Transformer-specific Interpretability

Hosein Mohebbi, **Jaap Jumelet**, Michael Hanna, Afra Alishahi, Willem Zuidema

Introduction

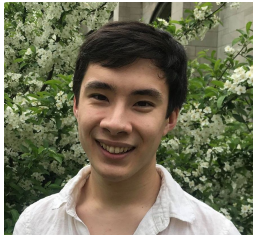
Tilburg University

Hosein Mohebbi

Afra Alishahi

University of Amsterdam

Jaap Jumelet



Michael Hanna

Willem Zuidema

Plan for Today

1. Model-agnostic

- Why interpretability?
- Faithfulness
- Probing
- Feature attributions
- Limitations

Plan for Today

1. Model-agnostic

- Why interpretability?
- Faithfulness
- Probing
- Feature attributions
- Limitations

2. Context Mixing

- Attention analysis
- Limits of attention
- Attention flow
- ALTI
- Value Zeroing

+ Hands-on **tutorial**

Plan for Today

1. Model-agnostic

- Why interpretability?
- Faithfulness
- Probing
- Feature attributions
- Limitations

2. Context Mixing

- Attention analysis
- Limits of attention
- Attention flow
- ALTI
- Value Zeroing

+ Hands-on **tutorial**

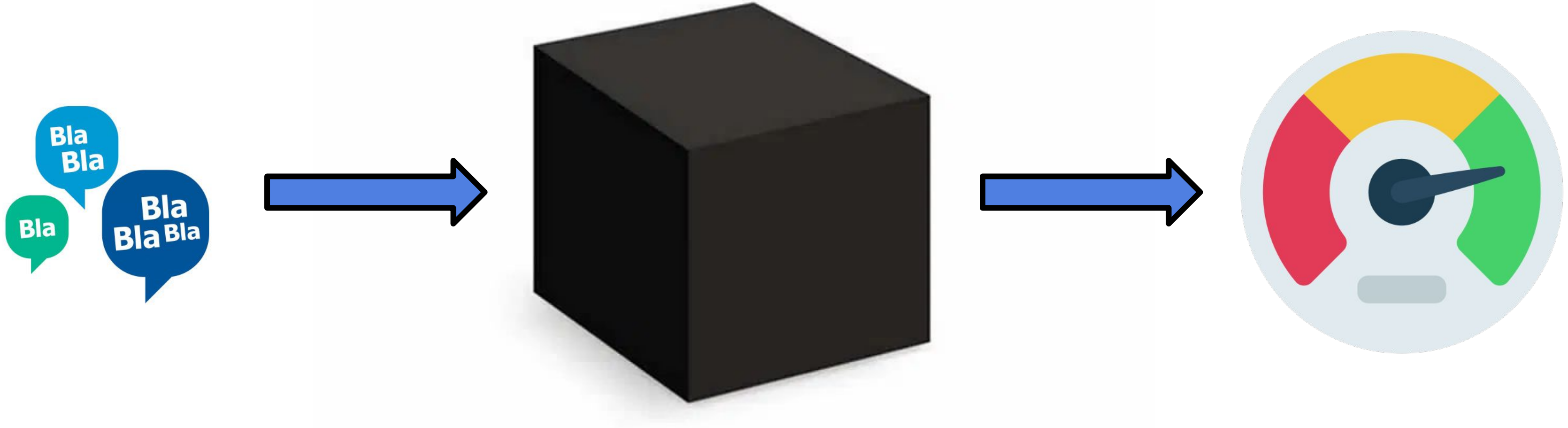
3. Mechanistic

- Residual stream
- Circuits
- Circuit Discovery
- Activation patching
- Logit Lens

+ Hands-on **tutorial**

+ General Discussion

Why do we need interpretability?



Why do we need interpretability?

The **desiderata** of algorithmic models:

1. **Fairness**
    - *What biases does it contain? Does it discriminate against particular groups?*

Lipton (2018) - *The Mythos of Model Interpretability*

Why do we need interpretability?

The **desiderata** of algorithmic models:

1. **Fairness**
   - *What biases does it contain? Does it discriminate against particular groups?*
2. **Trustworthiness**
   - *Models that are deployed carry a degree of responsibility, can we trust them?*

Lipton (2018) - *The Mythos of Model Interpretability*

Why do we need interpretability?

The **desiderata** of algorithmic models:

1. **Fairness**
    - *What biases does it contain? Does it discriminate against particular groups?*
2. **Trustworthiness**
    - *Models that are deployed carry a degree of responsibility, can we trust them?*
3. **Robustness**
    - *Does our model generalise robustly to unseen data?*

Lipton (2018) - *The Mythos of Model Interpretability*

Why do we need interpretability?

The **desiderata** of algorithmic models:

1. **Fairness**
    - *What biases does it contain? Does it discriminate against particular groups?*
2. **Trustworthiness**
    - *Models that are deployed carry a degree of responsibility, can we trust them?*
3. **Robustness**
    - *Does our model generalise robustly to unseen data?*
4. **Faithfulness**
    - *Is a model right for the right reasons?*

Lipton (2018) - *The Mythos of Model Interpretability*

Why do we nee

The **desiderata** of al

1. **Fairness**
   - *What biases doe* … *ainst particular groups?*
2. **Trustworthiness**
   - *Models that are* … *bility, can we trust them?*
3. **Robustness**
   - *Does our model*
4. **Faithfulness**
   - *Is a model right*

BBC

NEWS

Menu

Business | Market Data | New Tech Economy | Technology of Bu

Apple's 'sexist' credit card investigated by US regulator

11 November 2019

A US financial regulator has opened an investigation into claims Apple's credit card offered different credit limits for men and women.

It follows complaints - including from Apple's co-founder Steve Wozniak - that algorithms used to set limits might be inherently biased against women.

New York's Department of Financial Services (DFS) has contacted Goldman Sachs, which runs the Apple Card.

18) - *The Mythos of Model Interpretability*

Why do we nee

The **desiderata** of al

1. **Fairness**
   - *What biases doe* … *ainst particular groups?*
2. **Trustworthiness**
   - *Models that are* … *bility, can we trust them?*
3. **Robustness**
   - *Does our model*
4. **Faithfulness**
   - *How faithful are* … *oning?*

BBC

NEWS ☰ Menu

Tech

Facebook apology as AI labels black men 'primates'

6 September 2021

Facebook users who watched a newspaper video featuring black men were asked if they wanted to "keep seeing videos about primates" by an artificial-intelligence recommendation system.

Facebook told BBC News it "was clearly an unacceptable error", disabled the system and launched an investigation.

"We apologise to anyone who may have seen these offensive recommendations."

18) - *The Mythos of Model Interpretability*

Why do we nee

The **desiderata** of al

1. **Fairness**
   - *What biases doe* *ainst particular groups?*
2. **Trustworthiness**
   - *Models that are* *bility, can we trust them?*
3. **Robustness**
   - *Does our model*
4. **Faithfulness**
   - *How faithful are* *oning?*

503px × 920px

BBC

NEWS

Menu

Tech

Twitter finds racial bias in image-cropping AI

20 May 2021

Preferences for white people over black people and women over men were found in testing

Twitter's automatic cropping of images had underlying issues that favoured white individuals over black people, and women over men, the company said.

It comes months after its users highlighted potential problems with the algorithm, which cropped large photos.

