

Model Analysis and Interpretability



Jing Huang

Slides adapted from CS 224U

Lecture Plan

Motivations

Advanced behavioral analysis methods

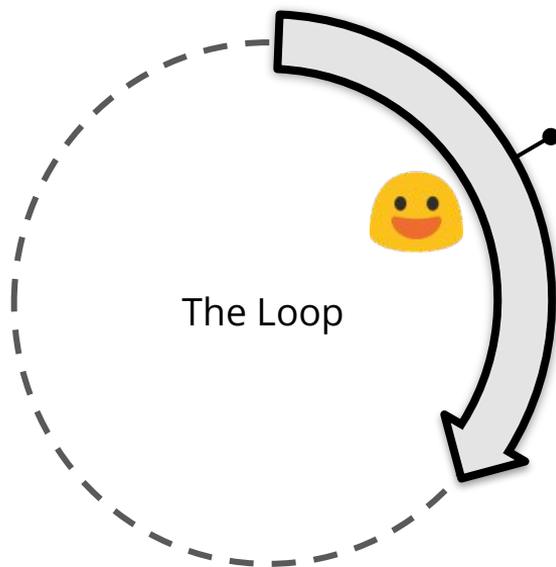
- Stress testing

Structural analysis methods

- Probing
- Attribution
- Causal intervention

Open problems in interpretability

Motivations: The LLM development cycle



Phase 1: Create a benchmark for a task that none of the existing models can solve.

WMT

GLUE SuperGLUE

SQuAD TriviaQA

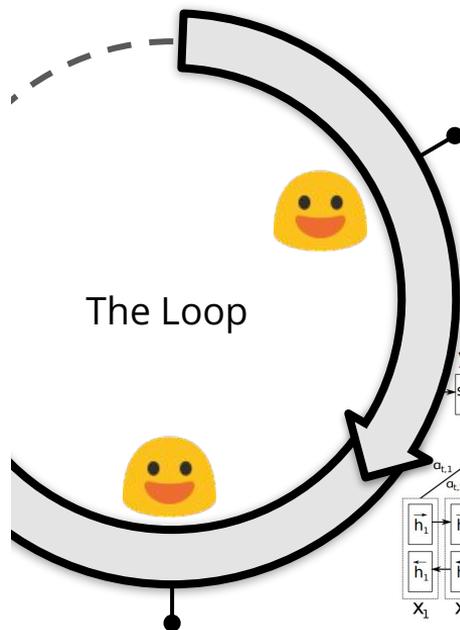
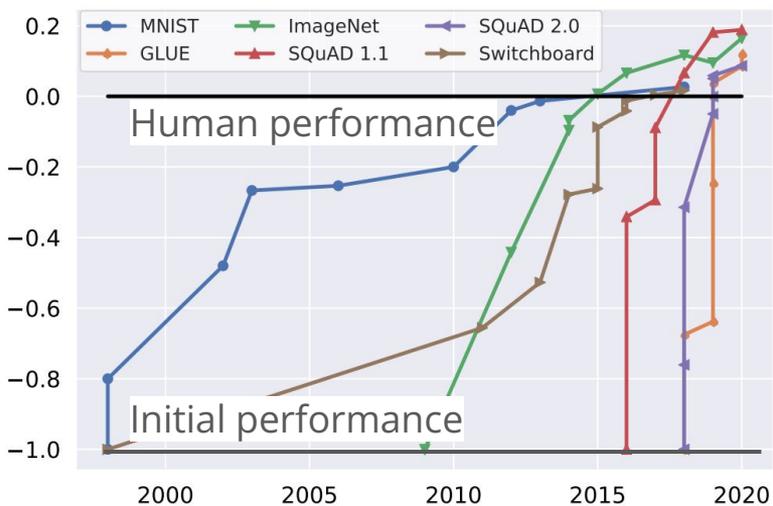
cais/mmlu

openai/gsm8k

Chatbot Arena

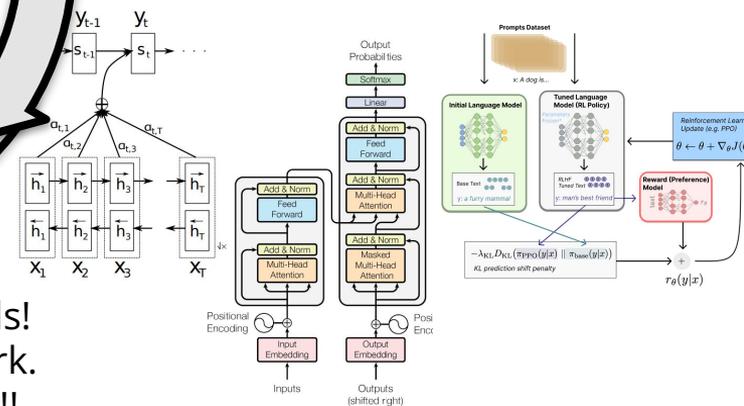
SWE-bench

Motivations: The LLM development cycle



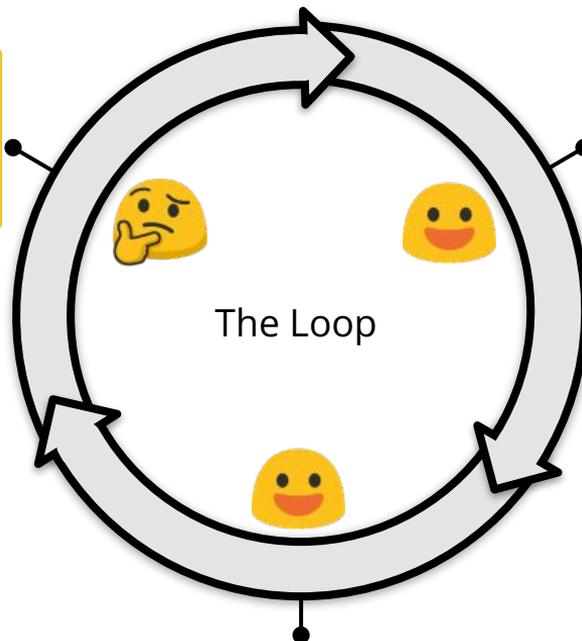
Phase 1: Create a benchmark for a task that none of the existing models can solve.

Phase 2: Build new models!
Evaluate on the benchmark.
The benchmark is solved!!



Motivations: The LLM development cycle

Phase 3: Discover that models can't solve simple variations of the benchmark.



Phase 1: Create a benchmark for a task that none of the existing models can solve.

Phase 2: Build new models!
Evaluate on the benchmark.
The benchmark is solved!!

Part II: Advanced Behavioral Analysis Methods

Method: Stress testing, i.e., creating a challenging set

- Assessing the compositionality of the model
 - Primitive **substitutions**
 - Primitive **permutations**
- Evaluating the robustness to noises
 - Adding semantics-preserving **noises**

Stress testing with primitive substitutions

GSM8K

When Sophie watches her nephew, she gets out a variety of toys for him. The bag of building blocks has 31 blocks in it. The bin of stuffed animals has 8 stuffed animals inside. The tower of stacking rings has 9 multicolored rings on it. Sophie recently bought a tube of bouncy balls, bringing her total number of toys for her nephew up to 62. How many bouncy balls came in the tube?

GSM Symbolic Template

When {name} watches her {family}, she gets out a variety of toys for him. The bag of building blocks has {x} blocks in it. The bin of stuffed animals has {y} stuffed animals inside. The tower of stacking rings has {z} multicolored rings on it. {name} recently bought a tube of bouncy balls, bringing her total number of toys she bought for her {family} up to {total}. How many bouncy balls came in the tube?

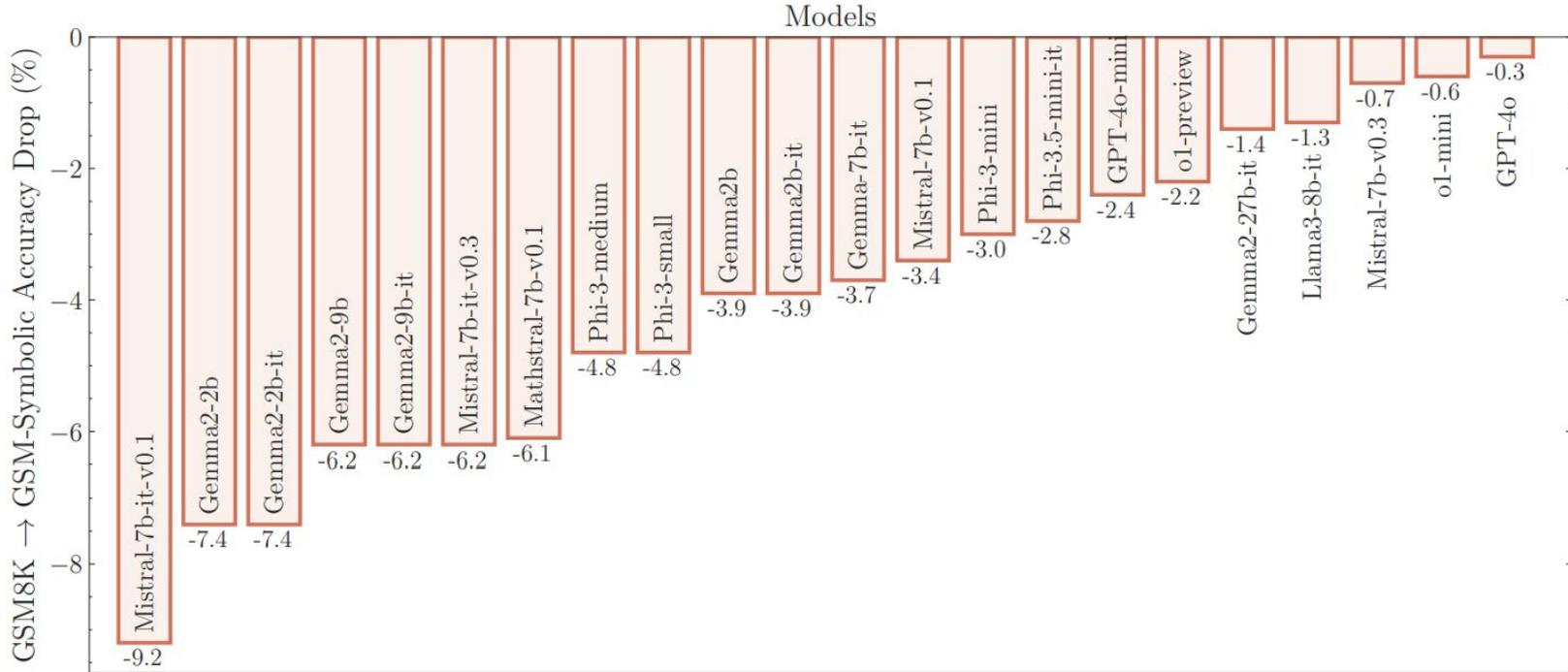
#variables:

