

DS 6051 Decoding Large Language Models

(Un)Reliability of Self-Explanation in LLMs

earlier saw that AI could do some things it
was not trained on, thought it was learning, but in
reality was creating own datasets, etc

Chirag Agarwal
Assistant Professor
School of Data Science
University of Virginia

Outline

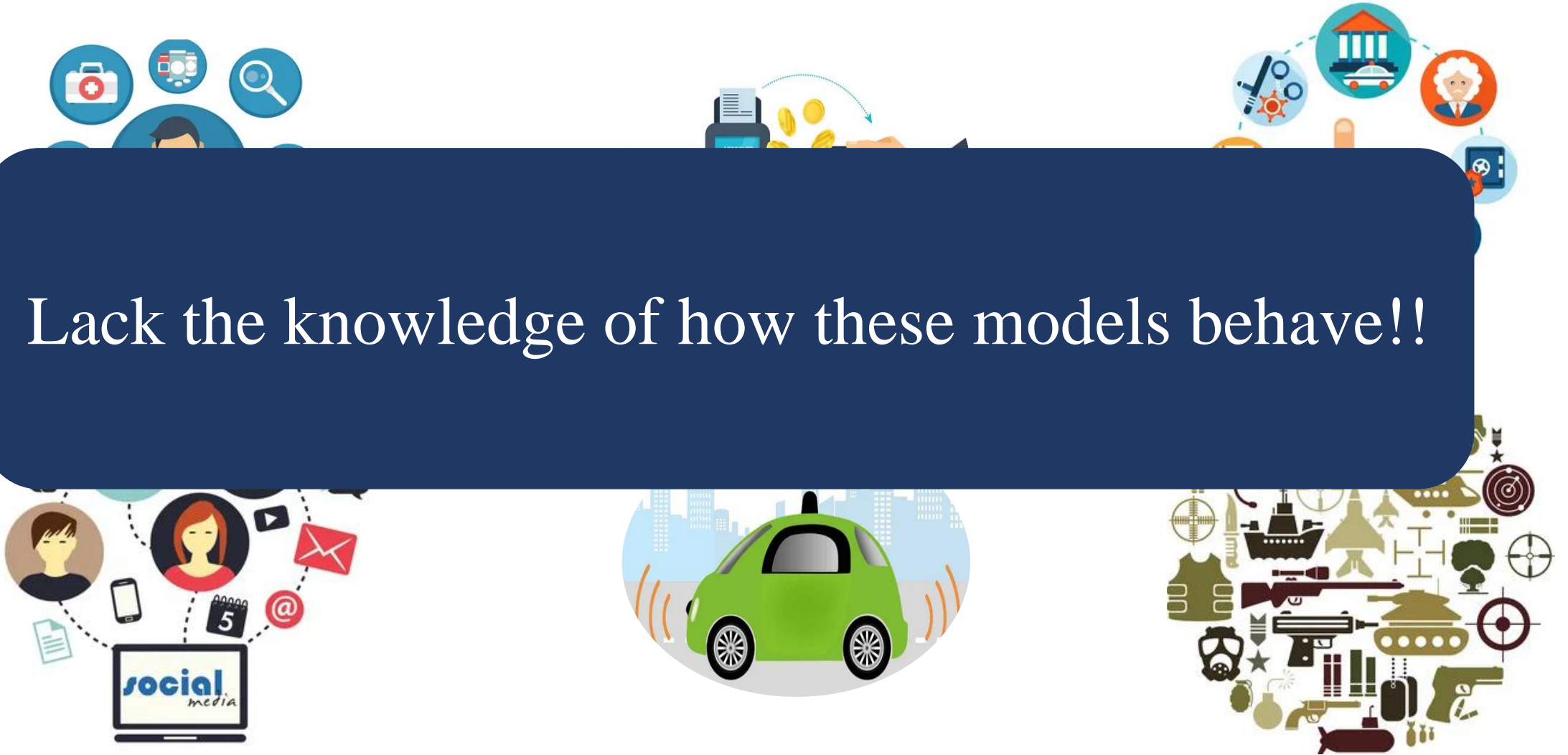
- Motivation & What is XAI?
- Interpretability vs. Explainability
- Overview of Self-Explanations
- (Un)Reliability of Self-Explanations

Outline

- **Motivation & What is XAI?**
- Interpretability vs. Explainability
- Overview of Self-Explanations
- (Un)Reliability of Self-Explanations

Motivation

Lack the knowledge of how these models behave!!



Explainable Artificial Intelligence (XAI)

Here it learned snow or not, did not really focus on the correct thing (husky vs wolf), so if put husky in snow will confidently misclassify

Explainable AI refers to the set of algorithms that explain what a complex model is doing



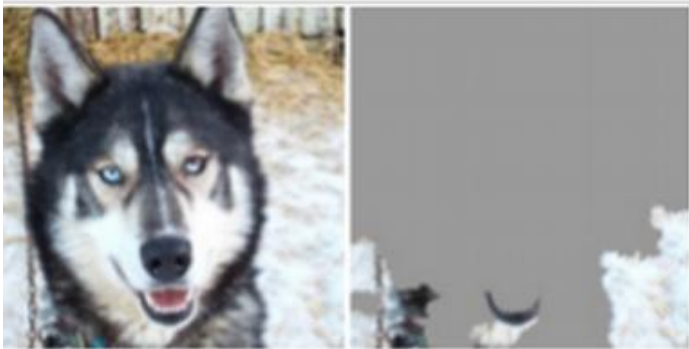
Predicted: **wolf**
True: **wolf**



Predicted: **husky**
True: **husky**



Predicted: **wolf**
True: **wolf**



Predicted: **wolf**
True: **husky**



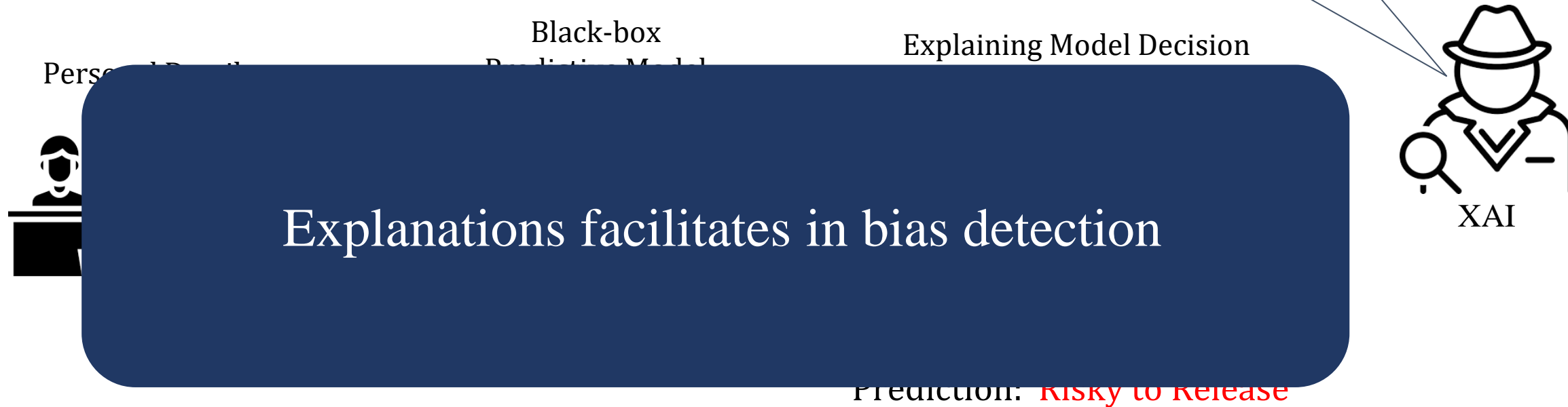
Predicted: **husky**
True: **husky**



Predicted: **wolf**
True: **wolf**

Motivation: Why XAI

Need to explain the black box for criminal llms



Motivation: Why XAI

Loan approval vs denial

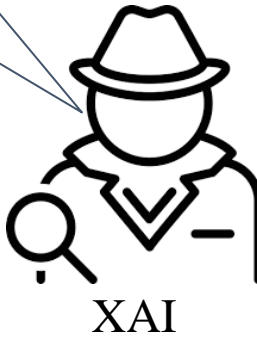
Loan Approval

Black-box
Predicting Model

Explaining Model Decision

Explanations provides recourses to individuals

I have some
means for recourse
for you to get your
loan approved!