The social network's follow-up research has now confirmed the problem

18) - *The Mythos of Model Interpretability*

The Guardian

Netherlands

Dutch government resigns over child benefits scandal

PM Mark Rutte will stay on in caretaker capacity until general elections scheduled for 17 March

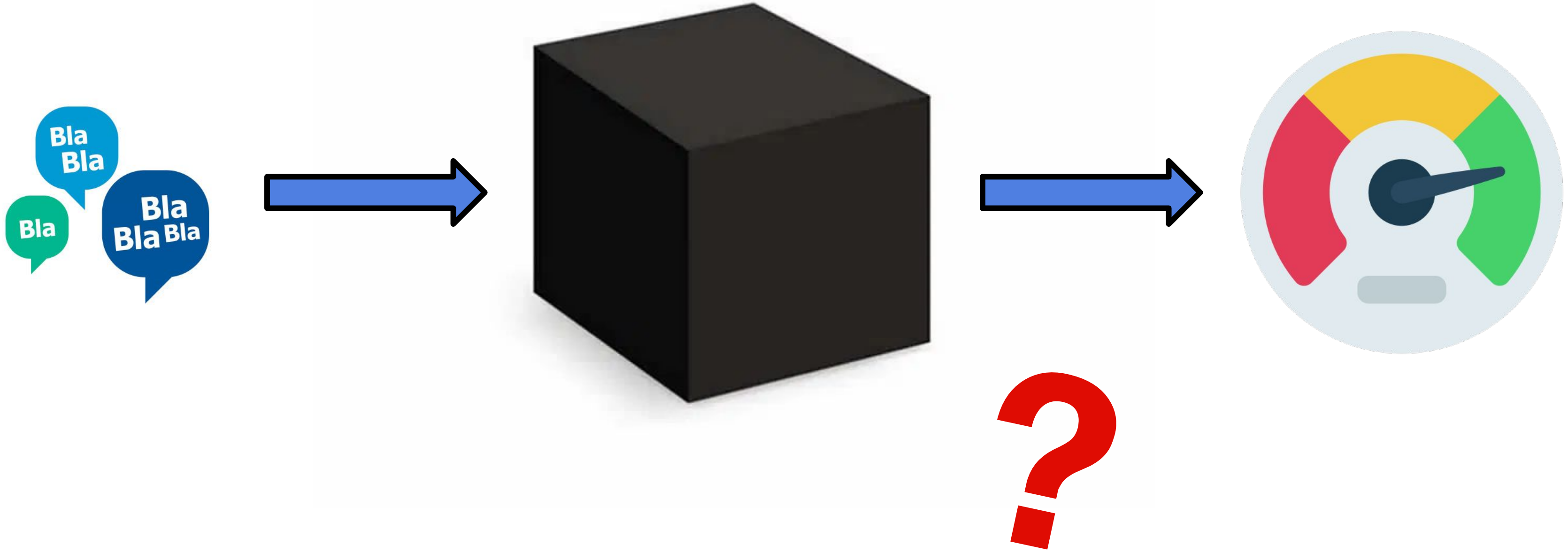
Jon Henley *Europe correspondent*
@jonhenley
Fri 15 Jan 2021 15.32 CET

Share

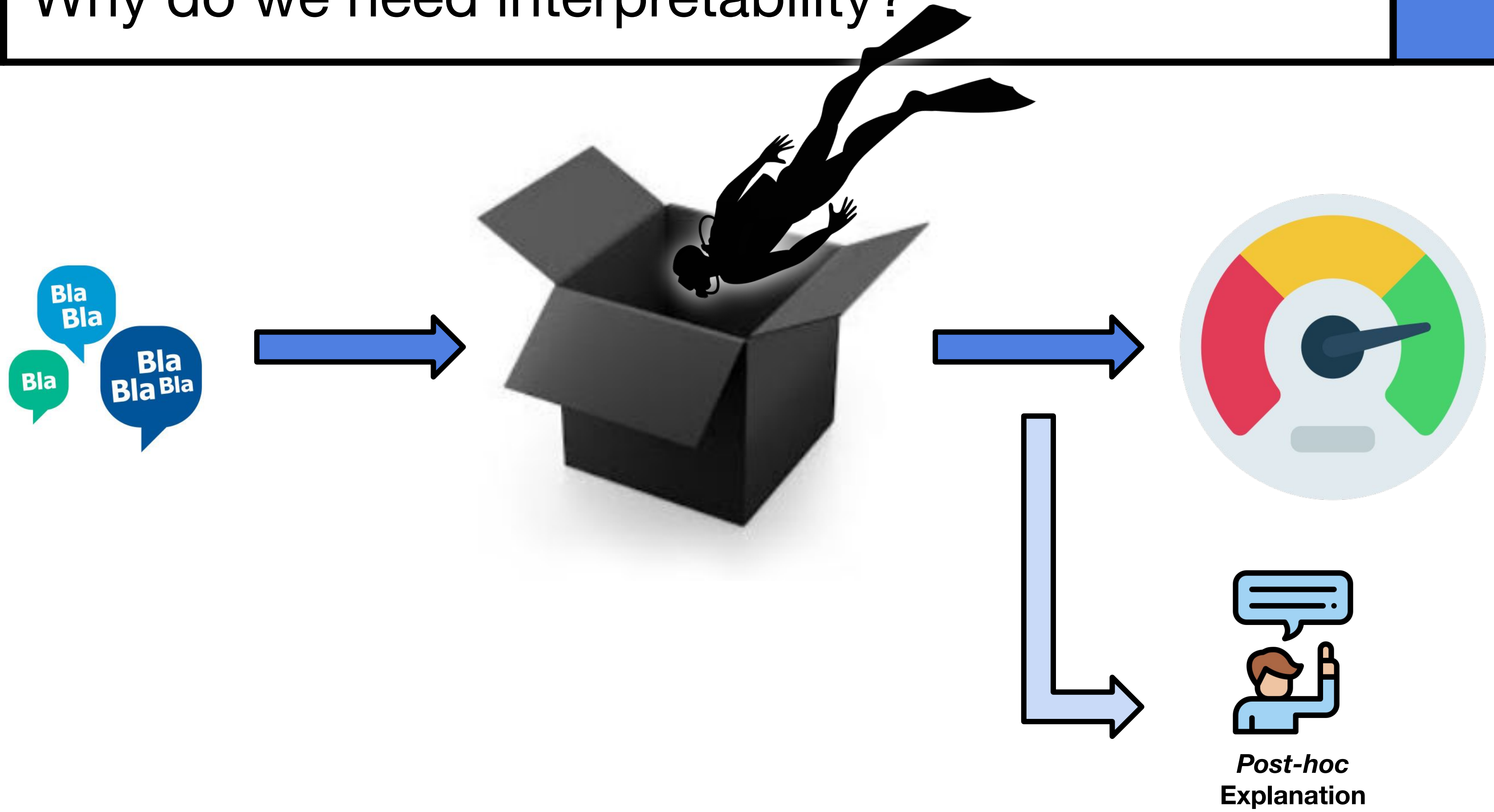
Mark Rutte appears at a press conference in The Hague after the resignation of the coalition. Photograph: Bart Maat/EPA

The Dutch government has resigned amid an escalating scandal over child benefits in which more than 20,000 families were wrongly accused of fraud by the tax authority.

Why do we need interpretability?



Why do we need interpretability?



Explanation Faithfulness

How do we ensure that an explanation **faithfully** represents a model's **reasoning**?

Plausibility does **not** imply faithfulness!

Models can be **right for the wrong reasons**!

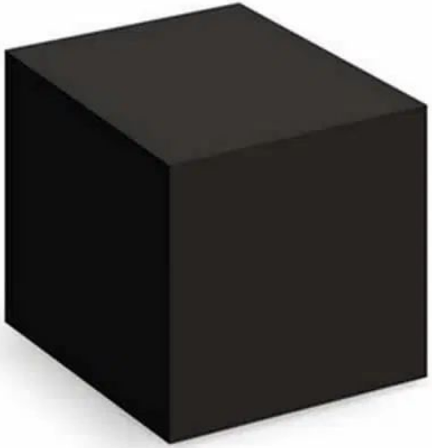
How do we ever know an explanation is truly **faithful** to the model?