- name = sample(names)
- family = sample(["nephew", "cousin", "brother"])
- x = range(5, 100)
- y = range(5, 100)
- z = range(5, 100)
- total = range(100, 500)
- ans = range(85, 200)

#conditions:

- x + y + z + ans == total

Models are not robust to primitive substitutions



Stress testing compositionality in VLMs

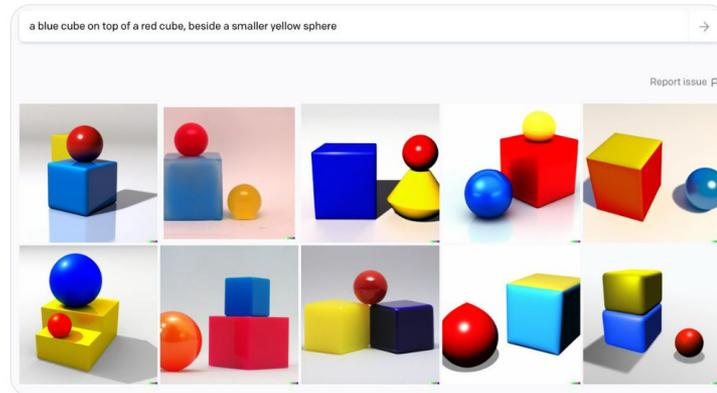


(a) some plants surrounding a lightbulb



(b) a lightbulb surrounding some plants

Prompt: a blue cube on top of a red cube, beside a smaller yellow sphere



Stress testing with adding noises

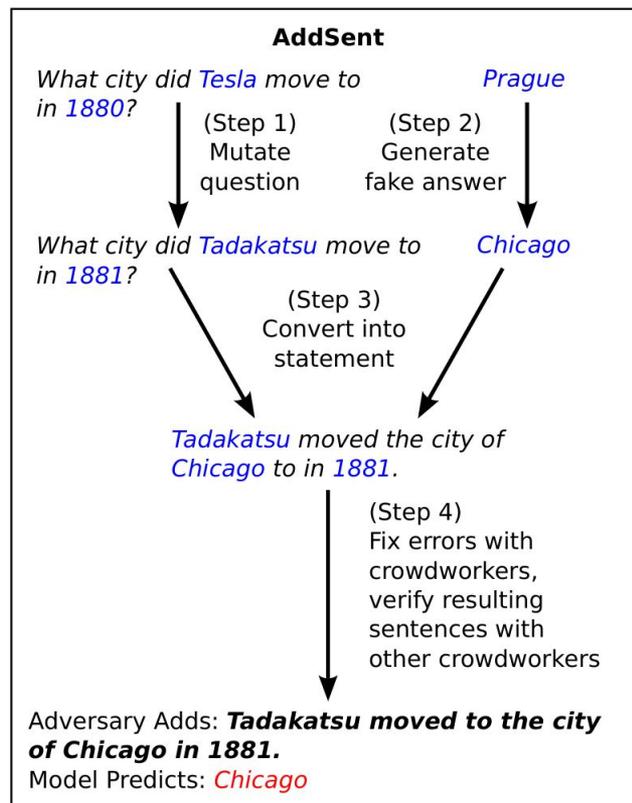
Article: Nikola Tesla

Paragraph: "In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study.

Unfortunately, he arrived too late to enroll at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses."

Question: "What city did Tesla move to in 1880?"

Prediction: Prague



Models are not robust to insertions of irrelevant content

Article: Nikola Tesla

Paragraph: "In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study.

Unfortunately, he arrived too late to enroll at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses."

Question: "What city did Tesla move to in 1880?"

Prediction: Prague

Article: Nikola Tesla

Paragraph: "In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study.

Unfortunately, he arrived too late to enroll at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses. Tadakatsu moved to the city of Chicago in 1881."

Question: "What city did Tesla move to in 1880?"

Prediction: Chicago

Stress testing with semantics-preserving noises

GSM-NoOp

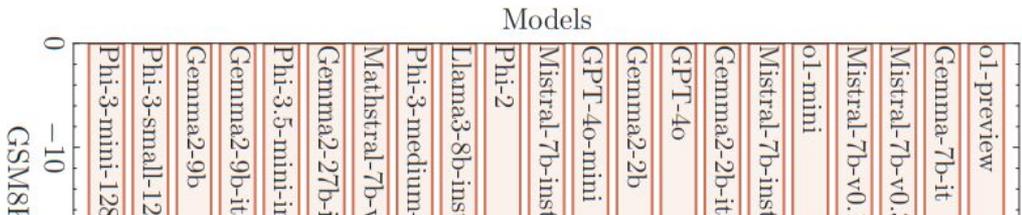
Oliver picks 44 kiwis on Friday. Then he picks 58 kiwis on Saturday. On Sunday, he picks double the number of kiwis he did on Friday, but five of them were a bit smaller than average. How many kiwis does Oliver have?

o1-mini: Oliver picks kiwis over three days as follows:

- Friday: 44 kiwis
- Saturday: 58 kiwis
- Sunday: Double the number he picked on Friday, which is $2 \times 44 = 88$ kiwis

However, on Sunday, 5 of these kiwis were smaller than average. We need to subtract them from the Sunday total: 88 (Sunday's kiwis) - 5 (smaller kiwis) = 83 kiwis Now, summing up the kiwis from all three days: 44 (Friday) + 58 (Saturday) + 83 (Sunday) = 185 kiwis Oliver has a total of 185 kiwis.

Models are not robust to semantics-preserving noises



We won't trust our models, unless the solution they have found is systematic.



Motivations: The LLM development cycle

Phase 3: Stress testing?

SYNTHETIC AND NATURAL NOISE BOTH BREAK NEURAL MACHINE TRANSLATION

WMT

Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference

lington.edu

Stress Test Evaluation for Natural Language Inference

GLUE SuperGLUE

Adversarial Examples for Evaluating Reading Comprehension Systems

SQuAD

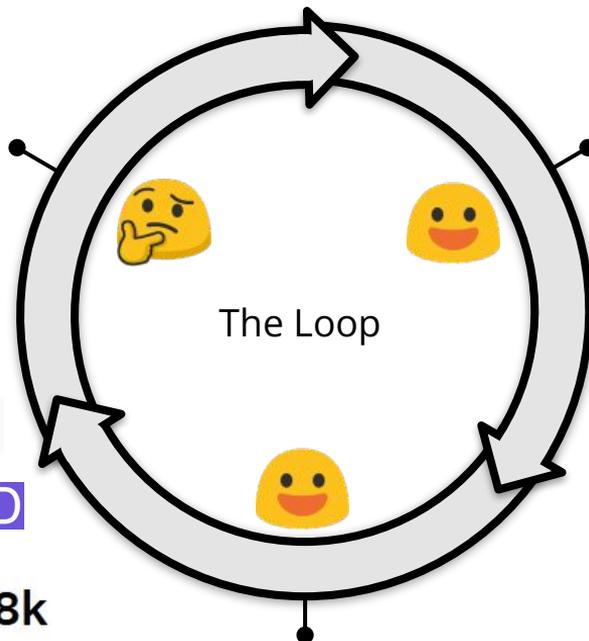
GSM-SYMBOLIC: UNDERSTANDING THE LIMITATIONS OF MATHEMATICAL REASONING IN LARGE LANGUAGE MODELS

openai/gsm8k

LARGE LANGUAGE MODELS ARE NOT ROBUST MULTIPLE CHOICE SELECTORS

cais/mmlu

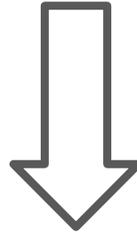
MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark



Phase 1: Create a benchmark for a task that none of the existing models can solve.

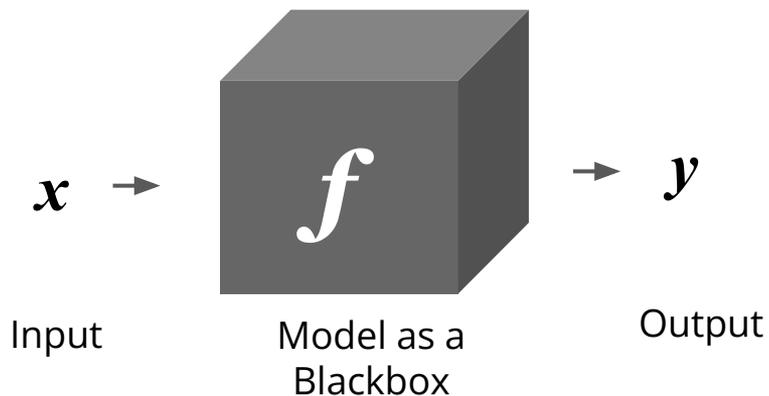
Phase 2: Build new models!
Evaluate on the benchmark.
The benchmark is solved!!

Does the system make a correct prediction?