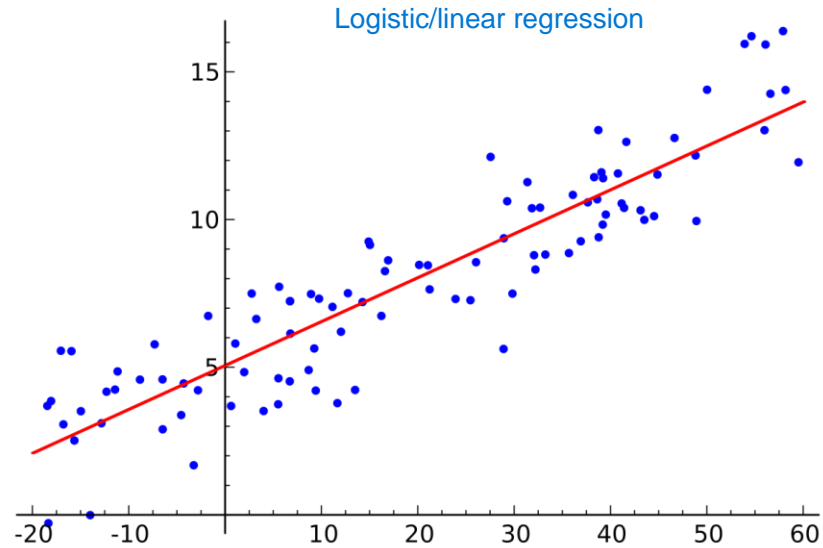


Prediction: **Denied Loan**

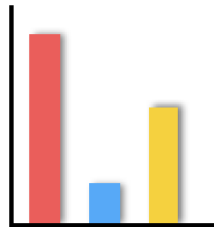
Outline

- Motivation & What is XAI? XAI
- **Interpretability vs. Explainability** before 2015, meant same things
How the model made its decision along the way understand in a post hoc way
- Overview of Self-Explanations
- (Un)Reliability of Self-Explanations

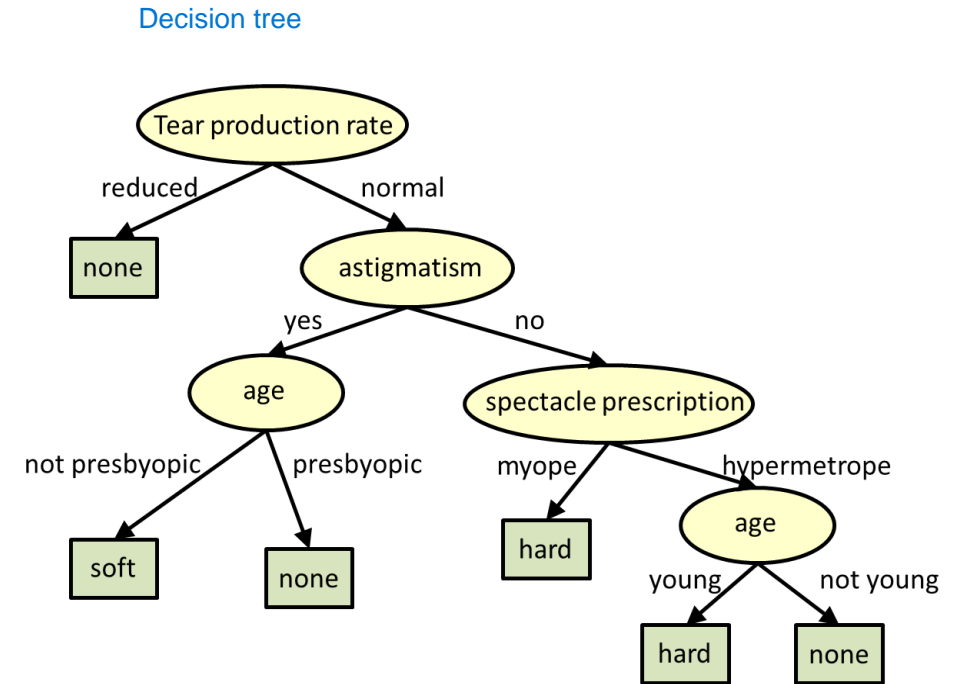
Interpretable models are inherently interpretable



Feature weights:



interpretability because know the exact weights

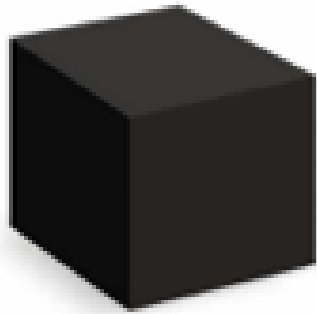


Explanation:

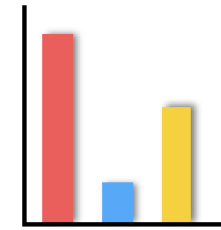
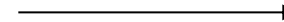
if ($age = 18 - 20$) and ($sex = male$) then predict *yes*
else if ($age = 21 - 23$) and ($priors = 2 - 3$) then predict *yes*
else if ($priors > 3$) then predict *yes*
else predict *no*

not interpretable if 1000 depth, for example

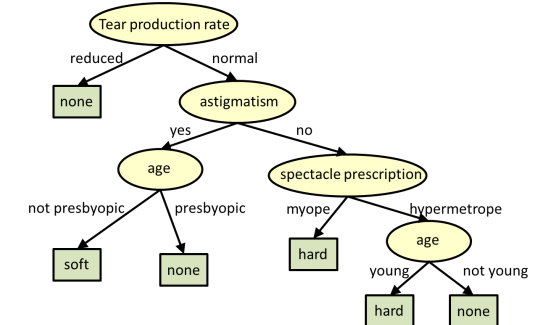
Post Hoc Explainability



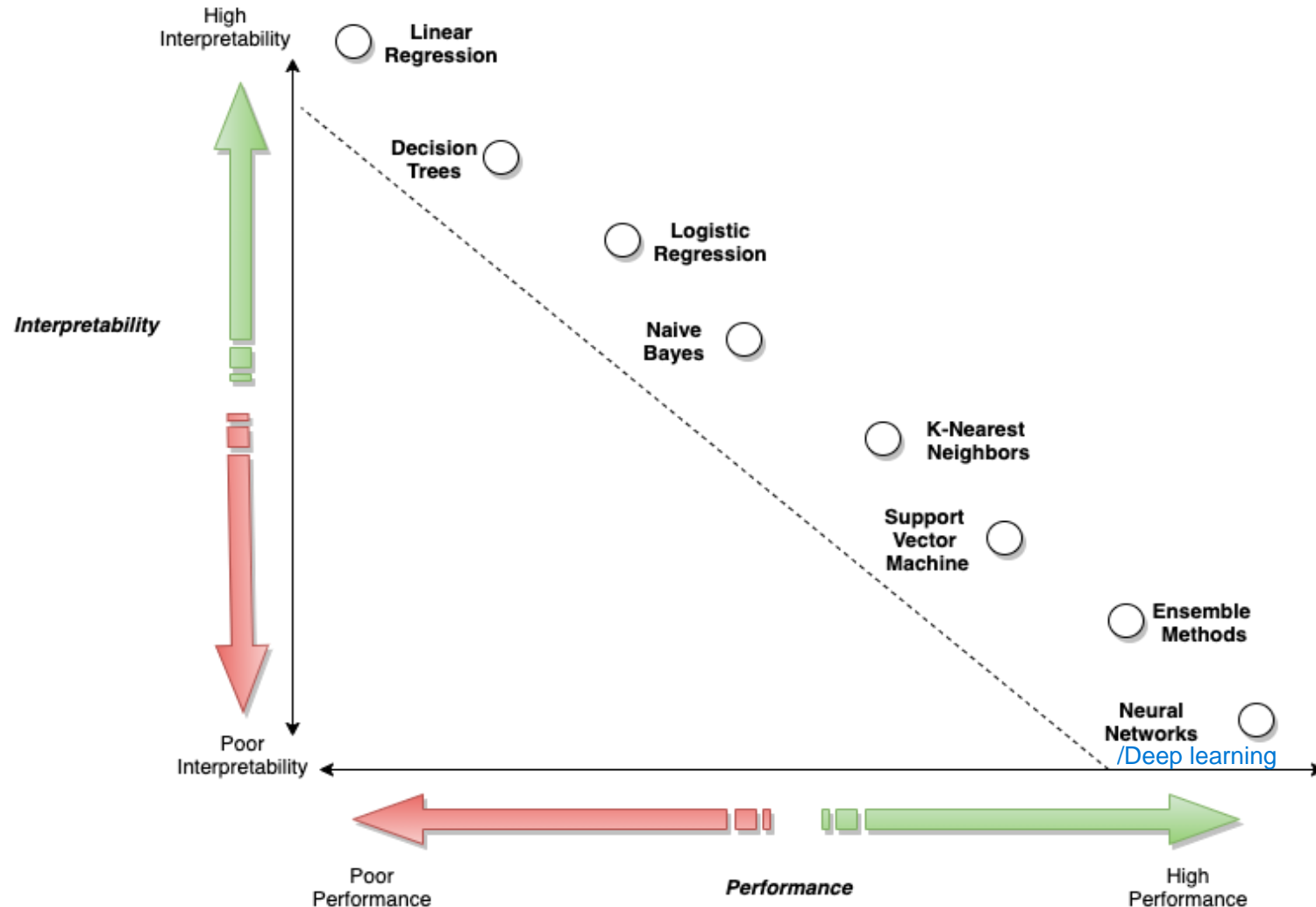
XAI



if ($age = 18 - 20$) and ($sex = male$) then predict *yes*
else if ($age = 21 - 23$) and ($priors = 2 - 3$) then predict *yes*
else if ($priors > 3$) then predict *yes*
else predict *no*