Clever Hans

McCoy et al. (2019)

Explanation Faithfulness

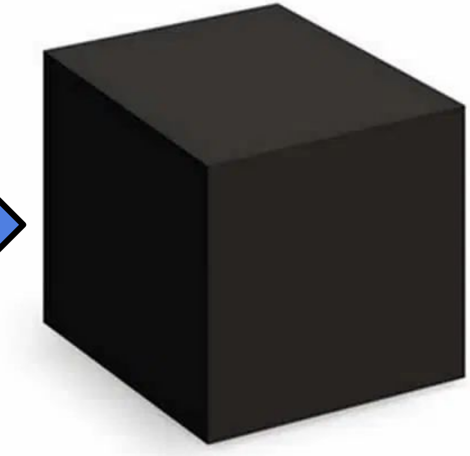
John is a 48 year old male lawyer from Amsterdam



€5000 credit

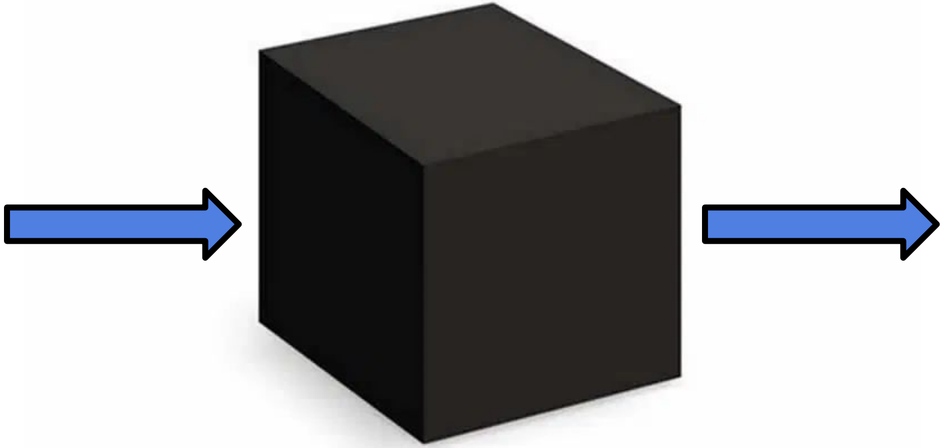
Explanation Faithfulness

Suzan is a 32 year old female doctor from Utrecht



Explanation Faithfulness

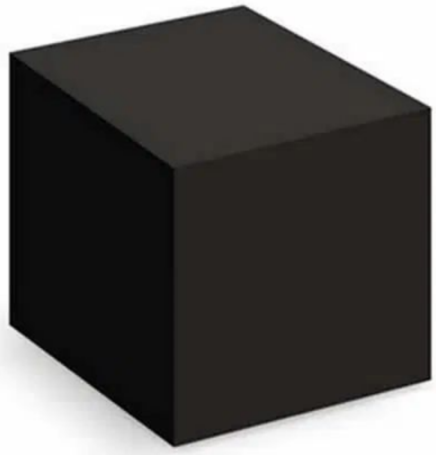
Suzan is a 32 year old female doctor from Utrecht



€1000 credit

Explanation Faithfulness

Suzan is a 32 year old female doctor from Utrecht

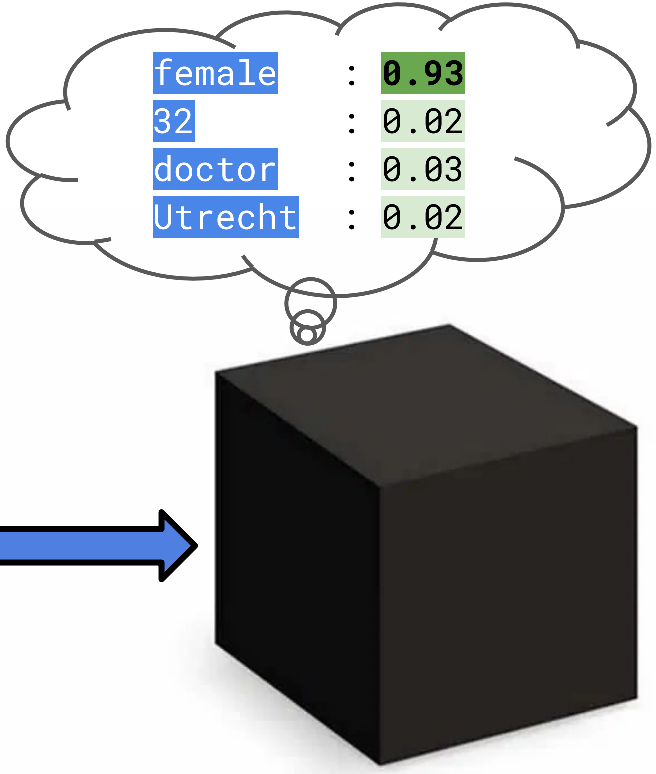


Why does Suzan get less than John? Because of her **age**? **Gender**? **Occupation**? **Location**?

€1000 credit

Explanation Faithfulness

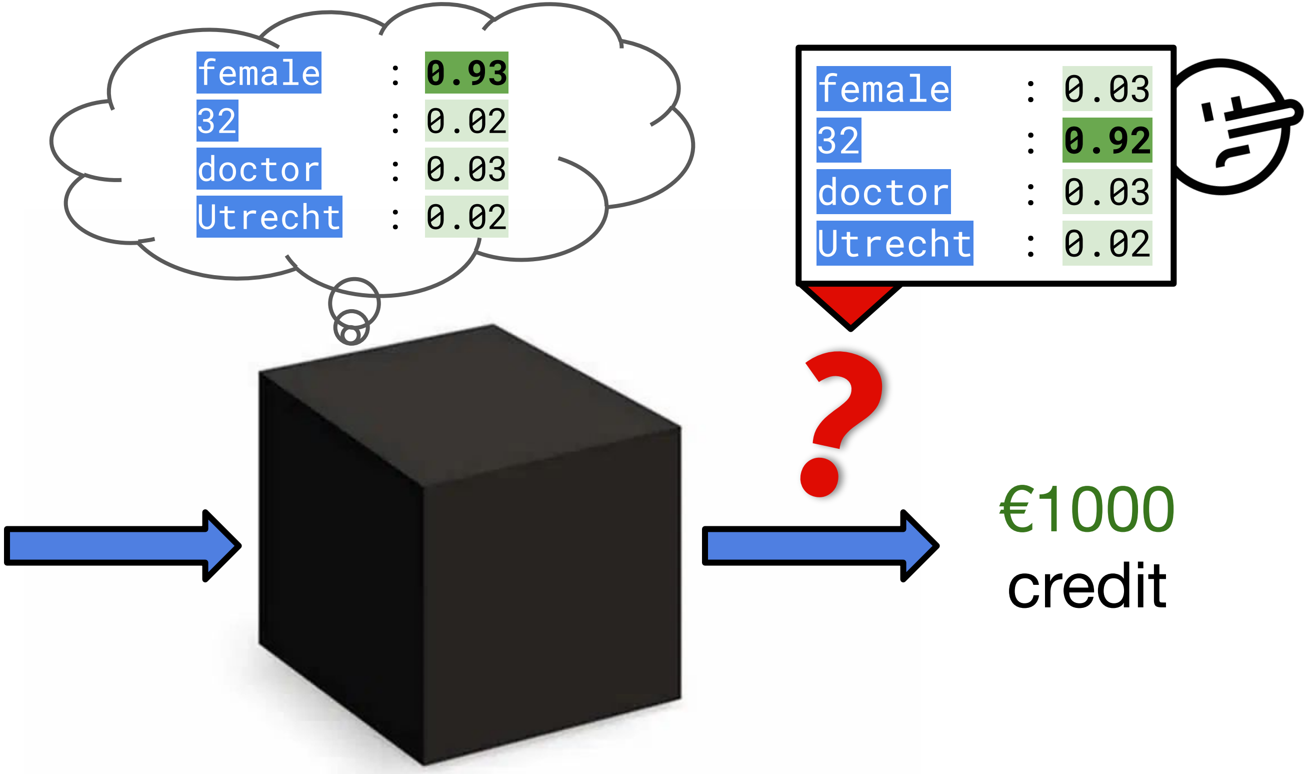
Suzan is a 32 year old female doctor from Utrecht



€1000 credit

Explanation Faithfulness

Suzan is a 32 year old female doctor from Utrecht



Marr's Tri-Level Hypothesis

Marr & Poggio (1976) - *From Understanding Computation to Understanding Neural Circuitry*

1. Computational

- What does the system do?
- What problems does it solve or overcome?