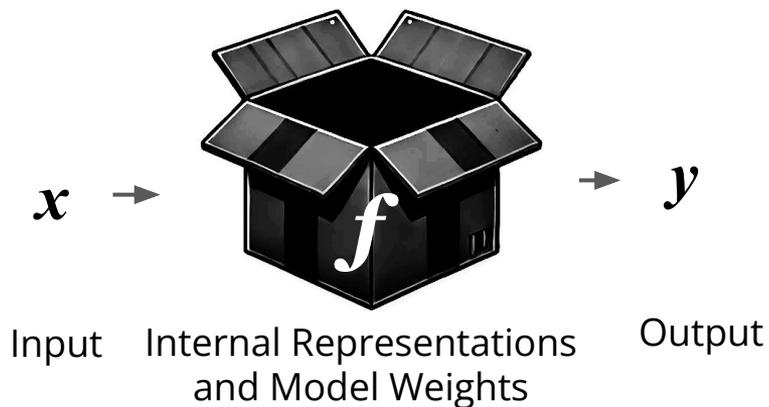
Why does the system make this prediction?

Going beyond behavioral testing



Limitation of behavioral testing: Cannot offer insights on model behaviors beyond the distribution tested.

Going beyond behavioral testing



A core thesis of interpretability: A model will succeed at a generalization task if and only if it has induced a mechanism that implements a “correct” algorithm for that task

Our field starts with a different model development cycle

Letter | Published: 09 October 1986

Learning representations by back-propagating errors

[David E. Rumelhart](#), [Geoffrey E. Hinton](#) & [Ronald J. Williams](#)

The ideal model development cycle

Task

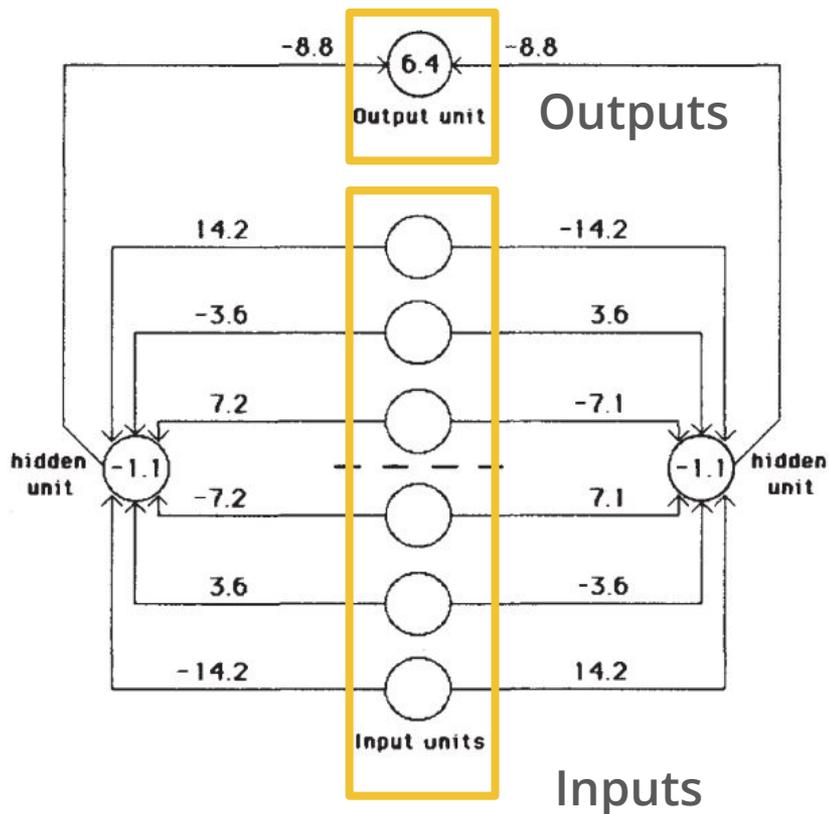
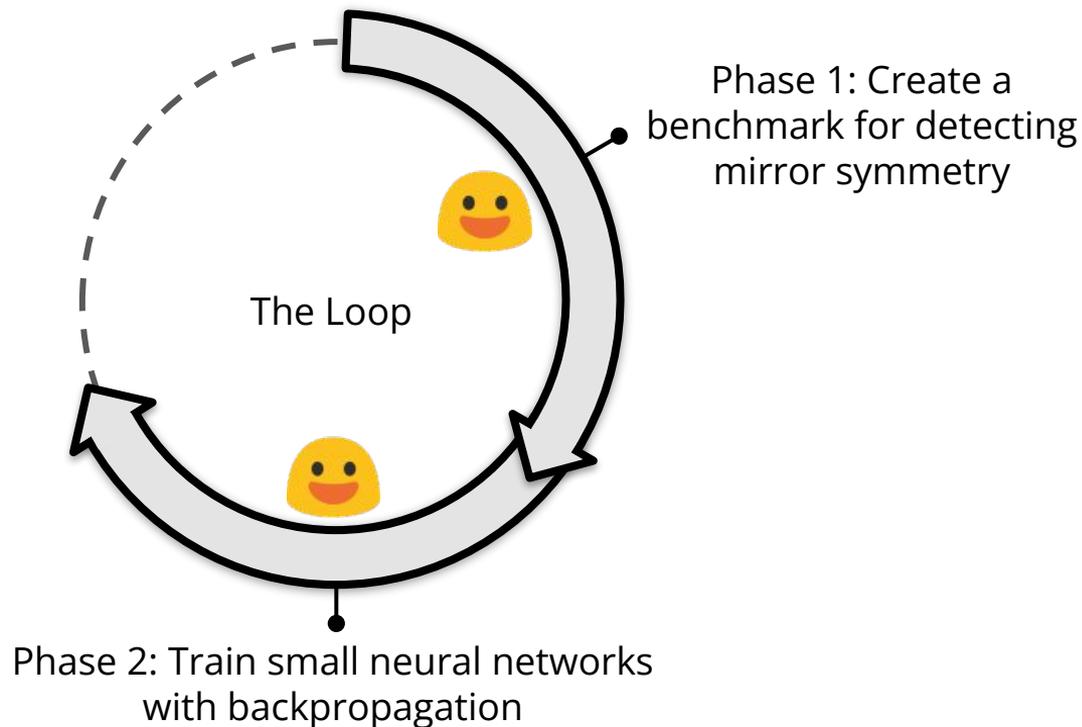


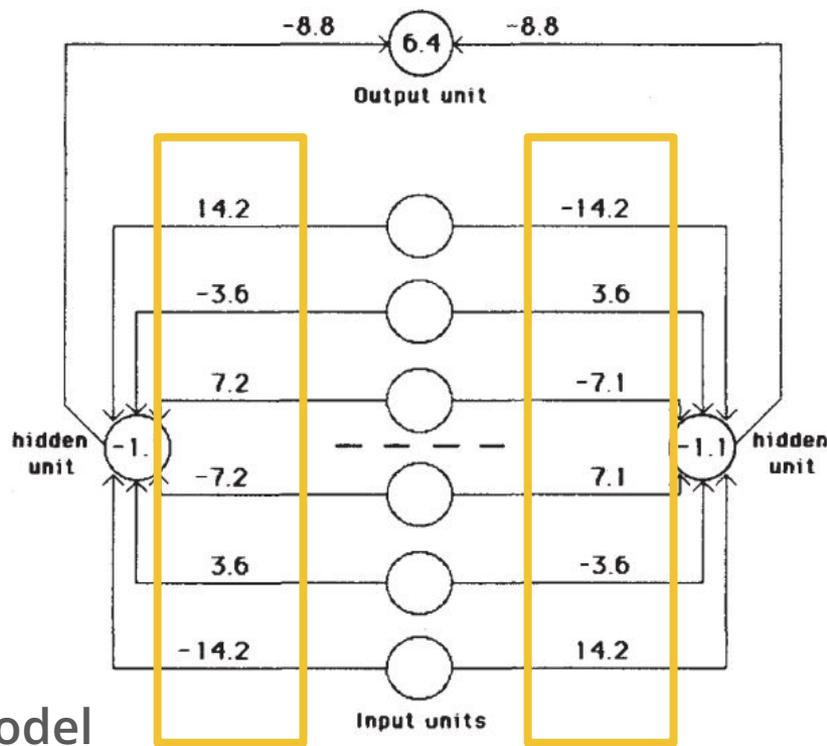
Fig. 1 A network that has learned to detect mirror symmetry in the input vector. The numbers on the arcs are weights and the numbers inside the nodes are biases. The learning required 1,425 sweeps through the set of 64 possible input vectors, with the weights being adjusted on the basis of the accumulated gradient after each sweep. The values of the parameters in equation (9) were $\epsilon = 0.1$ and $\alpha = 0.9$. The initial weights were random and were uniformly distributed between -0.3 and 0.3 . The key property of this solution

is that for a given hidden unit, weights the middle of the input vector are equal in sign. So if a symmetrical pattern is presented, the hidden units will receive a net input of 0 from the input units, and, because the hidden units have a negative bias, both will be off. In this case the output unit, having a positive bias, will be on. Note that the weights on each side of the midpoint are in the ratio 1:2:4. This ensures that each of the eight patterns that can occur above the midpoint sends a unique activation sum to each hidden unit, so the only pattern below the midpoint that can exactly balance this sum is the symmetrical one. For all non-symmetrical patterns, both hidden units will receive non-zero activations from the input units. The two hidden units have identical patterns of weights but with opposite signs, so for every non-symmetric pattern one hidden unit will come on and suppress the output unit.