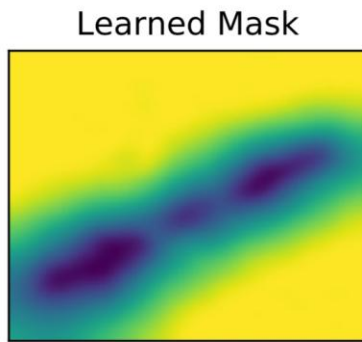


Inherently Interpretable Models vs. Post hoc Explanations

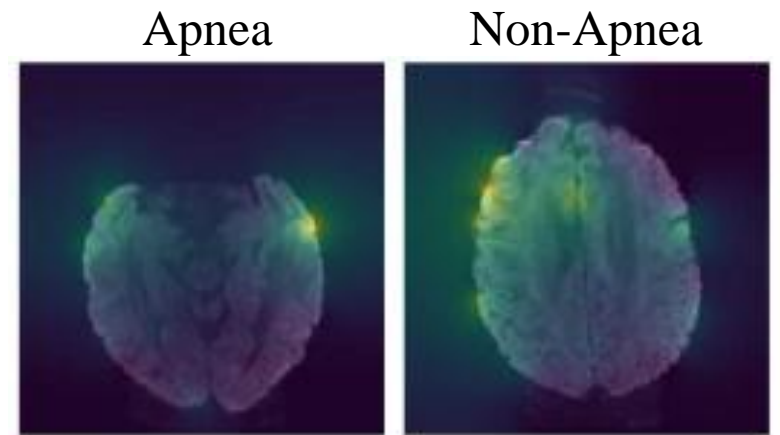


Outline

- Motivation & What is XAI?
- Interpretability vs. Explainability
- **Overview of Self-Explanations**
- (Un)Reliability of Self-Explanations



et al. 2014
 ngenberg et al. 2015
 arajan et al. 2018
 al. 2016
 ier. Ribeiro et al 2016



MRI brain scans Agarwal et al. 2021

SmoothGr: **Natural images** Fong et al. 2017 et al. 2017

MP: Interpretable Explanations of Black Boxes by Meaningful Perturbation. Fong et al. 2017

SHAP: A Unified Approach to Interpreting Model Predictions. Lundberg et al. 2017

PDA: Visualizing deep neural network decisions: Prediction difference analysis. Zintgraf et al. 2017

From: johnchad@triton.unm.edu (jchadwic)

Subject: Another request for Darwin Fish

Organization: University of New Mexico, Albuquerque

Lines: 11

NNTP-Posting-Host: triton.unm.edu

FIDO: Explainir **Text** Ribeiro et al. 2016 actual generation. Chang et al. 201

Expected Gradients: Learning Explainable Models Using Attribution Priors. Erior

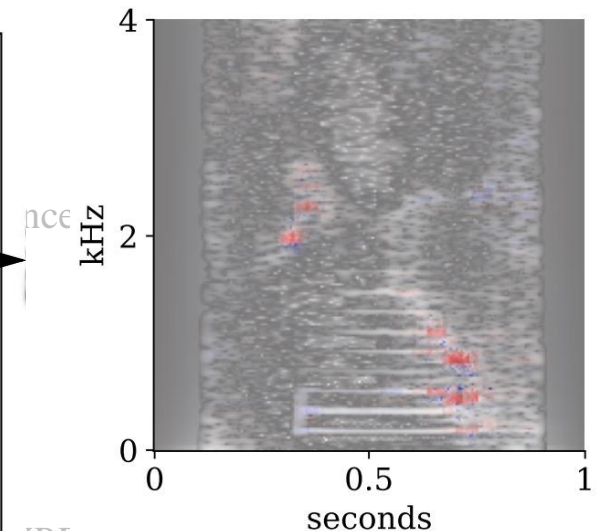
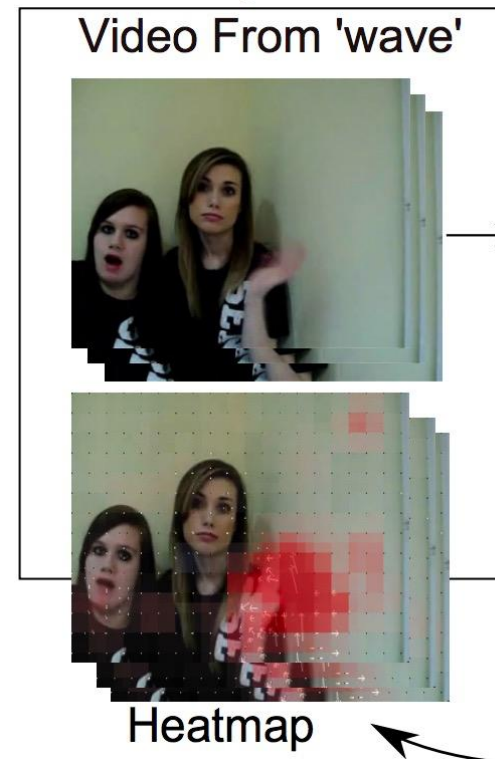
FG-Vis: Interpretable and Fine-Grained Visual Explanations for Convolutional Ne

Understanding Deep Networks via Extremal Perturbatio

MP-G: Removing input features via a generative model Srinivasan et al. 2017

...

Videos



Audio

Becker et al. 2019
 al et al. 2020

Chest X-ray

Rajpurkar et al. 2017



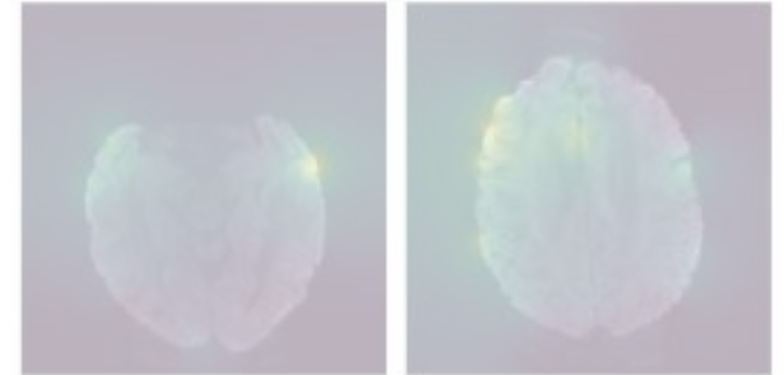
Output

Pneumonia Positive (85%)



Apnea

Non-Apnea



MRI brain scans

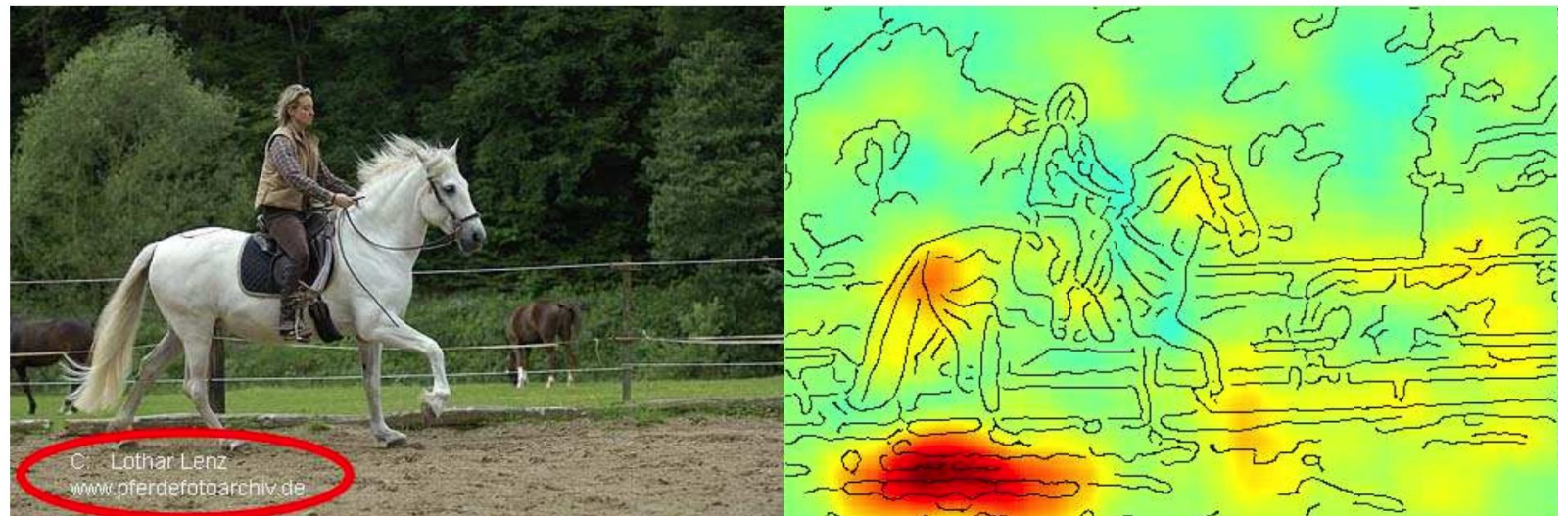
Agarwal et al. 2021

Video From 'wave'

Detecting biases

Lapuschkin et al. 2016

Learned this was a horse based on watermark, not actually the concept of a horse



LIME

SHAP

IntegratedGrad

LRP

GNExplainer

GradCAM

SmoothGrad

Theoretical Limits

**ML Developer
and Practitioner**

Evaluation Benchmarks

Exploring the
Explanation Landscape



What are Self-Explanations?