Marr's Tri-Level Hypothesis

Marr & Poggio (1976) - *From Understanding Computation to Understanding Neural Circuitry*

Explaining Complex Systems

Marr's Tri-Level Hypothesis

Marr & Poggio (1976) - *From Understanding Computation to Understanding Neural Circuitry*

1. Computational

- What does the system do?
- What problems does it solve or overcome?

2. Algorithmic

- How does the system do what it does?
- What representations does it use?

Explaining Complex Systems

Marr's Tri-Level Hypothesis

Marr & Poggio (1976) - *From Understanding Computation to Understanding Neural Circuitry*

1. Computational

- What does the system do?
- What problems does it solve or overcome?

2. Algorithmic

- How does the system do what it does?
- What representations does it use?

3. Implementational

- How is the system physically realised?
- What neural circuitry implements the system?

Explaining Neural Models

Levels of explanation *granularity*:

1. **Behavioural**
   - How does the model behave on certain phenomena?

Marr's Level

1. Computational

Explaining Neural Models

Levels of explanation *granularity*:

1. **Behavioural**
   - How does the model behave on certain phenomena?
2. **Attributional**
   - Which input features were most *important* for a prediction?

Marr's Level

1. Computational

2. Algorithmic

Explaining Neural Models

Levels of explanation *granularity*:

1. **Behavioural**
   - How does the model behave on certain phenomena?
2. **Attributional**
   - Which input features were most *important* for a prediction?
3. **Probing**
   - What *abstract features* are encoded by the model?

Marr's Level

1. Computational

2. Algorithmic

Explaining Neural Models

Levels of explanation *granularity*:

1. **Behavioural**
   - How does the model behave on certain phenomena?
2. **Attributional**
   - Which input features were most *important* for a prediction?
3. **Probing**
   - What *abstract features* are encoded by the model?
4. **Mechanistic**
   - Can we identify specific *circuits* responsible for a particular behaviour?

Marr's Level

1. Computational

2. Algorithmic

3. Implementational

Behavioural Interpretability

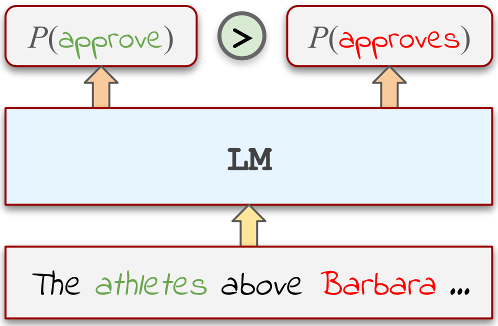
How can we understand a model better, without ‘opening the black box’?

- Using carefully crafted **minimal pairs** we can investigate a model’s performance on a specific phenomenon.

Behavioural Interpretability

How can we understand a model better, without 'opening the black box'?

- Using carefully crafted **minimal pairs** we can investigate a model's performance on a specific phenomenon.
- This type of experiment only requires access to the **output probabilities** of the model.



Behavioural Interpretability

- Assessing linguistic competence via minimal pairs:
  - **BLiMP** & **SyntaxGym**: Benchmark **suites** of different linguistic phenomena:

Phenomenon	N	Acceptable Example	Unacceptable Example
ANAPHOR AGR.	2	*Many girls insulted themselves.*	*Many girls insulted herself.*
ARG. STRUCTURE	9	*Rose wasn't disturbing Mark.*	*Rose wasn't boasting Mark.*
FILLER-GAP	7	*Brett knew what many waiters find.*	*Brett knew that many waiters find.*
IRREGULAR FORMS	2	*Aaron broke the unicycle.*	*Aaron broken the unicycle.*
ISLAND EFFECTS	8	*Which bikes is John fixing?*	*Which is John fixing bikes?*
NPI LICENSING	7	*The truck has clearly tipped over.*	*The truck has ever tipped over.*
QUANTIFIERS	4	*No boy knew fewer than six guys.*	*No boy knew at most six guys.*
SUBJECT-VERB AGR.	6	*These casseroles disgust Kayla.*	*These casseroles disgusts Kayla.*

Model	Overall	ANA. AGR	ARG. STR	BINDING	CTRL. RAIS.	D-N AGR	ELLIPSIS	FILLER. GAP	IRREGULAR	ISLAND	NPI	QUANTIFIERS	S-V AGR
5-gram	61.2	47.9	71.9	64.4	68.5	70.0	36.9	60.2	79.5	57.2	45.5	53.5	60.3
LSTM	69.8	91.7	73.2	73.5	67.0	85.4	67.6	73.9	89.1	46.6	51.7	64.5	80.1
TXL	69.6	94.1	72.2	74.7	71.5	83.0	77.2	66.6	78.2	48.4	55.2	69.3	76.0
GPT-2	83.0	99.3	81.8	80.9	81.9	95.8	89.3	81.3	91.9	72.7	76.8	79.0	86.4
Human	88.6	97.5	90.0	87.3	83.9	92.2	85.0	86.9	97.0	84.9	88.1	86.6	90.9

Warstadt et al. (2020)
Hu et al. (2020)

Limitations of Behavioural Tests

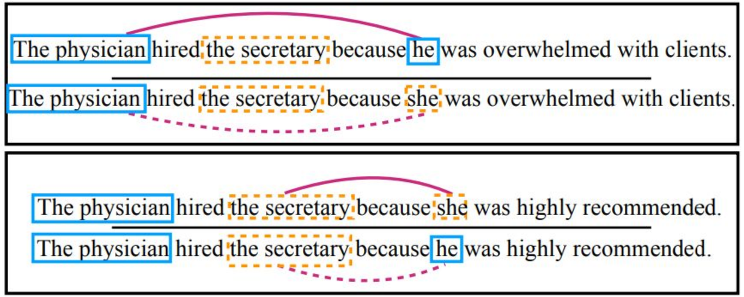
Behavioural tests show us a model's response to a particular input

- We now know roughly ***what*** a model can do.
- ***Why*** a model gave a particular response is not clear though!