The ideal model development cycle



The ideal model development cycle



Model
Weights

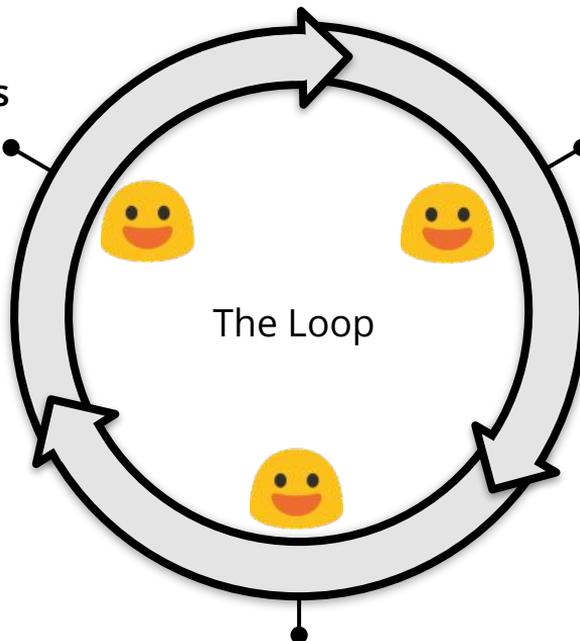
Fig. 1 A network that has learned to detect mirror symmetry in the input vector. The numbers on the arcs are weights and the numbers inside the nodes are biases. The learning required 1,425 sweeps through the set of 64 possible input vectors, with the weights being adjusted on the basis of the accumulated gradient after each sweep. The values of the parameters in equation (9) were $\epsilon = 0.1$ and $\alpha = 0.9$. The initial weights were random and were uniformly distributed between -0.3 and 0.3 . The key property of this solution is that for a given hidden unit, weights that are symmetric about the middle of the input vector are equal in magnitude and opposite in sign. So if a symmetrical pattern is presented, both hidden units will receive a net input of 0 from the input units, and, because the hidden units have a negative bias, both will be off. In this case the output unit, having a positive bias, will be on. Note that the weights on each side of the midpoint are in the ratio 1:2:4. This ensures that each of the eight patterns that can occur above the midpoint

Dissect the model and recover the underlying algorithms learned by the model.

will come on and suppress the output unit.

The ideal model development cycle

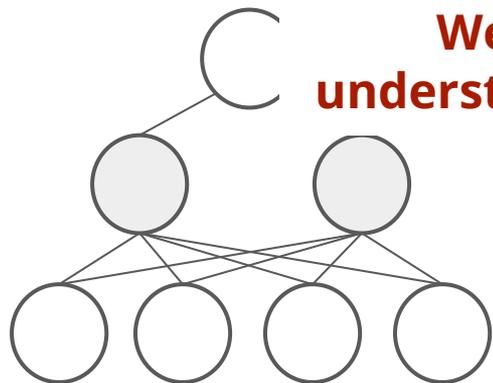
Phase 3: Analyze whether models learn a systematic solution that can generalize well.



Phase 1: Create a benchmark for a task that none of the existing models can solve.

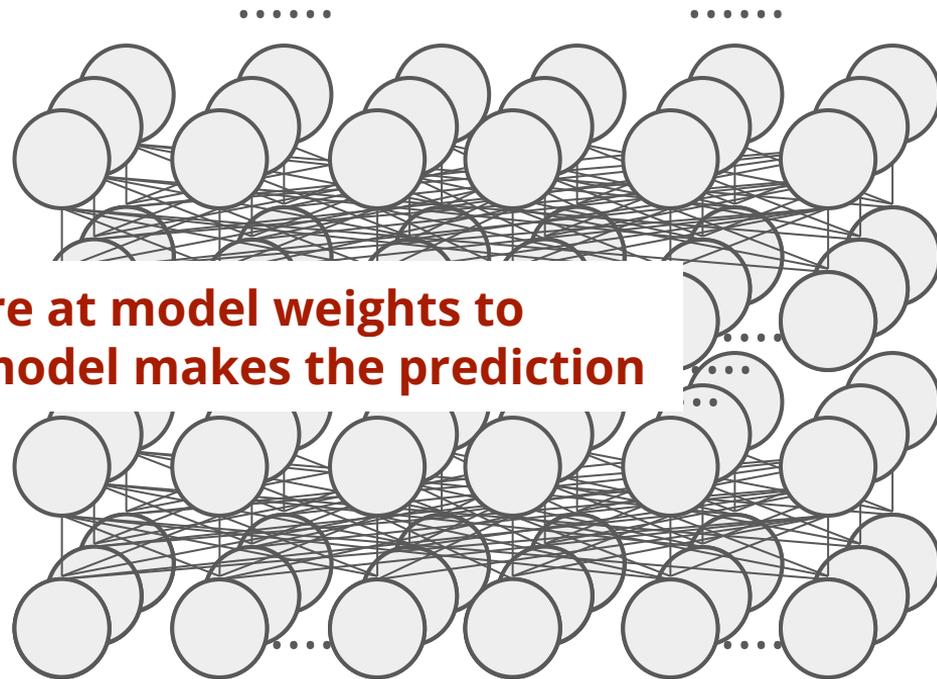
Phase 2: Build new models!
Evaluate on the benchmark.
The benchmark is solved!!

Today's LLMs have billions of neurons!



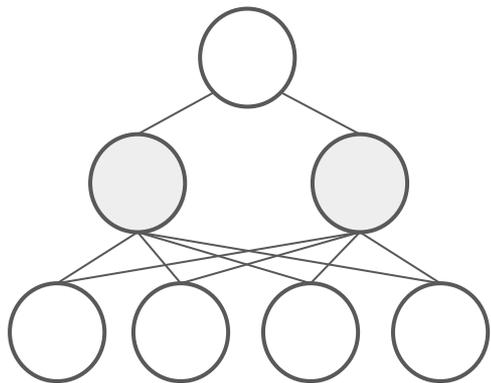
1986: 2 neurons

We can't just stare at model weights to understand how the model makes the prediction

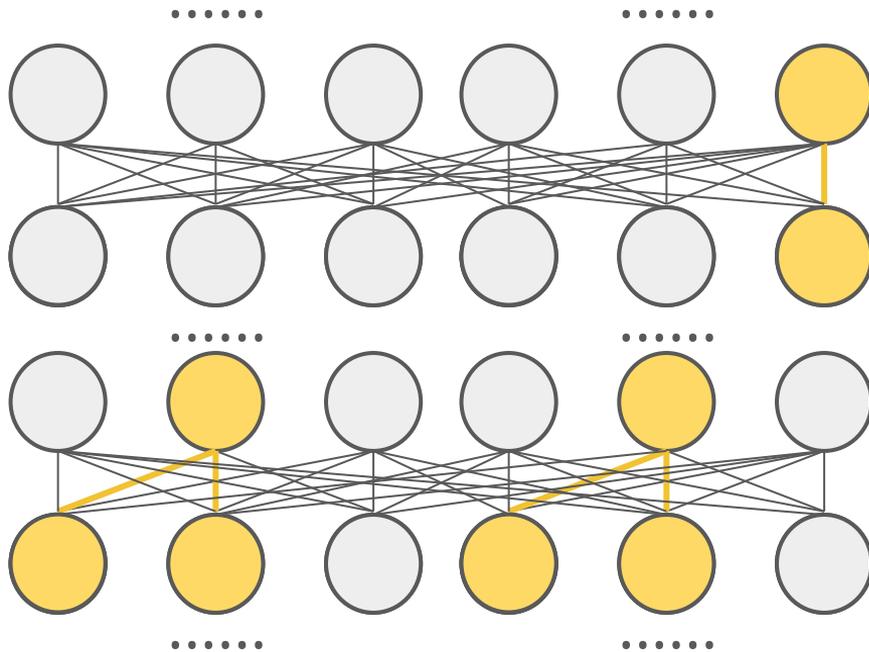


The LLM era: billions of neurons

Can we still find sparse and interpretable structures?



1986: 2 neurons



The LLM era: billions of neurons

Part III: Structural Analysis Methods

- Probing
- Feature attribution
- Causal intervention

Probing

Goal: Understand what concepts are encoded in the hidden representations

Method: Train supervised classifiers on internal representations and evaluate their accuracy

- **Intuition:** If a concept is encoded in the representations, the classifier should have higher accuracy
- **Full recipe**
 - Hypothesize a concept or a structure that is relevant to model behaviors
 - Construct a dataset of input sequences to concept-related labels
 - Train supervised classifiers on features extracted from certain model location
 - Evaluate classifier accuracy on a holdout set

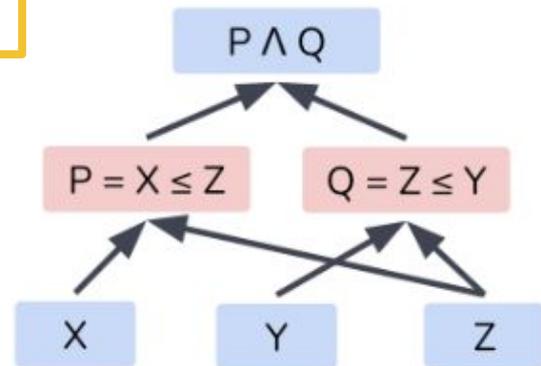
Hypothesize a concept or a structure

Task: Predicting whether a number is within a range

PROMPT: Does the following item cost between \$10.2 and \$18.7? Item: \$5.87.

OUTPUT: No, the item does not cost between \$10.2 and \$18.7.

Does the model actually compute the **difference** between the left/right boundary and item?



Construct a dataset

Input (Train Split)

Label

Does the following item cost between \$10.2 and \$18.7? Item: \$5.8.

0

Does the following item cost between \$3.7 and \$6.7? Item: \$3.9.