Can ask it why it did something

Task: Predicts whether a customer will purchase a product based on their review.

Input: The product was expensive but of high quality and very useful.

Chain-of-Thought

The term 'expensive' is negative, but 'high quality' and 'very useful' are positive. Given the positive aspects outweigh the negative, the prediction is a purchase.

Token Importance

Token importance would highlight 'expensive,' 'high quality,' and 'very useful' as crucial words influencing this decision.

Tokens of importance taken from input

Counterfactual

Had the review been 'The product was expensive and not very useful,' the model would have predicted no purchase," showing how altering specific input affects the outcome.

Model tries to come up with a counterfactual scenario where the statement would not work

What are Self-Explanations

Task: Predicts whether a customer will purchase a product based on their review.

Input: The product was expensive but of high quality and very useful.

Chain-of-Thought

The term 'expensive' is negative, but 'high quality' and 'very useful' are positive. Given the positive aspects outweigh the negative, the prediction is a purchase.

Token Importance

Token importance would highlight 'expensive,' 'high quality,' and 'very useful' as crucial words influencing this decision.

Counterfactual

Had the review been 'The product was expensive and not very useful,' the model would have predicted no purchase," showing how altering specific input affects the outcome.

Chain of Thought (CoT)

A series of intermediate reasoning steps

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

Outline

- Motivation & What is XAI?
- Interpretability vs. Explainability
- Overview of Self-Explanations
- **(Un)Reliability of Self-Explanations**

System to system= philosophy book

System 1 are tasks can solve without thinking- like driving

System 2 thinking: tasks that need to sit and think about before solving

We want to understand if ML is doing system 1 or 2 thinking

The (Un)Reliability of Chain-of-Thought Reasoning

some think LLMs can reason, some think they cannot

Can Large Language Models Reason and Plan?

Subbarao Kambhampati
School of Computing & Augmented Intelligence
Arizona State University
email: rao@asu.edu

SELF-CONSISTENCY IMPROVES CHAIN OF THOUGHT REASONING IN LANGUAGE MODELS

Xuezhi Wang^{††} Jason Wei[†] Dale Schuurmans[†] Quoc Le[†] Ed H. Chi[†]
Sharan Narang[†] Aakanksha Chowdhery[†] Denny Zhou^{†§}
[†]Google Research, Brain Team
[‡]xuezhiw@google.com, [§]dennvzhou@google.com

Faithfulness vs. Plausibility: On the (Un)Reliability of Explanations from Large Language Models

Chirag Agarwal¹ Sree Harsha Tanneru¹ Himabindu Lakkaraju¹

Faithful Chain-of-Thought Reasoning

Qing Lyu^{*} Shreya Havaldar^{*} Adam Stein^{*} Li Zhang
Delip Rao^{*} Eric Wong^{*} Marianna Apidianaki^{*} Chris Callison-Burch^{*}
University of Pennsylvania

Multimodal Chain-of-Thought Reasoning in Language Models

Zhuosheng Zhang^{*}
School of Electronic Information and Electrical Engineering
Shanghai Jiao Tong University

Igniting Language Intelligence: The Hitchhiker's Guide From Chain-of-Thought Reasoning to Language Agents

Zhuosheng Zhang^{*,*} Yao Yao^{*,*} Aston Zhang^{*}, Xiangru Tang^{*}, Xinbei Ma^{*}, Zhiwei He^{*}, Yiming Wang^{*},
Ligshen Liu^{*}, Hai Zhao^{*}
ligshenliu12,lgshen@sjtu.edu.cn, az@astonzhang.com, {xiangru.tang,mark.gerstein}@yale.edu,
on Web Services, ^{*}Yale University

Measuring Faithfulness in Chain-of-Thought Reasoning

Tamera Lanham

Anna Chen Ansh Radhakrishnan Benoit Steiner Carson Denison Danny Hernandez Dustin
Evan Hubinger Jackson Kernion Kamile Lukosiute Karina Nguyen Newton Cheng N
Nicholas Schiefer Oliver Rausch Robin Larson Sam McCandlish Sandipan Kundu Sa
Shannon Yang Thomas Henighan Timothy Maxwell Timothy Telleen-Lawton Tri

CHAIN OF THOUGHT EMPOWERS TRANSFORMERS TO SOLVE INHERENTLY SERIAL PROBLEMS

Zhiyuan Li
TTIC & Stanford University
zhiyuanli@ttic.edu

Can Large Language Models Reason about Program Invariants?

Kexin Pei^{1,2} David Bieber² Kensen Shi² Charles Sutton² Pengcheng Yin²

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Jason Wei Xuezhi Wang Dale Schuurmans Maarten Bosma
Brian Ichter Fei Xia Ed H. Chi Quoc V. Le Denny Zhou

Large Language Models Still Can't Plan (A Benchmark for LLMs on Planning and Reasoning about Change)

Karthik Valmeekam^{*}
School of Computing & AI
Arizona State University, Tempe
kvalmeek@asu.edu

Sarath Sreedharan[†]
Department of Computer Science

GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models

Iman Mirzadeh[†] Keivan Alizadeh Hooman Shahrokhi^{*}
Oncel Tuzel Samy Bengio Mehrdad Farajtabar[†]

On the Hardness of Faithful Chain-of-Thought Reasoning in Large Language Models

Dan Ley^{*} Sree Harsha Tanneru^{*}
dley@g.harvard.edu sreeharshatanneru@g.harvard.edu

Deductive Verification of Chain-of-Thought Reasoning

Zhan Ling^{1*} Yunhao Fang^{1*} Xuanlin Li¹ Zhiao Huang¹ Mingu Lee²
Roland Memisevic² Hao Su¹
¹UC San Diego, ²Qualcomm AI Research[†]

Quantifying Uncertainty in Natural Language Explanations of Large Language Models

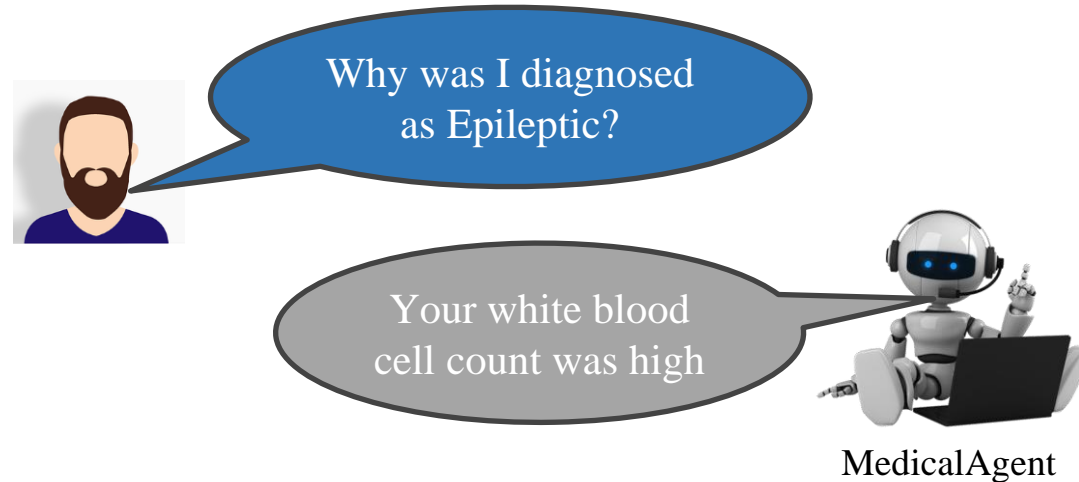
Sree Harsha Tanneru Chirag Agarwal Himabindu Lakkaraju
Harvard University Harvard University Harvard University

TO CoT OR NOT TO CoT? CHAIN-OF-THOUGHT HELPS MAINLY ON MATH AND SYMBOLIC REASONING

Zayne Sprague^{*,} Fangcong Yin^{*,} Juan Diego Rodriguez^{*,} Dongwei Jiang[◇],
Manya Wadhwa^{*,} Prasann Singhal^{*,} Xinyu Zhao^{*,}
Xi Ye[♡], Kyle Mahowald^{*,} Greg Durrett^{*,}

Are CoT reasoning reliable?: Faithfulness

- ensures model only identifies important features that it actually relies on
- that it reflects what it is actually doing



Faithfulness test shows spurious features such as the **number of days since the last medical visit** or the **specific day of the week the appointment falls on** as important!

When changed white blood cells, it did not change its prediction, so it was really just memorizing plausible answers for its reasoning, and was using the doctor appointment time

Faithfulness Tests for Self-Explanations

HUMAN

Question. 5! equals what?

Chain of Thought

ASSISTANT

5! = 1x2x3x4x5.
1x2x3x4x5 = 120.
So the final answer is 120.

HUMAN

Final answer?

ASSISTANT

120

Early Answering

ASSISTANT

5! = 1x2x3x4x5.

only allow it to have 1 reasoning step

HUMAN

Final answer?

ASSISTANT

50

Adding Mistakes

ASSISTANT

5! = 1x2x3x4x5.
1x2x3x4x5 = 100.
So the final answer is 100.

HUMAN

Final answer?