Limitations of Behavioural Tests

Behavioural tests show us a model's response to a particular input

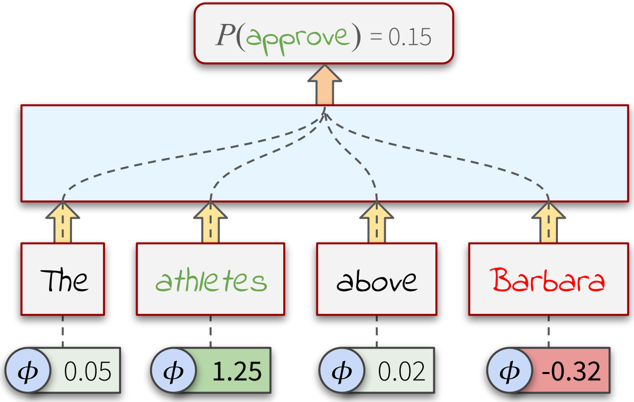
- We now know roughly ***what*** a model can do.
- ***Why*** a model gave a particular response is not clear though!
- Complex phenomena require more complex explanations
- E.g. *coreference resolution*:



Zhao et al. (2018), Jumelet et al. (2019)

Feature Attribution Methods

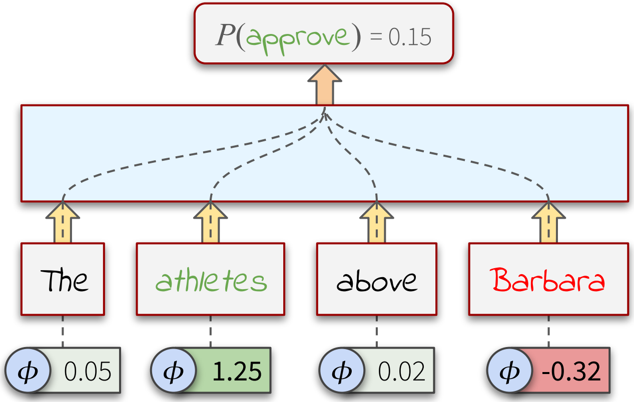
- **Feature attribution methods** explain model predictions in terms of the strongest *contributing* features.
- By normalizing such scores we get an insight into the **relative importance** of each feature.
- Shows us the **rationale** of a model behind a prediction → useful for uncovering biases!



Feature Attribution Methods

How do we compute the relative importance of a feature?

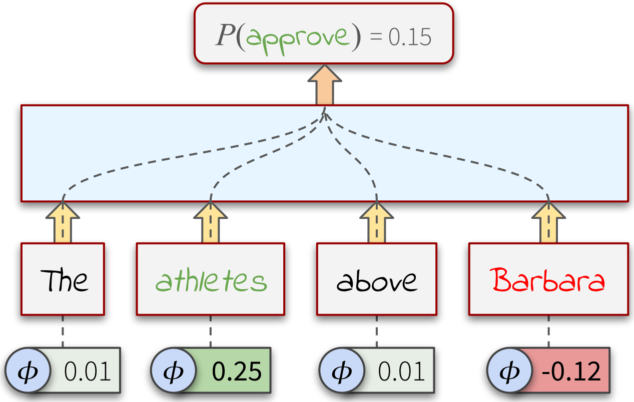
- Often this is done by **perturbing** parts of the input, and measuring the *change* in model output.



Feature Attribution Methods

How do we compute the relative importance of a feature?

- Often this is done by **perturbing** parts of the input, and measuring the *change* in model output.
- How should we perturb?
- How can we represent the *missingness* of a feature?
- How should we measure the change?



Baselines

- We often explain events by pointing out the most **important** factors
- This is often done in **contrast** to a neutral **baseline**:

Static Baseline

Model being explained

Features still present

$$v(\mathbf{x}_S) = f(\mathbf{x}_S \cup \mathbf{x}'_{\setminus S})$$

Value function for partial input

Removed features

x	= *This movie is not bad*
x’	= <unk>
\S	= *{not}*
$x_S \cup x'_{\setminus S}$	= *This movie is* <unk> *bad*

Baselines

- We often explain events by pointing out the most **important** factors
- This is often done in **contrast** to a neutral **baseline**:

Interventional Baseline

$$v(\mathbf{x}_S) = \mathbb{E}_{\mathbf{x}'_{\setminus S}}\left[f(\mathbf{x}_S \cup \mathbf{x}'_{\setminus S})\right]$$

Expectation over removed features

x = “This movie is not bad”

\S = {*not*}

$x_S \cup x'_{\setminus S}$ = “This movie is *the* / *is* / *walk* / *...* bad”

Baselines

- We often explain events by pointing out the most **important** factors
- This is often done in **contrast** to a neutral **baseline**:

Observational Baseline

Conditioned on present features

$$v(\mathbf{x}_S) = \mathbb{E}_{\mathbf{x}'_{\setminus S}}\left[f(\mathbf{x}_S \cup \mathbf{x}'_{\setminus S}) \mid \mathbf{x}_S\right]$$

Expectation over removed features

x = "This movie is not bad"

$\setminus S$ = {*not*}

$x_S \cup x'_{\setminus S}$ = "This movie is *very* / *that* / *quite* / *...* bad"

Baselines

- More targeted baselines allow for precise **counterfactual** explanations:

Input: *Can you stop the dog from*
Output: barking

1. Why did the model predict "barking"?
Can you stop the dog from

Importance of feature x: difference of output when removing x

$$S_E(x_i) = q(y_t|\boldsymbol{x}) - q(y_t|\boldsymbol{x}_{\neg i})$$

Baseline

Yin & Neubig (2022) - *Interpreting Language Models with Contrastive Explanations*

Baselines

- More targeted baselines allow for precise **counterfactual** explanations:

Input: *Can you stop the dog from*   **Output:** barking
1. Why did the model predict "barking"?   Can you stop the dog from
2. Why did the model predict "barking" *instead of* "crying"?   Can you stop the dog from
3. Why did the model predict "barking" *instead of* "walking"?   Can you stop the dog from

Importance of feature: difference of output when removed

$$S_E(x_i) = q(y_t|\boldsymbol{x}) - q(y_t|\boldsymbol{x}_{\neg i})$$

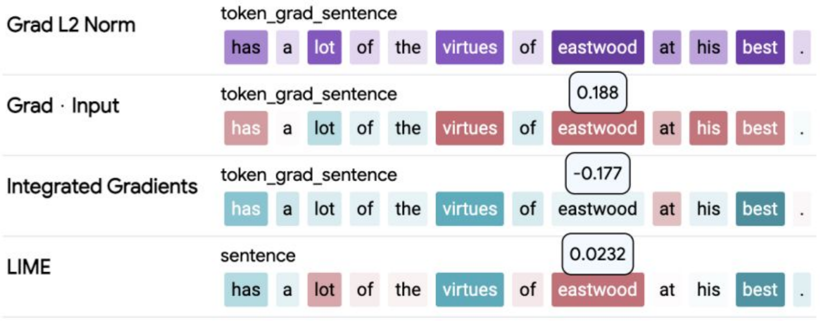
$$S_E^*(x_i) = (q(y_t|\boldsymbol{x}) - q(y_t|\boldsymbol{x}_{\neg i})) - (q(y_f|\boldsymbol{x}) - q(y_f|\boldsymbol{x}_{\neg i}))$$

Explanation with respect to ***foil***

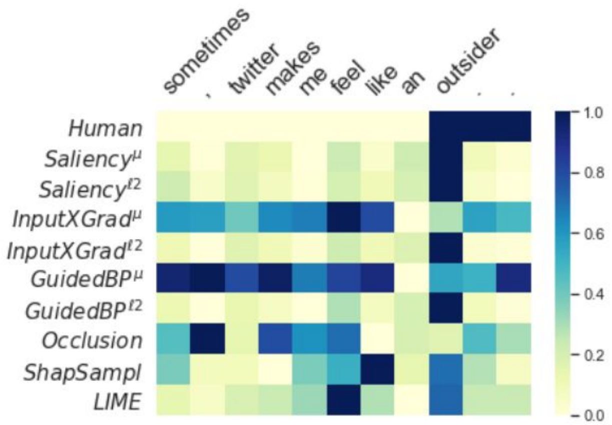
Yin & Neubig (2022) - *Interpreting Language Models with Contrastive Explanations*

Limitations of Feature Attributions

- Attribution methods **disagree** strongly
- Which explanation is the right one?
- Can we simplify model behaviour to a single explanation?