1

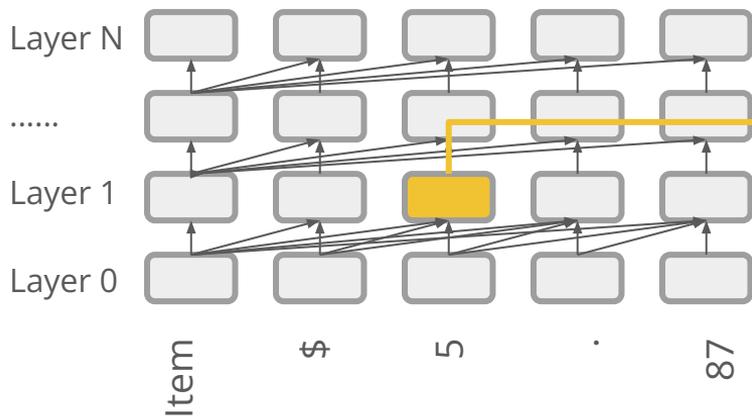
Does the following item cost between \$9.5 and \$12.8? Item: \$15.5.

1

.....

Train supervised classifiers

1. Select a feature location



2. Define a probe

$$\phi(x) = \sigma(Wx + b)$$

3. Train with cross-entropy loss

$$\ell_{W_\phi} = \mathbb{E}[-y \cdot \log(\phi(x))]$$

Label: X < Z

Evaluate probe accuracy

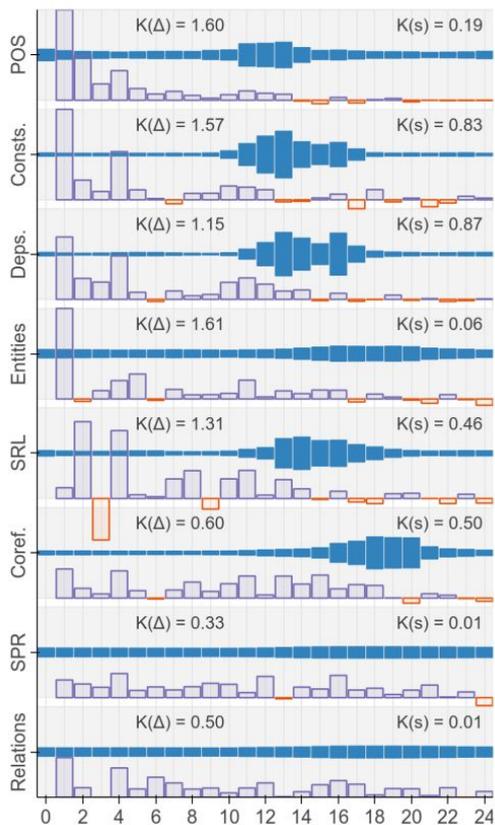
| Layer | gin_of_text > | _header_id > | _user_header_id > | _header_id > | \n\n' | Does | the | following | item | cost | between | \$ | 10 | . | 2 | and | \$ | 18 | ? | Item | . | \$ | 5 | . | 87 | . | </eot_id > | _header_id > | _assistant_header_id > | _header_id > | \n\n' | | | | |
|-------|---------------|--------------|-------------------|--------------|-------|------|-----|-----------|------|------|---------|-----|-----|-----|-----|-----|-----|-----|-----|------|-----|-----|-----|-----|-----|-----|------------|--------------|------------------------|--------------|-------|-----|-----|-----|-----|
| 31 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.6 | 0.7 | 0.6 | 0.6 | 0.6 | 0.7 | 0.7 | 0.6 | 0.6 | 0.7 | 1 | 1 | 0.9 | 0.9 | 0.9 | 1 | 0.9 | 0.9 | 0.9 | |
| 30 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.6 | 0.7 | 0.6 | 0.6 | 0.6 | 0.7 | 0.7 | 0.6 | 0.6 | 0.7 | 1 | 1 | 0.9 | 0.9 | 1 | 1 | 0.9 | 0.9 | 0.9 | |
| 29 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.6 | 0.7 | 0.6 | 0.6 | 0.6 | 0.6 | 0.7 | 0.6 | 0.6 | 0.7 | 1 | 1 | 0.9 | 0.9 | 0.9 | 1 | 0.9 | 0.9 | 0.9 | |
| 28 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.6 | 0.7 | 0.6 | 0.6 | 0.6 | 0.6 | 0.7 | 0.6 | 0.6 | 0.7 | 1 | 0.9 | 0.9 | 0.9 | 0.9 | 1 | 0.9 | 0.9 | 0.9 | |
| 27 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.6 | 0.7 | 0.6 | 0.6 | 0.6 | 0.6 | 0.7 | 0.6 | 0.6 | 0.7 | 1 | 0.9 | 0.9 | 0.9 | 0.9 | 1 | 0.9 | 0.9 | 0.9 | |
| 26 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.6 | 0.7 | 0.6 | 0.6 | 0.6 | 0.6 | 0.7 | 0.6 | 0.6 | 0.7 | 1 | 1 | 0.9 | 0.9 | 0.9 | 1 | 0.9 | 0.9 | 0.9 | |
| 25 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.6 | 0.7 | 0.6 | 0.7 | 0.7 | 0.6 | 0.7 | 0.6 | 0.6 | 0.7 | 1 | 0.9 | 0.9 | 0.9 | 0.9 | 1 | 0.9 | 0.9 | 0.9 | |
| 24 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.6 | 0.7 | 0.6 | 0.7 | 0.7 | 0.6 | 0.7 | 0.6 | 0.6 | 0.7 | 1 | 0.9 | 0.9 | 0.9 | 0.9 | 1 | 0.9 | 0.9 | 0.9 | |
| 23 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.6 | 0.7 | 0.6 | 0.7 | 0.7 | 0.7 | 0.7 | 0.6 | 0.6 | 0.7 | 1 | 0.9 | 0.9 | 0.9 | 0.9 | 1 | 0.9 | 0.9 | 0.9 | |
| 22 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.6 | 0.7 | 0.6 | 0.7 | 0.7 | 0.7 | 0.6 | 0.6 | 0.6 | 0.7 | 1 | 0.9 | 0.9 | 0.9 | 0.9 | 1 | 0.9 | 0.9 | 0.9 | |
| 21 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.6 | 0.7 | 0.6 | 0.7 | 0.6 | 0.6 | 0.7 | 0.6 | 0.6 | 0.6 | 1 | 0.9 | 0.9 | 0.9 | 0.9 | 1 | 0.9 | 0.9 | 0.9 | |
| 20 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.6 | 0.7 | 0.6 | 0.6 | 0.7 | 0.7 | 0.7 | 0.6 | 0.6 | 0.6 | 1 | 0.9 | 0.9 | 0.9 | 0.9 | 1 | 0.9 | 0.9 | 0.9 | |
| 19 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.6 | 0.7 | 0.6 | 0.6 | 0.7 | 0.7 | 0.7 | 0.6 | 0.6 | 0.6 | 1 | 1 | 0.9 | 0.9 | 0.9 | 1 | 0.9 | 0.9 | 0.9 | |
| 18 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.6 | 0.7 | 0.6 | 0.6 | 0.6 | 0.7 | 0.7 | 0.6 | 0.6 | 0.6 | 1 | 0.9 | 1 | 0.9 | 0.9 | 1 | 0.9 | 0.9 | 0.9 | |
| 17 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.6 | 0.7 | 0.6 | 0.6 | 0.6 | 0.7 | 0.7 | 0.7 | 0.6 | 0.6 | 1 | 0.9 | 0.9 | 0.9 | 0.9 | 1 | 0.9 | 0.9 | 0.9 | |
| 16 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.6 | 0.6 | 0.6 | 0.6 | 0.6 | 0.7 | 0.7 | 0.7 | 0.6 | 0.6 | 1 | 0.9 | 0.9 | 0.9 | 0.9 | 1 | 0.9 | 0.9 | 0.9 | |
| 15 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.6 | 0.6 | 0.6 | 0.6 | 0.7 | 0.7 | 0.7 | 0.7 | 0.6 | 0.6 | 1 | 0.9 | 0.9 | 0.9 | 0.9 | 1 | 0.9 | 0.9 | 0.9 | |
| 14 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.6 | 0.6 | 0.6 | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 | 0.6 | 0.6 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 1 | 0.9 | 0.9 | 0.9 | |
| 13 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.6 | 0.6 | 0.6 | 0.7 | 0.7 | 0.7 | 0.7 | 0.6 | 0.6 | 0.6 | 1 | 0.9 | 0.9 | 1 | 0.9 | 1 | 0.9 | 0.9 | 0.9 | |
| 12 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.6 | 0.6 | 0.6 | 0.7 | 0.7 | 0.7 | 0.7 | 0.6 | 0.6 | 0.6 | 1 | 0.9 | 0.9 | 0.9 | 0.9 | 1 | 0.8 | 0.9 | 0.9 | |
| 11 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.6 | 0.6 | 0.6 | 0.7 | 0.7 | 0.7 | 0.7 | 0.6 | 0.6 | 0.6 | 1 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 0.8 | 0.9 | 0.9 | |
| 10 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.6 | 0.6 | 0.6 | 0.7 | 0.7 | 0.7 | 0.7 | 0.6 | 0.6 | 0.6 | 1 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 0.7 | 0.7 | 0.8 | |
| 9 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.6 | 0.7 | 0.6 | 0.7 | 0.7 | 0.7 | 0.6 | 0.6 | 0.6 | 0.6 | 1 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 0.7 | 0.7 | 0.8 | |
| 8 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.6 | 0.7 | 0.6 | 0.7 | 0.7 | 0.7 | 0.6 | 0.6 | 0.6 | 0.6 | 1 | 1 | 0.9 | 0.9 | 0.8 | 0.7 | 0.5 | 0.5 | 0.5 | |
| 7 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.6 | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 | 0.6 | 0.6 | 0.5 | 0.6 | 0.9 | 0.9 | 0.8 | 0.8 | 0.8 | 0.8 | 0.6 | 0.5 | 0.5 | 0.5 |
| 6 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.6 | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 | 0.6 | 0.6 | 0.5 | 0.6 | 0.9 | 0.8 | 0.8 | 0.8 | 0.8 | 0.6 | 0.5 | 0.5 | 0.5 | |
| 5 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.6 | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 | 0.6 | 0.6 | 0.5 | 0.6 | 0.8 | 0.8 | 0.8 | 0.8 | 0.7 | 0.5 | 0.5 | 0.5 | 0.5 | |
| 4 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.6 | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 | 0.6 | 0.6 | 0.5 | 0.6 | 0.8 | 0.8 | 0.8 | 0.8 | 0.7 | 0.5 | 0.5 | 0.5 | 0.5 | |
| 3 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.7 | 0.6 | 0.7 | 0.7 | 0.7 | 0.7 | 0.6 | 0.6 | 0.5 | 0.5 | 0.8 | 0.8 | 0.8 | 0.8 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | |
| 2 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.6 | 0.6 | 0.6 | 0.7 | 0.6 | 0.6 | 0.5 | 0.5 | 0.5 | 0.5 | 0.8 | 0.8 | 0.8 | 0.8 | 0.6 | 0.5 | 0.5 | 0.5 | 0.5 | |
| 1 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.6 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.8 | 0.5 | 0.7 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | |
| 0 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.7 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.8 | 0.5 | 0.6 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | |

Probing accuracy >0.95 on a holdout test set

Highly likely that model encodes the difference between the left boundary and the item



Probing: BERT Rediscovered the Classical NLP Pipeline



*“POS tags processed earliest, followed by constituents, dependencies, semantic roles, and coreference. That is, it appears that **basic syntactic information appears earlier in the network, while high-level semantic information appears at higher layers.**”*

What could go wrong with probing?