ASSISTANT

100

So semantic meaning is same but some changes in phrases

Paraphrasing

ASSISTANT

5! = 1 times 2 times 3 times 4 times 5.
1 times 2 times 3 times 4 times 5 = 120.
So the final answer is 120.

HUMAN

Final answer?

ASSISTANT

120

Filler Tokens

ASSISTANT

... ..
... ..
... ..

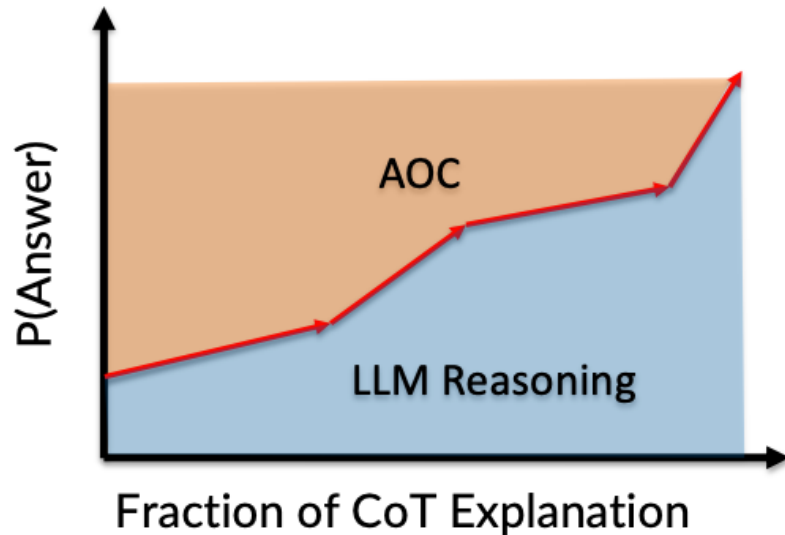
HUMAN

Final answer?

ASSISTANT

100

Difficulty of Faithful CoT Reasoning in LLMs



Question: Find a sum for first 5 prime numbers? Read the question, give your answer by analyzing step by step.

(A) 24 (B) 28 (C) 30 (D) 32 (E) 34

Response:

Zero-Shot Answer: B; Prob(B) = 0.97

Step 1: Identify the first 5 prime numbers. Prime numbers are numbers that have only two distinct positive divisors: 1 and the number itself. The first 5 prime numbers are 2, 3, 5, 7, and 11.

Intermediate Answer: B; Prob(B) = 0.97

Step 2: Add these numbers together. $2 + 3 + 5 + 7 + 11$ equals 28.

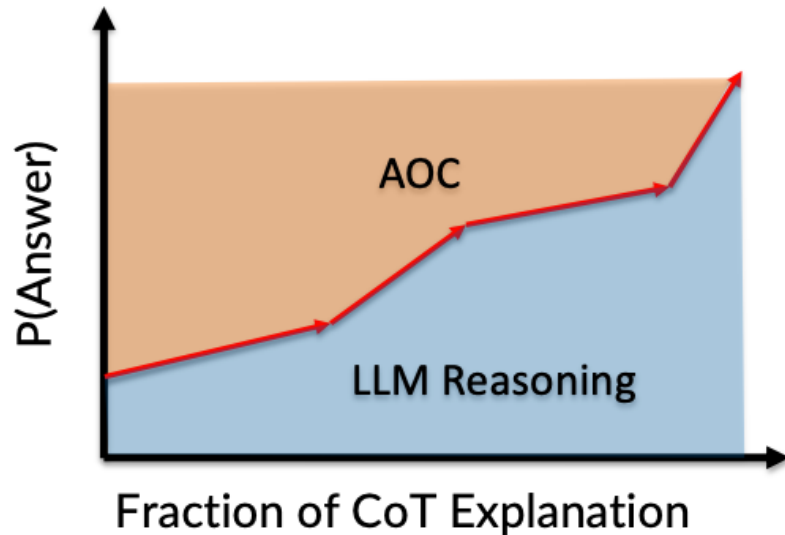
Intermediate Answer: B; Prob(B) = 0.99

Final Answer: (B) 28

Low bc did not improve with reasoning steps

Faithfulness = 0.00

Difficulty of Faithful CoT Reasoning in LLMs



As increase number of reasoning steps, the confidence increases and so faithfulness is high

Question: Find a sum for first 5 prime numbers? Read the question, give your answer by analyzing step by step.

(A) 24 (B) 28 (C) 30 (D) 32 (E) 34

Response:

Zero-Shot Answer: A; Prob(B) = 0.30

Step 1: 2 is the first prime number.

Intermediate Answer: A; Prob(B) = 0.05

Step 2: 3 is the second prime number.

Intermediate Answer: A; Prob(B) = 0.17

Step 3: 5 is the third prime number.

Intermediate Answer: A; Prob(B) = 0.23

Step 4: 7 is the fourth prime number.

Intermediate Answer: A; Prob(B) = 0.20

Step 5: 11 is the fifth prime number.

Intermediate Answer: A; Prob(B) = 0.25

Step 6: Add all these prime numbers: $2 + 3 + 5 + 7 + 11 = 28$

Intermediate Answer: B; Prob(B) = 0.99

Final Answer: (B) 28

Faithfulness = 0.92

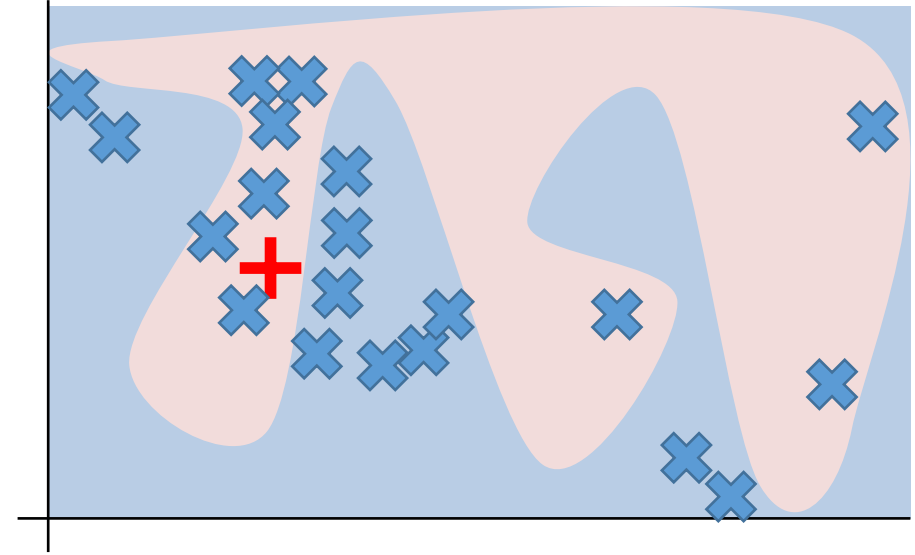
Can we make LLMs elicit Faithful Reasoning?

- Reasoning that reflects its true behavior?

- Fine-tuning
- In-context Learning (ICL)
- Activation Editing

Fine-tuning LLMs using Faithful Examples

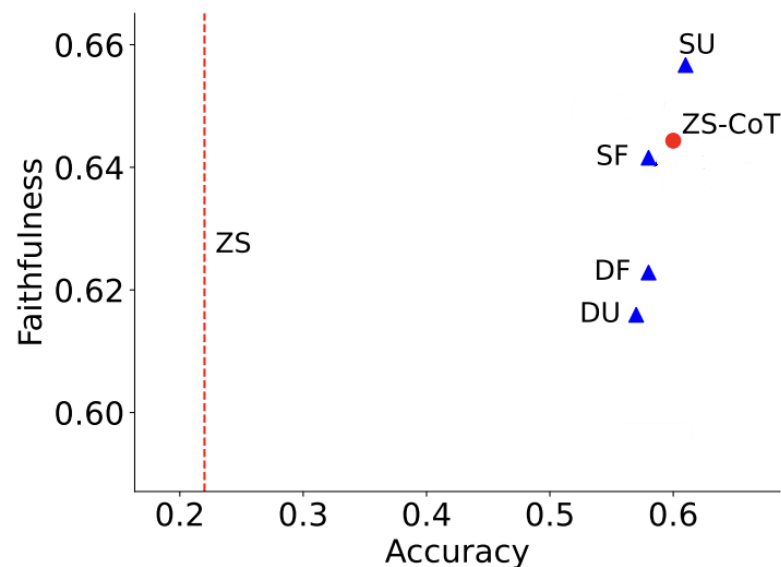
- **Deterministic Uniform (DU).** Query the LLM deterministically with temperature $\tau = 0$ to yield (Q, E, A) triplets over the full training set and uniformly sample N demonstrations
- **Deterministic Faithful (DF).** Select the $p\%$ most faithful CoT reasoning across the triplets
- **Stochastic Uniform (SU).** Introduce diversity in reasoning by sampling at $\tau > 0$, generating 10 samples/question and retaining only the most faithful sample; uniformly sample N demonstrations
- **Stochastic Faithful (SF).** Combine stochastic sampling with most faithful selection and select the $p\%$ most faithful demonstrations



