:





Methods Leveraging Language Model Strengths - Vocabulary Projection

◆ Patchscopes





Methods Leveraging Language Model Strengths - Vocabulary Projection

◆ 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

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

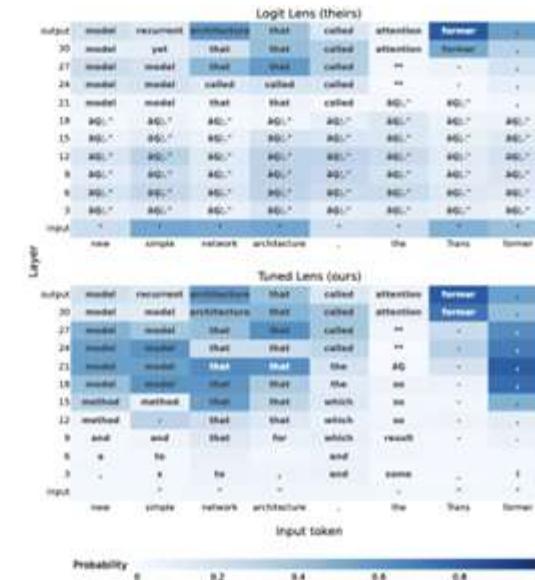


Figure 1. Comparison of our method, the *tuned lens* (bottom), with the “logit lens” (top) for GPT-Neo-2.7B prompted with an except from the abstract of [Vaswani et al. \(2017\)](#). Each cell shows the top-1 token predicted by the model at the given layer and token index. The logit lens fails to elicit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



Methods Leveraging Language Model Strengths - Vocabulary Projection

◆ 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?



Methods Leveraging Language Model Strengths - Decoding Natural Language Explanations from Representations

- ◆ 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



Methods Leveraging Language Model Strengths - Decoding Natural Language Explanations from Representations

- ◆ Using GPT-4 to Explain Neurons of GPT-2

May 9, 2023

Language models can explain neurons in language models

[Read paper ↗](#) [View neurons ↗](#) [View code and dataset ↗](#)



Methods Leveraging Language Model Strengths - Decoding Natural Language Explanations from Representations

- ◆ Using GPT-4 to Explain Neurons of GPT-2

Activations of a neuron in **GPT-2**

a very special event in collaboration with ArenaNet to give away 20 Scarlet Briar t-shirts. Oh, and they're quite lovely. Scarlet Briar began her reign of terror months ago, launching assault after **assault** upon Tyria and its people. Together with the Aetherblade pirates, she unleashed world bosses and catastrophic inv

. Once inside Himkok, you are greeted by an interior that is an even cross between a Prohibition hideout and modern laboratory. Featuring prominently to your eyes upon entry will be jar after **jar** of pickled fruits and vegetables, which is an homage to the days of Prohibition when secret bars would set up elaborate fronts of legitimate

just so. But do you think they call me Roberts the Cathedral Builder? No."He points out the other window. "You see that pier on the lake out there? I built that pier with my bare hands, driving the pilings 10-feet into the sand, laying the pier plank by **plank** but

Health Statistics and based on a sample of 58,488 women and 24,652 men in the United States. To reach his findings, he then ran projections for the Millennial Generation as they age, comparing people who were born between 1940 and 1990 decade-by-decade. "To me the most surprising

Explanation: X by / after X



Methods Leveraging Language Model Strengths - Decoding Natural Language Explanations from Representations

- ◆ Using GPT-4 to Explain Neurons of GPT-2

Propose hypothesis: few-shot prompting

Step 1 Explain the neuron's activations using GPT-4

Show neuron activations to GPT-4:

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 inadvertently blows up an

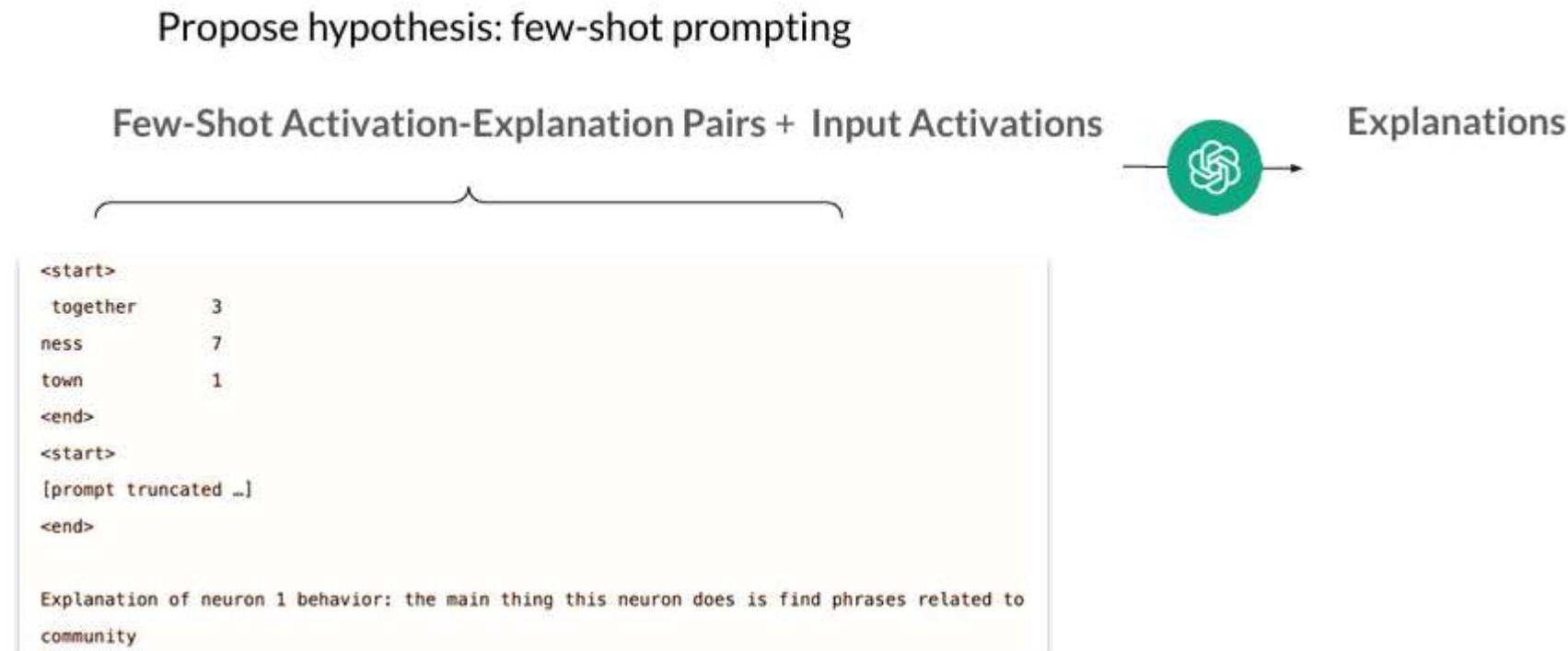
GPT-4 gives an explanation, guessing that the neuron is activating on

references to movies, characters, and entertainment.



Methods Leveraging Language Model Strengths - Decoding Natural Language Explanations from Representations

- ◆ Using GPT-4 to Explain Neurons of GPT-2





Methods Leveraging Language Model Strengths - Decoding Natural Language Explanations from Representations

◆ Using GPT-4 to Explain Neurons of GPT-2

Verify hypothesis:

simulate activations based on explanations; compare simulated and actual activations

Activations Obtained by GPT-4

Assuming that the neuron activates on

references to movies, characters, and entertainment.

GPT-4 guesses how strongly the neuron responds at each token:

: 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

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.

offbeat , Screenshots | Follow This Author @KartikMdgl We have two images from **Skyrim**, which totally stumped us. They show a walking barrel, and we're not sure how exactly that happened. Check out these two images below. Some people really do some weird

ultimate in lightweight portability. Generating chest-thumping lows and crystal clear highs, the four models in the series – the XLS1000, XLS1500, XLS2000, and XLS2500 – are engineered to meet any demanding audio requirements – reliably and within budget. Every XLS

Actual Activations

: 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

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.

offbeat , Screenshots | Follow This Author @KartikMdgl We have two images from **Skyrim**, which totally stumped us. They show a walking barrel, and we're not sure how exactly that happened. Check out these two images below. Some people really do some weird

ultimate in lightweight portability. Generating chest-thumping lows and crystal clear highs, the four models in the series – the XLS1000, XLS1500, XLS2000, and XLS2500 – are engineered to meet any demanding audio requirements – reliably and within budget. Every XLS

compare



Methods Leveraging Language Model Strengths - Decoding Natural Language Explanations from Representations

◆ 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 revamp the website to attract more

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



Methods Leveraging Language Model Strengths - Decoding Natural Language Explanations from Representations

- ◆ Takeaways

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

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



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 be 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?



Explanation methods to address the challenges

	Prompt-based	Influence function	Mech interp
Lack of transparency	Improve communication, faithfulness, etc	Understand data	Many ways
Computation	Language understanding and instruction following	Understand and refine data	Understand model
Unstability	Improve reasoning, grounding and instructability	Improve stability	Facilitate targeted control
Over-reliance	Improve communication, faithfulness, etc	Help understand the supports for model decisions	Help understand mechanisms of model decisions
Hard to evaluate	New evaluation methods and paradigms	New evaluation methods and paradigms	New evaluation methods and paradigms

AIAA 2290: Ethics, Privacy and Security in AI

Thanks!!

Yibo YAN

yyan047@connect.hkust-gz.edu.cn

The Hong Kong University of Science and Technology (Guangzhou)

2025 Spring