Probes might be too powerful

- Is the information learned by the probe or stored in the original model?

Probes do not distinguish between concepts encoded vs. used

Feature attribution

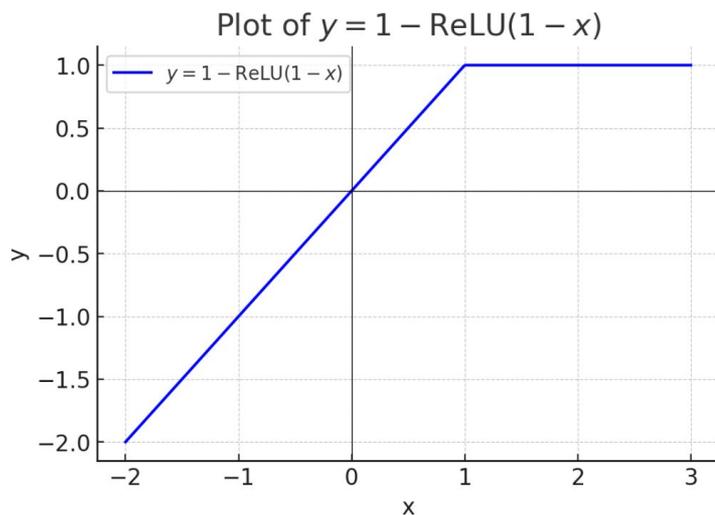
Goal: Understand which input features contribute to the model predictions

Method: Integrated gradient (and many others!)

- **Intuition:** A larger gradient at the feature value means the feature value has large impact on the output
- **Axioms of feature attribution**
 - Sensitivity: If the representations of two inputs x and x' only differ at dimension i , and the two inputs lead to different predictions, then the feature at dimension i has non-zero attribution
 - Implementation invariance: If two models have identical input-output behaviors, they should have the same attribution

Feature attribution

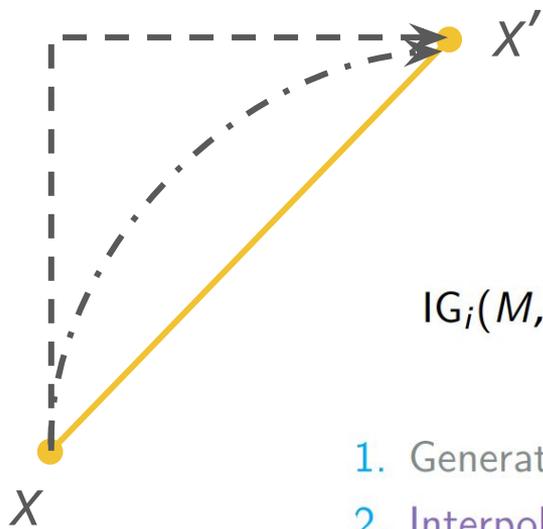
$$\text{InputXGradient}_i(M, x) = \frac{\partial M(x)}{\partial x_i} \cdot x_i$$



Consider a baseline with $x=0$,
what happens when we set $x = 2$?

**Gradient being zero does
not necessarily mean the
attribution should be zero!**

Feature attribution



$$IG_i(M, x, x') = \underbrace{(x_i - x'_i)}_5 \cdot \underbrace{\sum_{k=1}^m}_4 \frac{\underbrace{\partial M(x' + \frac{k}{m} \cdot (x - x'))}_3}{\underbrace{\partial x_i}_2} \cdot \underbrace{\frac{1}{m}}_4$$

1. Generate $\alpha = [1, \dots, m]$
2. Interpolate inputs between baseline x' and actual input x
3. Compute gradients for each interpolated input
4. Integral approximation through averaging
5. Scaling to remain in the space region as the original

What's new in the past 2 years?

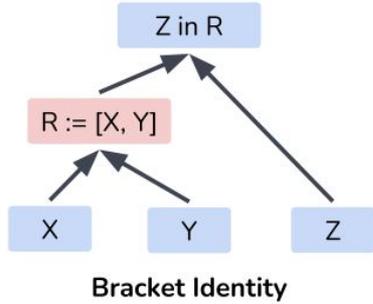
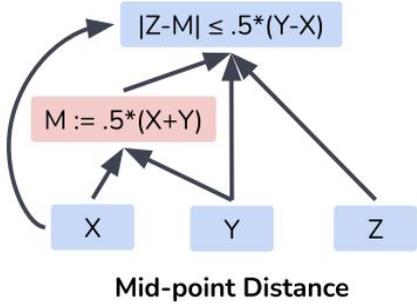
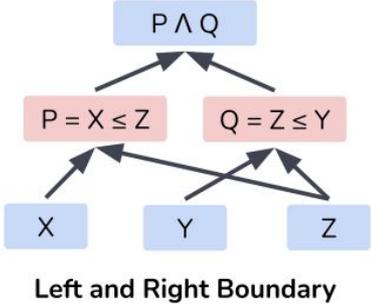
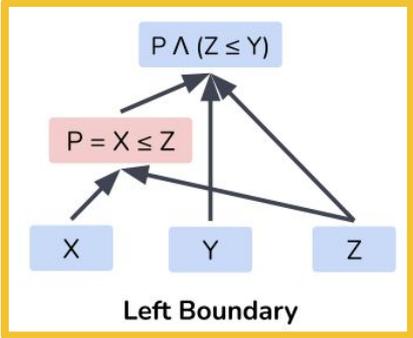
Causal intervention

Goal: Understand which concepts are used in model predictions.

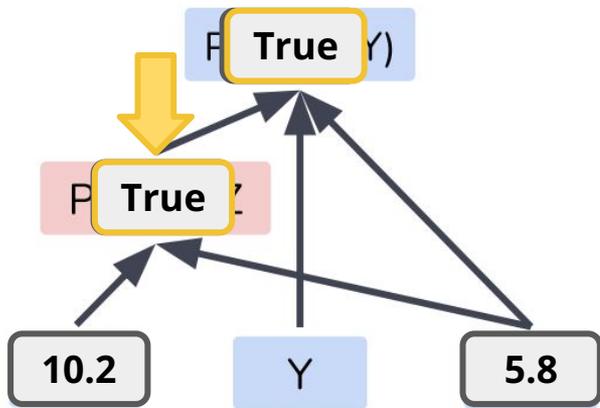
Methods: Causal abstractions

- **Intuition:** If a model component has causal effects on the prediction, intervening on the component should lead to changes in the prediction.
- **Full recipe**
 - Hypothesize a high-level causal model
 - Generate counterfactual data based on the high-level model
 - Search for an alignment between the high-level causal model and the neural model using interchange interventions
 - Evaluation alignment with interchange intervention accuracy

Hypothesize a high-level causal model



Generate counterfactual data



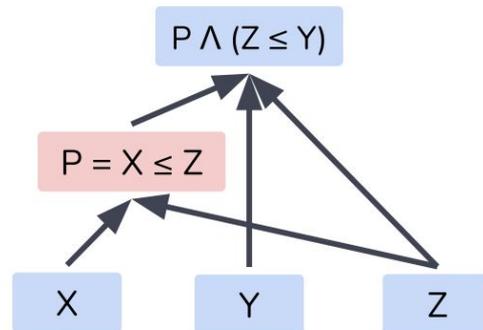
Does the following item cost between \$10.2 and \$18.7? Item: \$5.8.

How do we know these causal relationships hold?

A counterfactual state: what happens if P is set to True

An intervention: An operation that sets the value of a variable

Generate counterfactual data



Input (Train Split)

Base: Does the following item cost between \$10.2 and \$18.7? Item: \$5.8.

Source: Does the following item cost between \$1.5 and \$9.5? Item: \$3.5.

Base: Does the following item cost between \$3.7 and \$6.7? Item: \$12.9.

Source: Does the following item cost between \$9.5 and \$12.8? Item: \$15.5.

Base: Does the following item cost between \$3.5 and \$8.5? Item: \$6.2.

Source: Does the following item cost between \$10.5 and \$15.1? Item: \$1.5.

.....