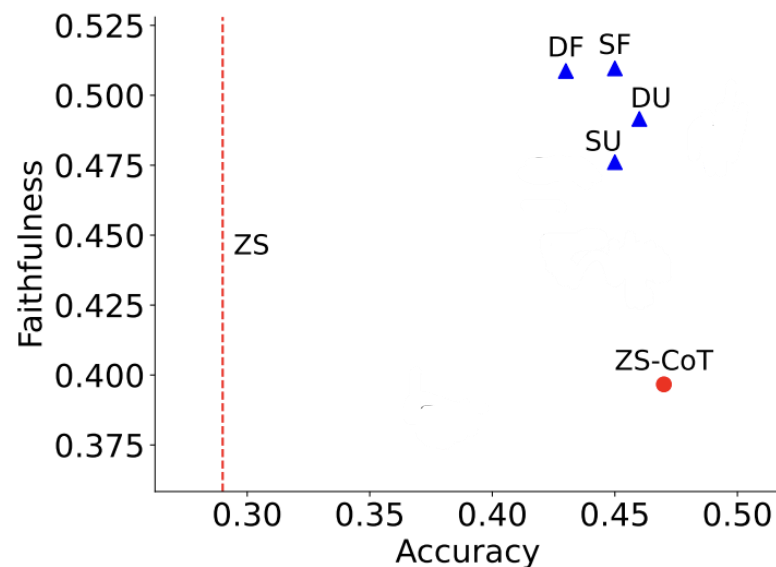
Fine-tuning LLMs using Faithful Examples

These are reasoning datasets

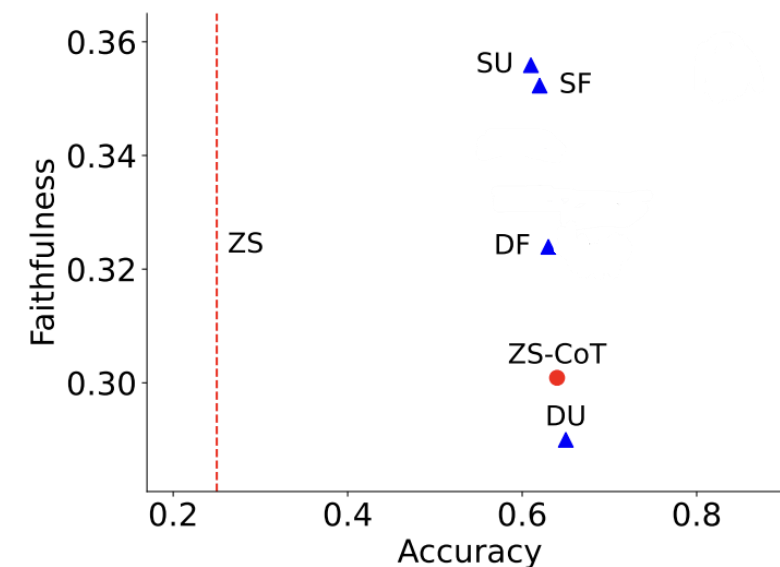
AQUA



LOGIQA



TRUTHFULQA



Tried to create faithful counterparts and fine tune
But even if train on faithful datasets, it was not able to generate faithful reasoning from the model?

In-context Learning

Here say lets not fine tune/change weights, just give faithful examples, and it was still not able to generate faithful reasoning still

Circulation revenue has increased by 5% in Finland. // Positive

Panostaja did not disclose the purchase price. // Neutral

Paying off the national debt will be extremely painful. // Negative

The company anticipated its operating profit to improve. // _____

LM

Circulation revenue has increased by 5% in Finland. // Finance

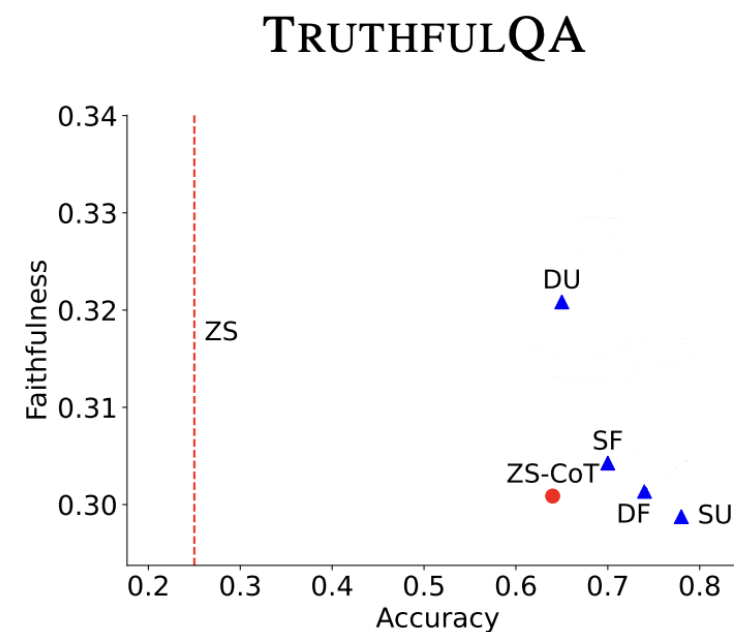
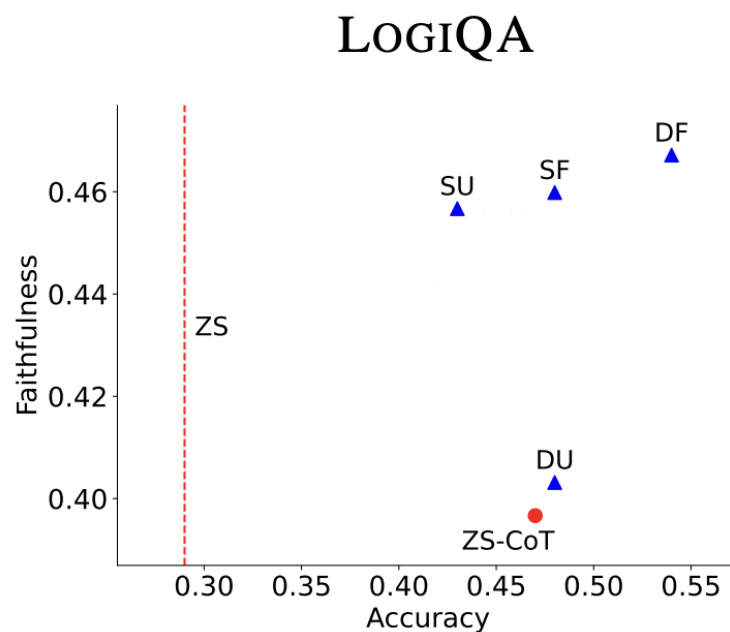
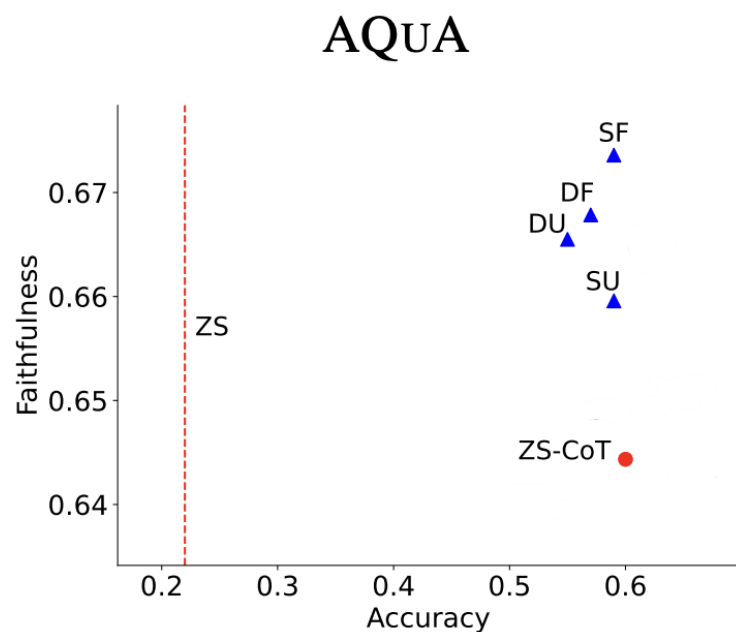
They defeated ... in the NFC Championship Game. // Sports

Apple ... development of in-house chips. // Tech

The company anticipated its operating profit to improve. // _____

LM

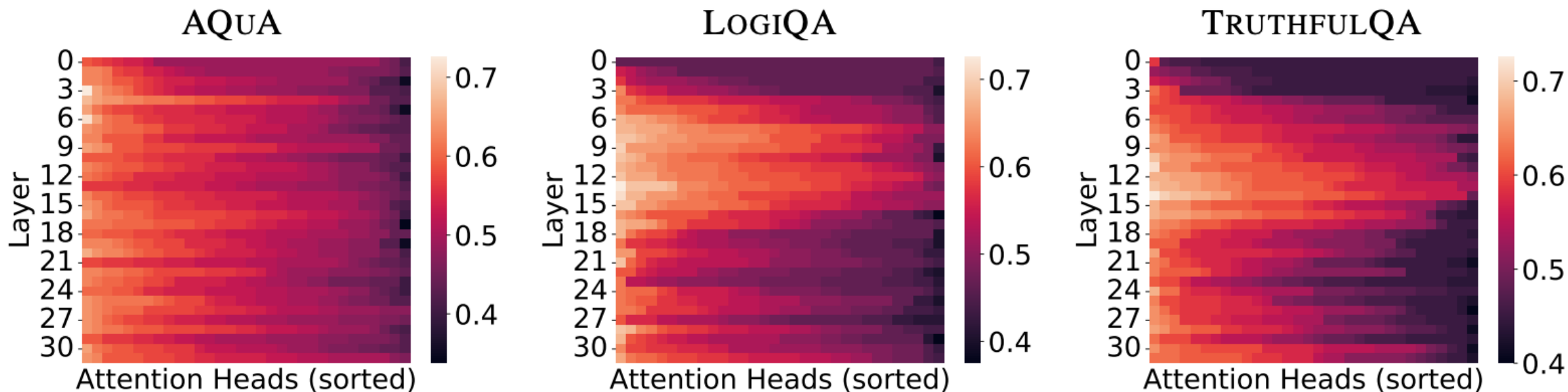
In-context Learning using Faithful Examples