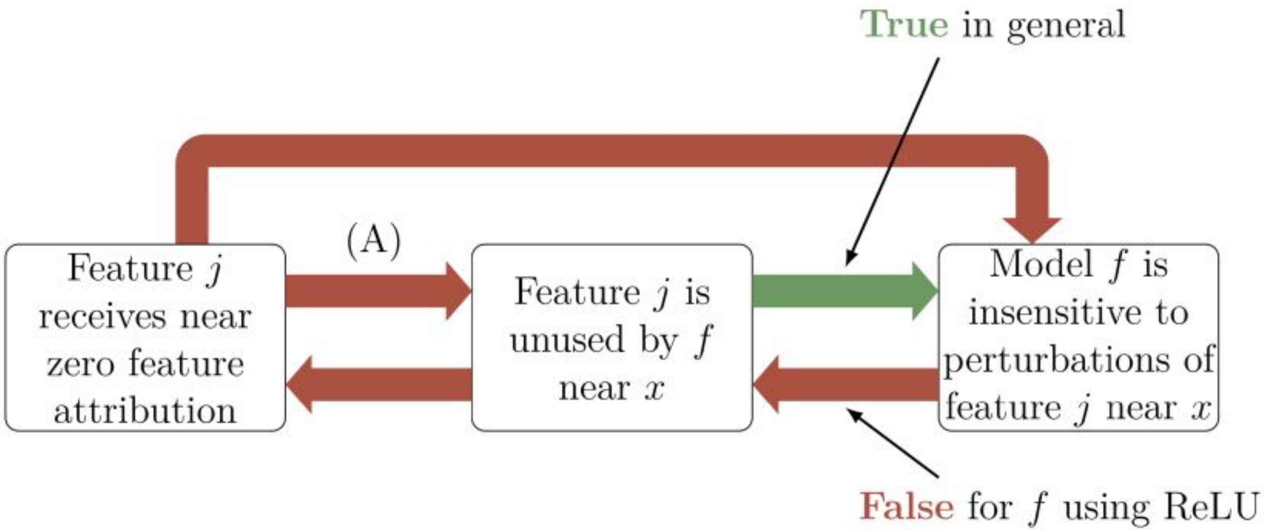
Bastings et al. (2022)



Atanasova et al. (2020)

Limitations of Feature Attributions

Bilodeau et al. (2024, PNAS): *Feature attributions can* ***provably fail*** *to improve on random guessing for inferring model behavior*



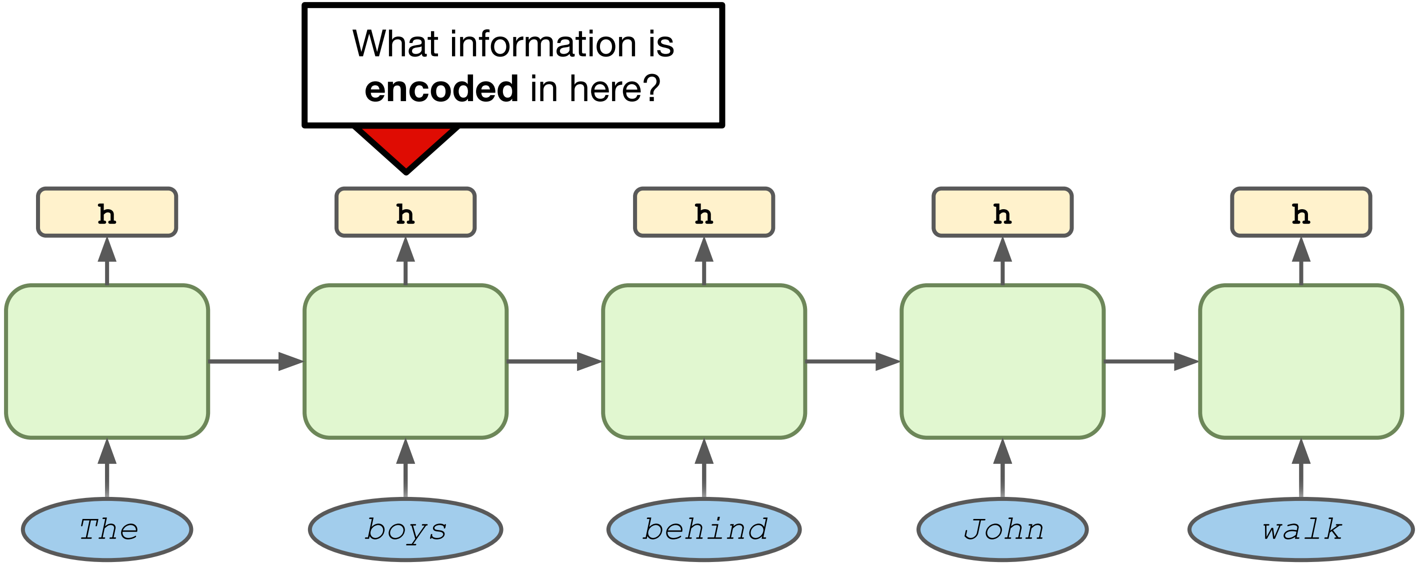
Probing

Feature attribution methods showed us which input features were important for a prediction.

- ✗ They do not show *how* in the model representations are formed
- ✗ They give no insight into **higher-level** concepts such as *‘gender’*, *‘number’*, or *‘part-of-speech’* class.

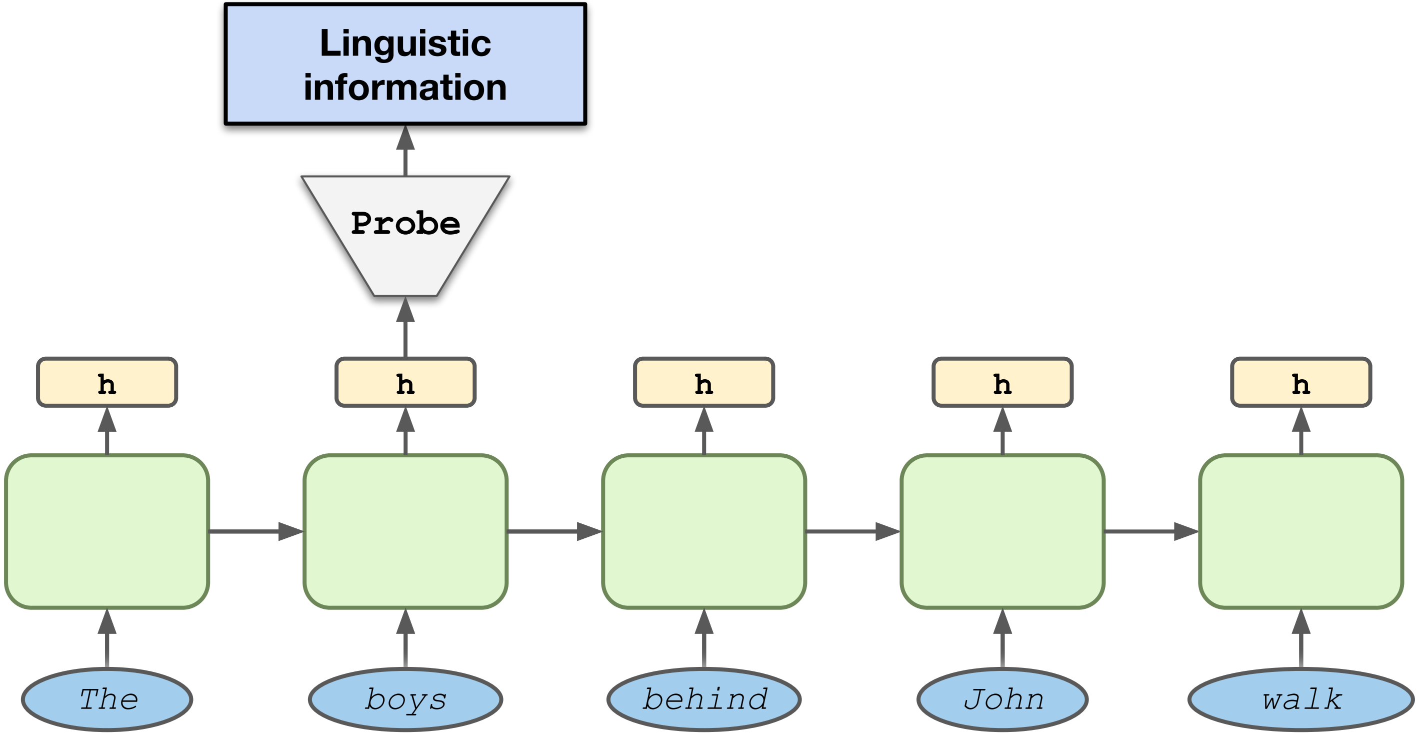
Instead, we can turn to **probing**, in which we train classifiers on top of model representations!