Counterfactual

Label

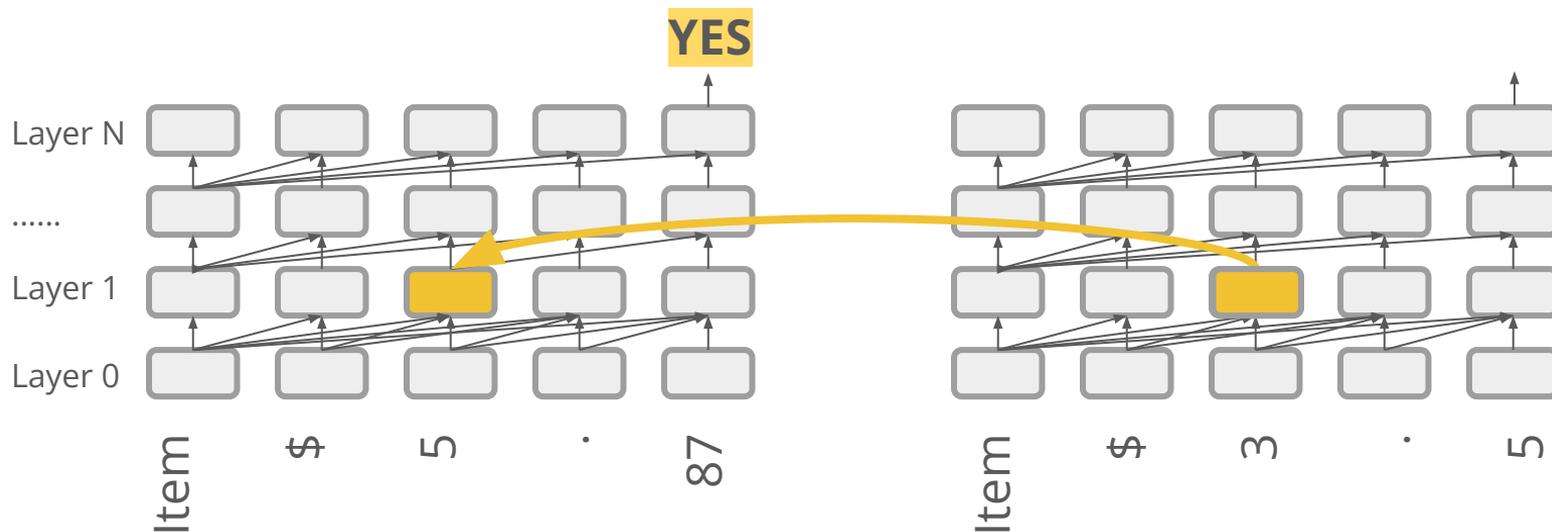
Yes

No

No

Search for an alignment

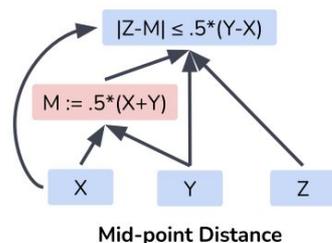
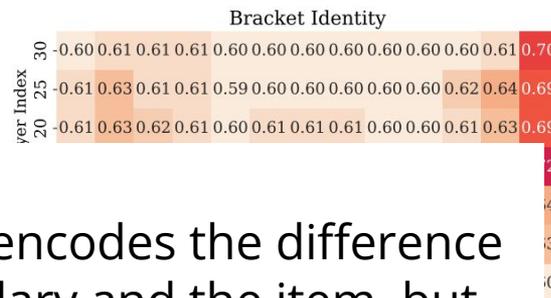
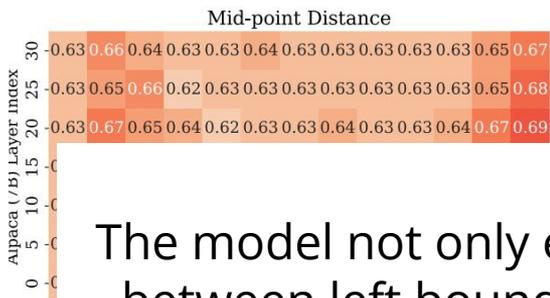
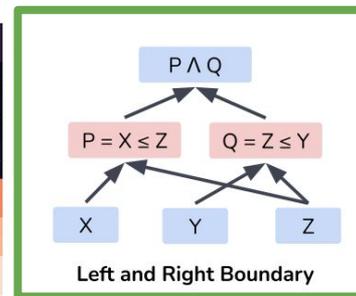
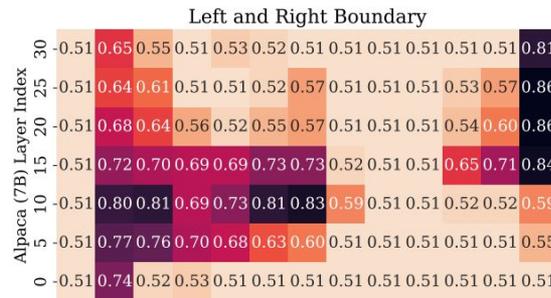
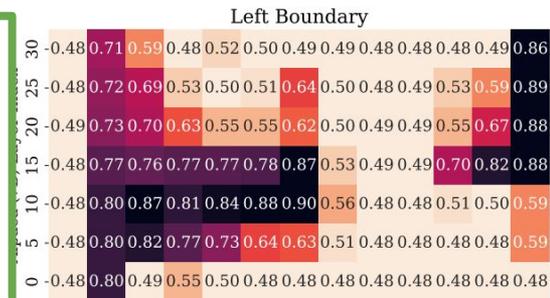
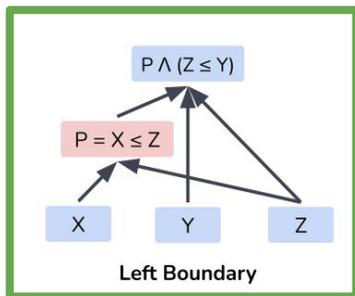
The key operation: Interchange Intervention



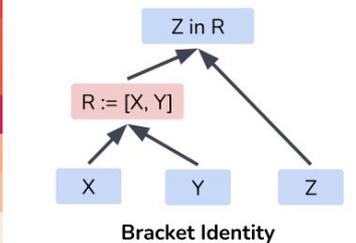
Does the following item cost between \$10.2 and \$18.7? Item: \$5.87.

Does the following item cost between \$1.5 and \$9.5? Item: \$3.5.

Evaluate interchange intervention accuracy



The model not only encodes the difference between left boundary and the item, but **uses** this variable for final prediction.



A comparison of structural analysis methods

| | Characterize Representations | Causal Inference |
|---------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
| Probing |  |  |
| Feature Attribution |  |  |
| Causal Intervention |  |  |

Part IV: Open Problems in Interpretability

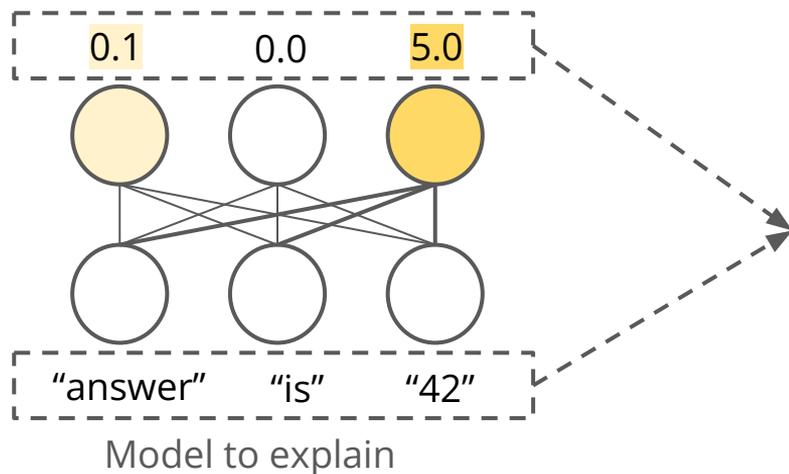
- Evaluation
- Broader applications

Evaluation

- **Faithful:** Interpretability methods should accurately represent the model to be explained.
- **Causal:** Interpretability methods should identify causal effects of model components on model input-output behaviors. **Causal interventions**
- **Generalizable:** The causal effects of the identified components should generalize to similar inputs. **Train-val-test Paradigm** It makes correct predictions for.
- **Isolating individual concepts:** Interpretability methods should isolate causal effects of individual concepts. **“Stress testing”:** Include hard negative examples

A case study: evaluating the auto-interpretability pipeline

What is an auto-interpretability pipeline?

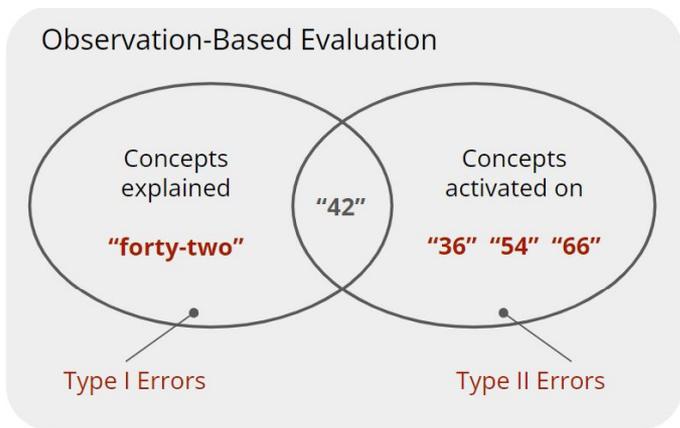


The neuron activates on words related to 42.

Natural language explanations generated by LLMs

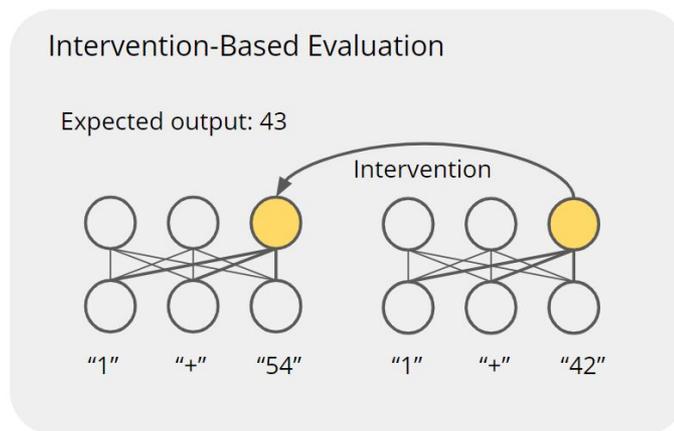
A case study: evaluating the auto-interpretability pipeline

The proposed explanation E : The feature f activates on words related to 42.