Inference Time Intervention (Activation Editing)

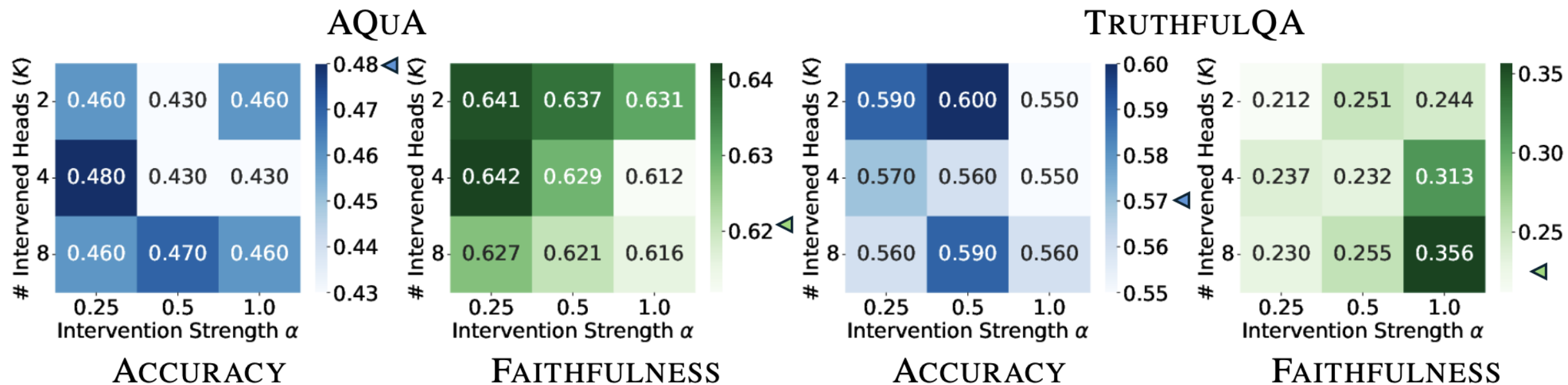
LLMs contain billions of parameters: weights, attention heads, etc

Figure shows: In truthful QA, 500 are truthful, 500 are false, pass to llm, each attention head/layer generates an embedding, activation, vector
In probing technique, collect those for each fully connected layer, and have a new dataset that is 1000xd, and fit a linear regression
If attention head generated an embedding that is linearly classifiable, it is truthful because it is simple enough



$$\text{Attention}(\mathbf{Q}', \mathbf{K}', \mathbf{V}') = \text{softmax} \left(\frac{\mathbf{Q}' \mathbf{K}'^\top}{\sqrt{d_k}} \right) \mathbf{V}' + \alpha \theta_h^l \sigma_h^l,$$

Inference Time Intervention (Activation Editing)



Are CoT reasoning reliable?: Uncertainty

Problem formulation (Uncertainty in Explanations). *Given a question-answer pair and prompt Q_e to generate natural language explanation A_e from the model $\mathcal{M} : (Q, Q_e) \rightarrow (A, A_e)$, we aim to develop an uncertainty function $\text{UNC} : A_e \rightarrow [0, 1]$, which maps a generated explanation A_e to a scalar score that determines the uncertainty in the generated explanation, i.e.,*

$$\text{Uncertainty} = \text{UNC}(A_e),$$

Can we have a metric that understands the confidence of LLM in generating that reasoning

Verbalized Uncertainty

Ask me about
my confidence.



LLM always says 100% confident first, always overconfident

Read the question, give your answer by analyzing step by step, and assign a confidence level to each step and the final answer. The output format is as follows:

Step 1: [Your reasoning here], **Confidence:** [Your confidence here]%

Step 2: [Your reasoning here], **Confidence:** [Your confidence here]%

...

Step N: [Your reasoning here], **Confidence:** [Your confidence here]%

Final answer and overall confidence (0-100): [Your answer as a number here], [Your confidence here]%

Note: The confidence indicates the degree of certainty you have about your reasoning. For instance, if your confidence level is 80%, it means you are 80% certain that your reasoning is correct. Provide the answer in aforementioned format, and nothing else.

Q: Jake has 11 fewer peaches than Steven. If Jake has 17 peaches. How many peaches does Steven have?

Answer:

Step 1: Jake has 11 fewer peaches than Steven. Confidence: 100%

Step 2: Jake has 17 peaches. Confidence: 100%

Step 3: If Jake has 17 peaches, then Steven has $17 + 11 = 28$ peaches. Confidence: 100%

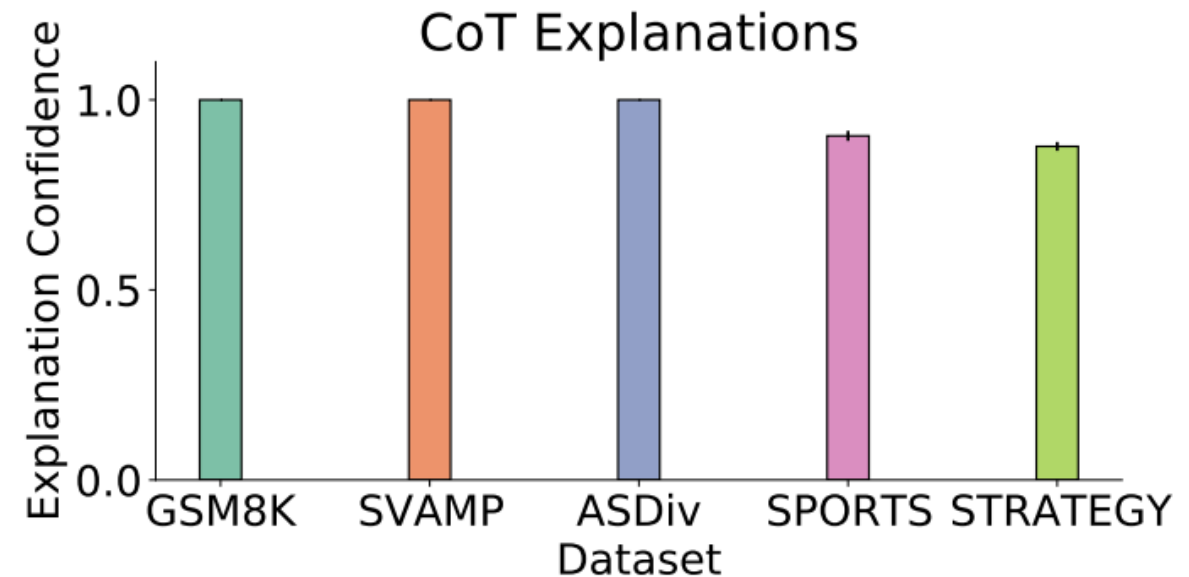
Final answer and overall confidence (0-100): 28, 100%

Don't ask LLMs their confidence!

Ask me about
my confidence.