Probing



What information is **encoded** in here?
h
h
h
h
h
The
boys
behind
John
walk

Probing



Linguistic information
Probe
h
h
h
h
h
The
boys
behind
John
walk

Probing

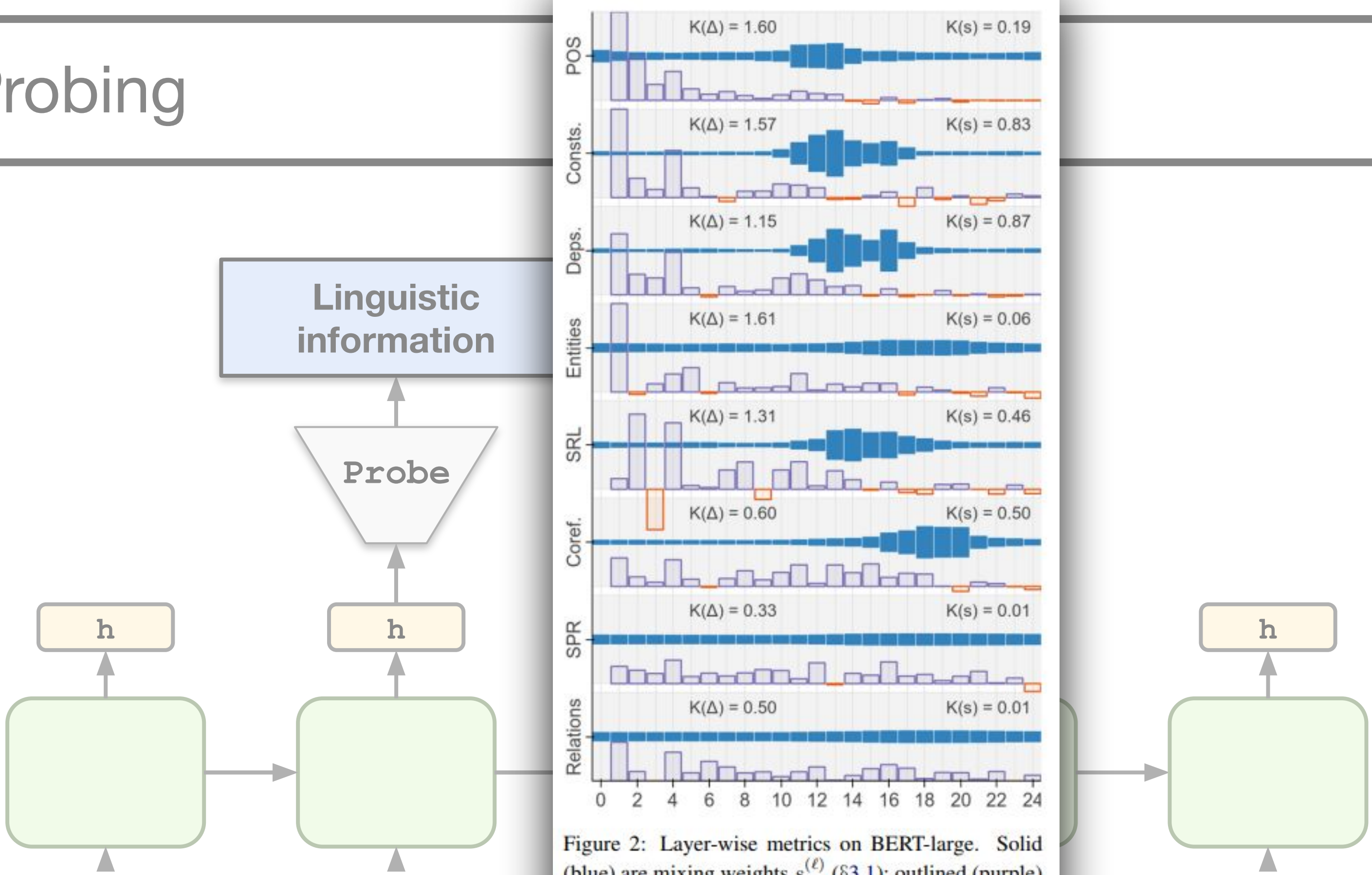


Figure 2: Layer-wise metrics on BERT-large. Solid (blue) are mixing weights $s_\tau^{(\ell)}$ (§3.1); outlined (purple) are differential scores $\Delta_\tau^{(\ell)}$ (§3.2), normalized for each task. Horizontal axis is encoder layer.

BERT Rediscovers the Classical NLP Pipeline

Tenney et al. (2019)

Limitations of Probing

Probing shows us whether abstract concepts are decodable.

✗ It does not show us whether the model actually *uses* these concepts for its predictions

For this we need a **causal** methodology.

Measure the impact on model performance after **removing** a concept from the representation.

Beyond Probing

Probe Ranking

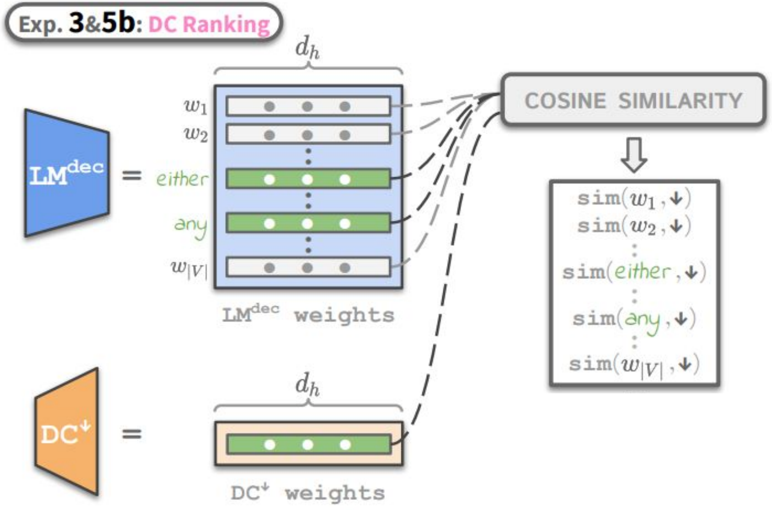
*Opens up a **downward monotone** environment in which **Negative Polarity Items** can occur*

He doesn't like fish ***either*** vs. **He does like fish* ***either***

Jumelet et al. (2021) - *Language Models Use Monotonicity to Assess NPI Licensing*