Observational: Test if a feature f activates on all and only input strings that refer to a concept picked out by E .

Probing



Interventional: Test if a feature f is a causally active representation of the concept denoted by E .

Causal interventions

A case study: evaluating the auto-interpretability pipeline

Evaluate 300 high confident explanations of GPT2-XL neurons generated and scored by GPT-4.

Neuron 1:3028

Explanation

days of the week.

score: 0.89

from 6:30 p.m. – 9:00 p.m.

Thursday November 15, from 6:30 p.m. – 9:00 p.m.

Thursday November 22, from 6:30 p.m. – 9:00 p.m.

Thursday November

lineup tonight, sitting Alex Avila for James McCann and Victor Martinez (a switch-hitter) on the disabled list. Worse yet from a fantasy standpoint is that another dud would likely tank his value in redraft formats given his injury history and lack of track record as a stud. Watch the fastball location,

championship Saturday, and after Austin Vandegrift won its first-ever title 1-0, fans flooded the field. A uniformed Georgetown police office was caught on film pushing and tripping students as they ran to celebrate with their classmates.

The Georgetown Police Department told KXAN that it was investigating the matter

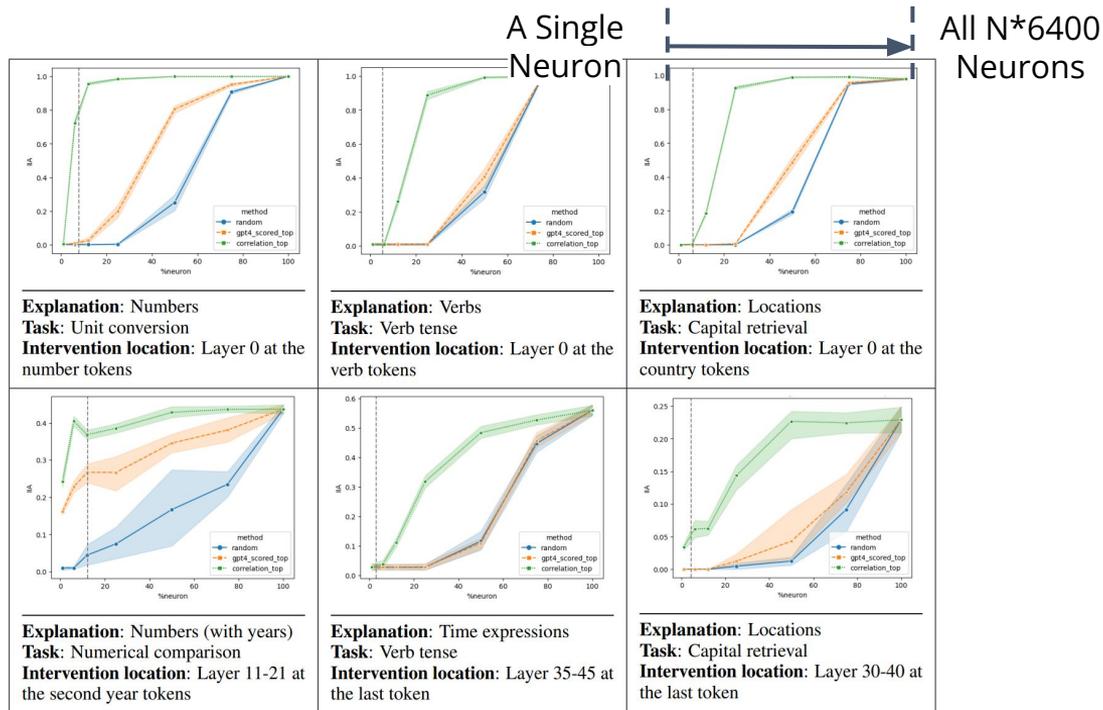
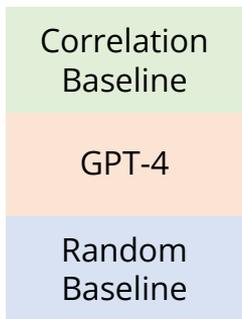
A case study: evaluating the auto-interpretability pipeline

Under observational mode: precision 0.64 and recall 0.50 (F1-score 0.56)

| Explanation | True Positives | Type I Errors | Type II Errors |
|-------------------------------------------------|--------------------------------------------------------------|-------------------------------------------------------|--------------------------------------------------------------|
| days of the week | I have a music class every Wednesday evening | <u>Thursday</u> 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 <u>2003</u> . | We need to rev amp the website to attract more |
| the word "most" and words related to comparison | lottery is a singular event for most people . | She is the <u>most</u> talented artist in the group | Their hostility towards each other was palpable . |
| color-related words | the sky in vibrant shades of violet and pink . | garden bloom ed in shades of <u>mag</u> <u>enta</u> . | her lifelong dream , she opened her own bakery |

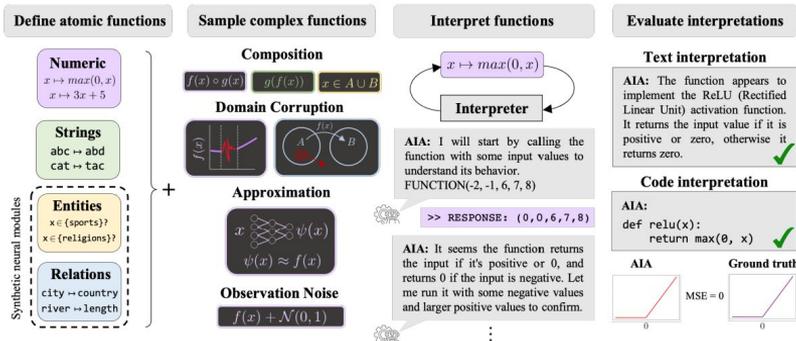
A case study: evaluating the auto-interpretability pipeline

Under interventional mode: Little to no causal efficacy, about same as the random baseline.

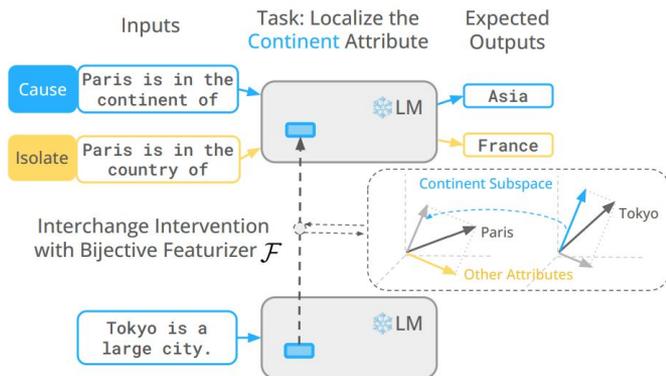


An active research area: interpretability benchmarks

FIND



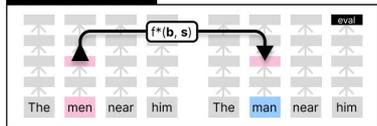
RAVEL



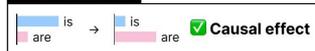
1. Minimal pairs



2. Intervention

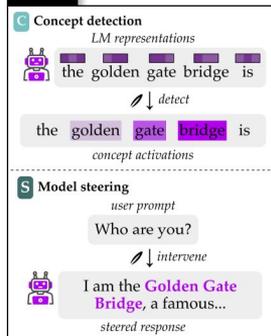


3. Evaluation

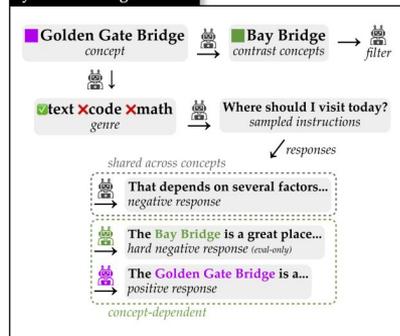


CausalGym

AxBench



Synthetic data generation

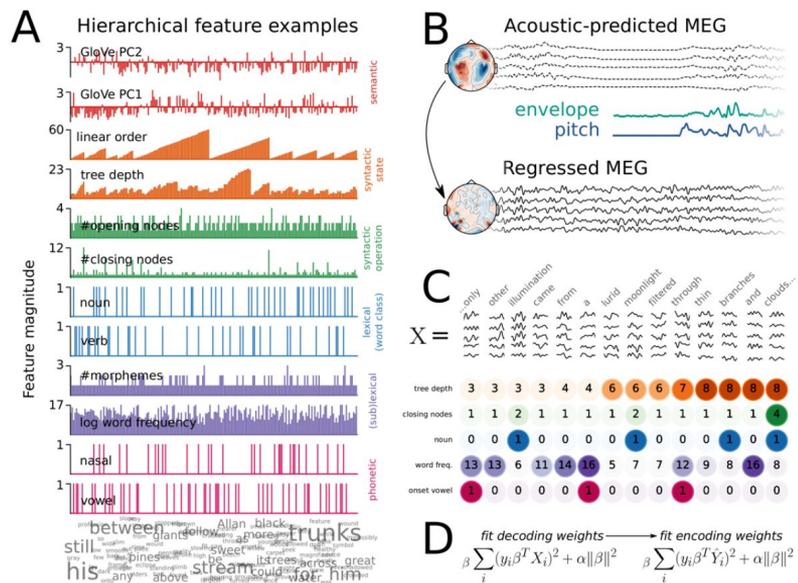


AxBench

Broader applications

Applications beyond machine learning model analysis and evaluation

- LLMs as test beds for linguistic theories
- Applications in cogsci/neurosciences



Recap

Motivations

Advanced behavioral analysis methods

- Stress testing

Structural analysis methods

- Probing
- Attribution
- Causal intervention

Open problems in interpretability