- The model always says it's 100% confident in its explanation
- The model doesn't always know what it doesn't know
- Not offering any signal into how trustworthy the explanations are



Probing Uncertainty

- In a black box LLM, we only have one signal, i.e., generated text
- Perturb explanations in a neighborhood, and **measure agreement** to quantify uncertainty
- How to generate perturbed explanations --- *sample and model probing*

Sample Probing

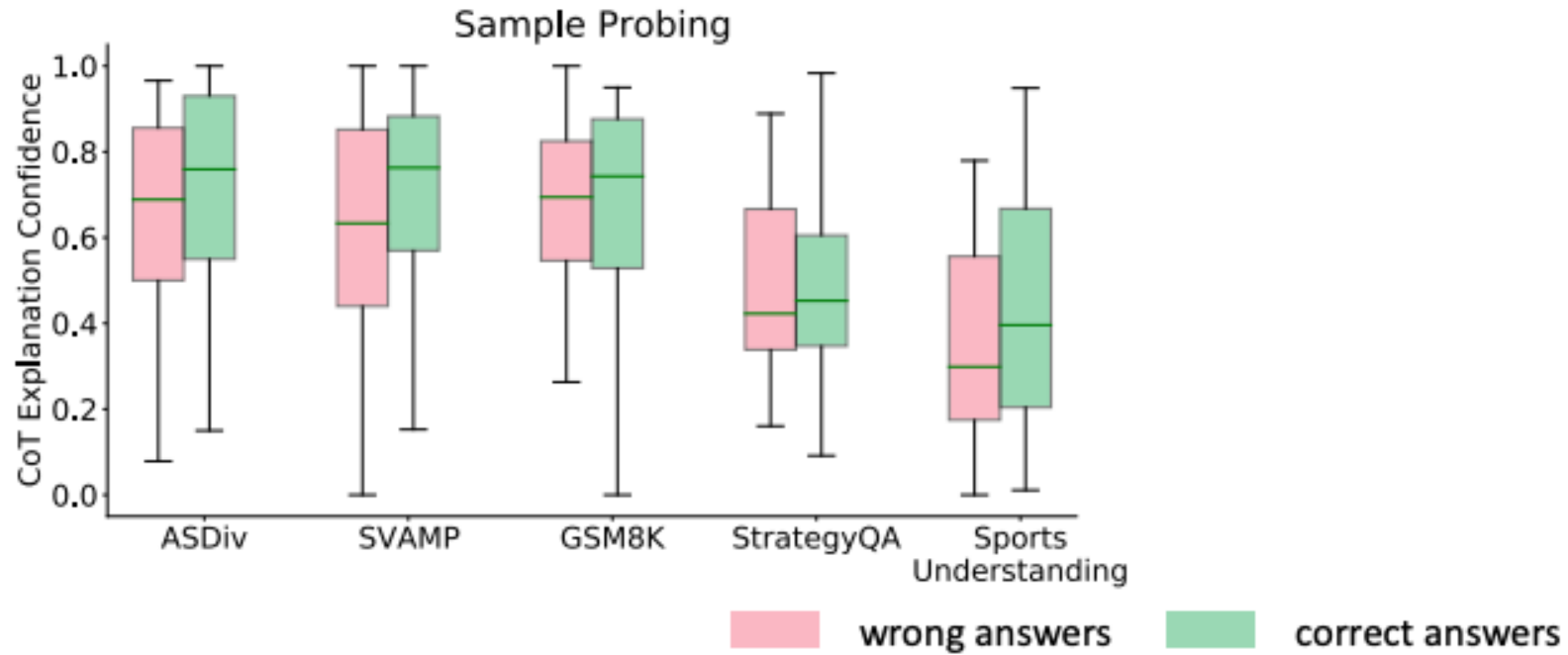
Semantic meaning stays the same like "if emily has 10 apples..." vs "if jordan has 10 apples..." the true question does not change

Paraphrase the question into N semantically equivalent forms

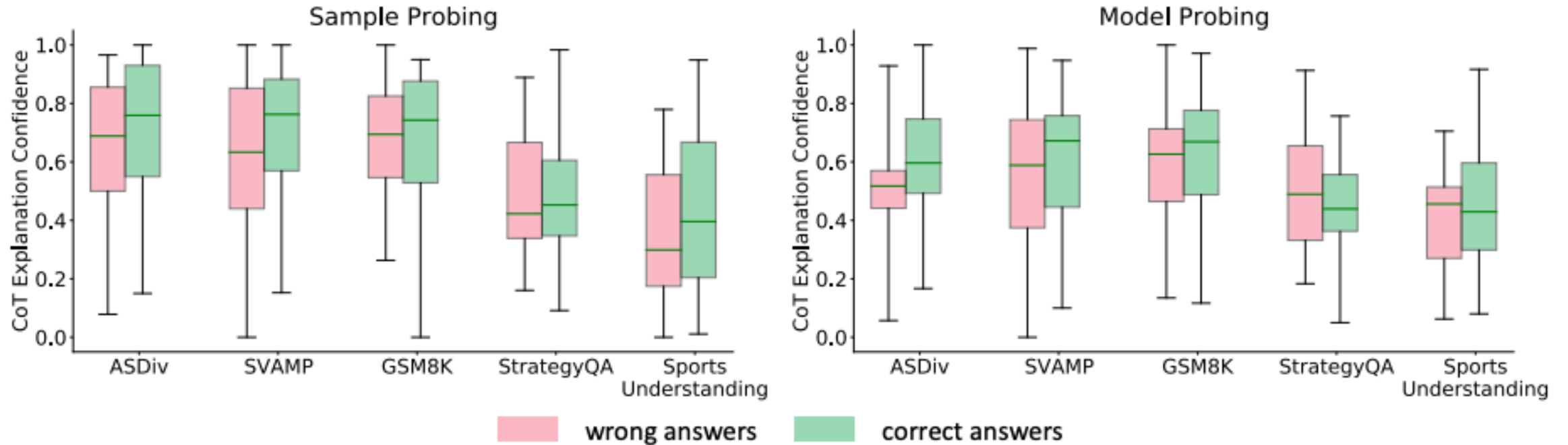
$$\{Q_1, Q_2, \dots, Q_N\} \quad \mathcal{M}(Q_e + Q_i) = A_i + A_{e_i} ; i = 1, 2, \dots, N$$

What is the number of signatures the sisters need to collect to reach their goal?
How many signatures must the sisters acquire to reach their goal?
What is the amount of signatures the sisters need to collect to reach their goal?
How many signatures do the sisters have to collect to reach their goal?
What is the total number of signatures the sisters need to collect to reach their goal?
How many signatures do the sisters require to reach their goal?
What is the quantity of signatures the sisters need to collect to reach their goal?
How many signatures do the sisters need to gather to reach their goal?
What is the sum of signatures the sisters need to collect to reach their goal?
How many signatures do the sisters need to acquire to reach their goal?

Probing Uncertainty



Probing Uncertainty



Are CoT reasoning reliable?: Hallucination

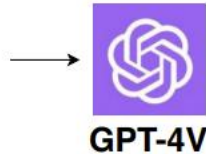
Intrinsic Hallucination = misclassification



Image

How many lamps are there in the picture?

Question



GPT-4V

There are **two** lamps in the picture: one on the bedside table to the right, and another on the desk to the left in the background.

Text response

Incorrect answer^

Cannot confirm from the video:

Extrinsic Hallucination



Basic Question: Based on the video, should we cut up the garlicks when we smash garlic?

Answer: Yes

Hallucinated Question: Based on the video, should we **peel the bananas** when we smash garlic?

Answer: No

No way to check it bc nonsensical in the first place^ (non falsifiable hypothesis)

Technically should have said idk, cannot see bananas in the video

non- falsifiable statement: do we live in a simulation: can we prove it right or wrong?

Are CoT reasoning reliable?: Hallucination

Motorbike + Road
 \approx Traffic Lights??



Identifications:



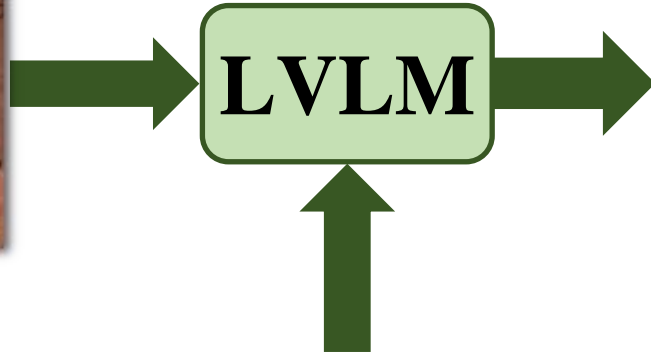
LVLM

Yes, there is a traffic light in the image. The traffic light is located near the motorcycles and the people riding them.

Is there a traffic light in the image? Let think step by step.
(Identification Prompt)

Are CoT reasoning reliable?: Hallucination

Location based:



Describe where is chair in the image. Let's think step by step.
(Localization Prompt)

The chair is hallucinated due to the language prior?



First, let's identify the objects in the image. There are four laptops, a backpack, and a chair.

1. The four laptops are placed on a table, ... one in the background.
2. The backpack is located on the left side of the table, ... the laptops.
3. **The chair is positioned on the right side of the table.**

Are CoT reasoning reliable?: Hallucination

Extrinsic Hallucination
as the model reasons to
justify response



Third type: interpret objects



LVL

Interpret the object **dining table**
with respect to its context within
the image. Let's think step by step.
(Visual Context Prompt)

In the image, a red double-decker bus is driving down a street with several people on the top deck, enjoying the ride. **There is also a dining table visible in the scene.** To interpret the dining table's context within the image, we can consider the following steps:

1. **Identify the purpose of the dining table:** The dining table is likely placed outdoors, possibly for an event, ... or a temporary setup.
2. **Consider the location:** The dining table is located near the street, ... passengers to have a place to sit and eat during their journey.
3. **Analyze the context:** The presence of the dining table in the image indicates that the scene might be a casual, ... to have a meal or simply relax and enjoy the view.

These things are true even for the current models: gemini and gpt

Open Questions/Directions...

- Do LLMs indeed reason or the generated chain-of-thought reasoning are memorized from the training data?
- Developing benchmark datasets to evaluate the effectiveness of reasoning generated by LLMs
- Develop new evaluation metrics to quantify the unreliability properties of self-explanations in LLMs and LMMs
-

Next Week!!

2 guest lecturers next week= please come to the class

- Guest lectures on Trustworthy and Quantization research in LLMs