Beyond Probing

Probe Ranking

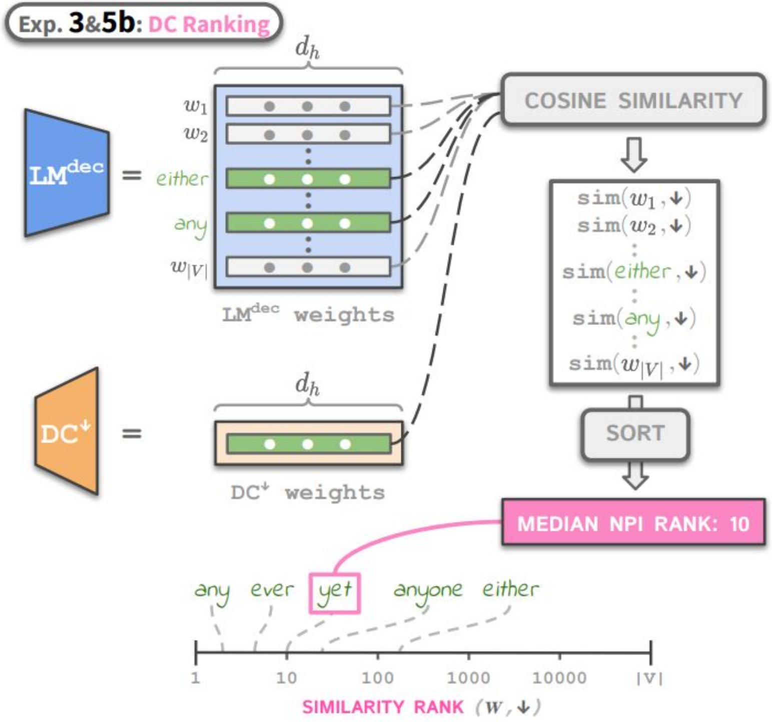


*He does**n't** like fish* ***either*** vs. **He does like fish* ***either***

Jumelet et al. (2021) - *Language Models Use Monotonicity to Assess NPI Licensing*

Beyond Probing

Probe Ranking



Jumelet et al. (2021) - *Language Models Use Monotonicity to Assess NPI Licensing*

Beyond Probing

Amnesic Probing

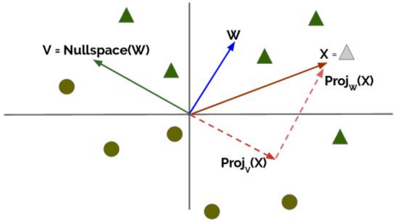
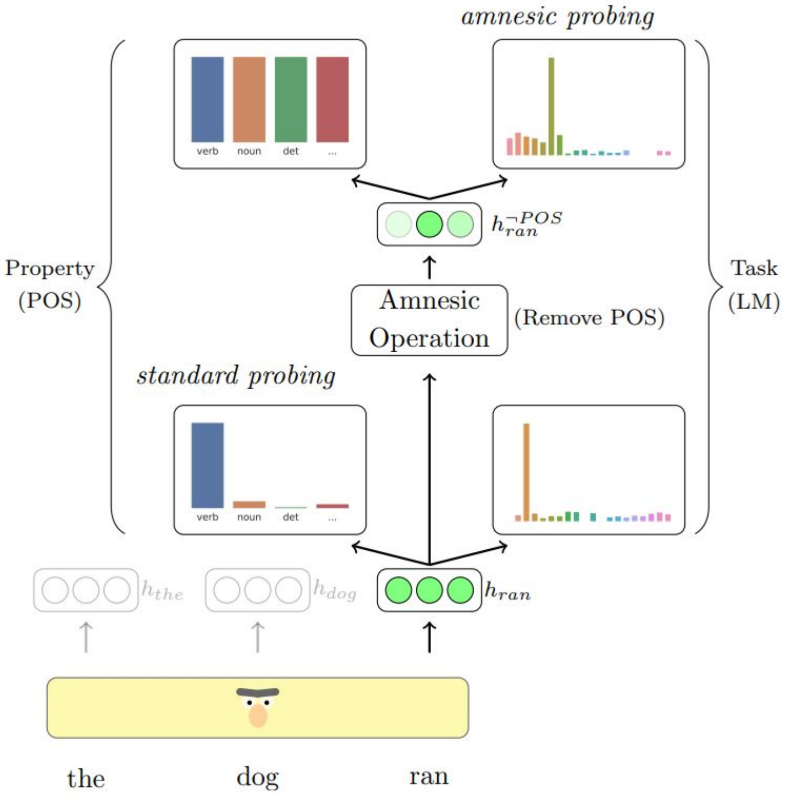


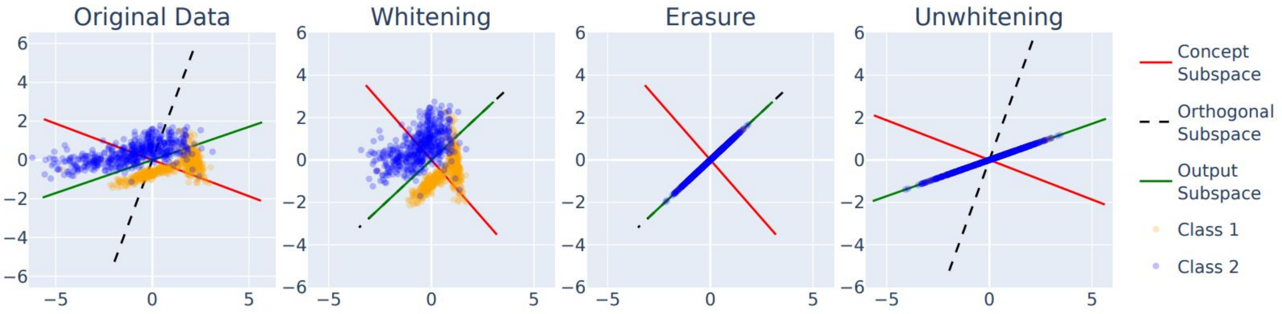
Figure 2: Nullspace projection for a 2-dimensional binary classifier. The decision boundary of W is W 's null-space.



Elazar et al. (2020)

Beyond Probing

LEACE



Belrose et al. (2023)

CausalGym

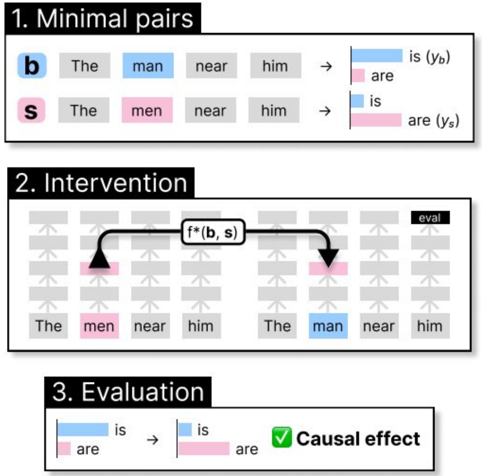


Figure 1: The **CausalGym** pipeline: **(1)** take an input minimal pair (b, s) exhibiting a linguistic alternation that affects next-token predictions (y_b, y_s); **(2)** intervene on the base forward pass using a pre-defined intervention function that operates on aligned representations from both inputs; **(3)** check how this intervention affected the next-token prediction probabilities. In aggregate, such interventions assess the causal role of the intervened representation on the model's behaviour.

Arora et al. (2024)

Recap

- The huge size of current NLP models has made us lose **transparency**
- Interpretability is **vital** for gaining trust in black-box models
- Interpretability is also vital for understanding the **linguistic capacities** of NLP models
- We can explain a model at increasing levels of granularity
  - Behavioural tests
  - Feature attributions
  - Probing
- However, each of these levels come with certain **limitations** regarding **faithfulness**
- Next: zooming in on Transformer components for more **faithful** explanations