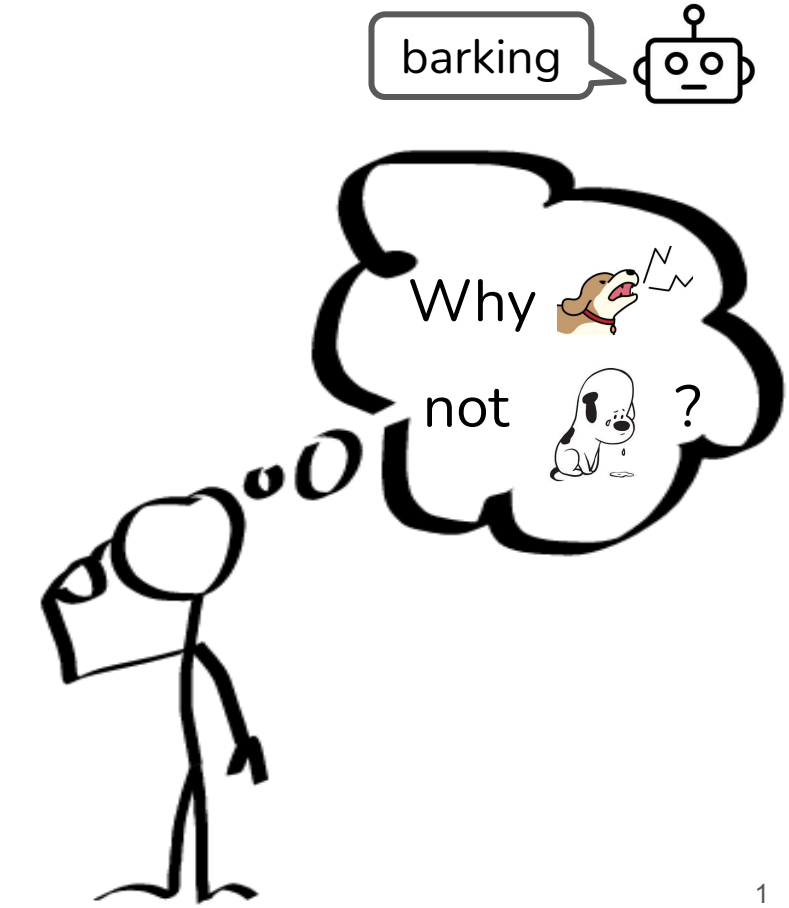
Carnegie Mellon University
Language Technologies Institute

Berkeley
N L P

Interpreting Language Models with Contrastive Explanations

Kayo Yin and Graham Neubig



EMNLP 2022

Interpreting Language Models with Contrastive Explanations

Kayo Yin

Graham Neubig

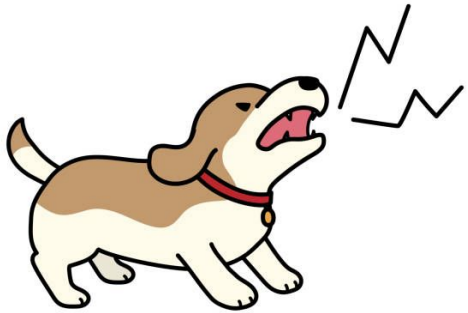
Berkeley NLP

Carnegie Mellon University
Language Technologies Institute

Why did the LM predict “barking”?

Input: Can you stop the dog from

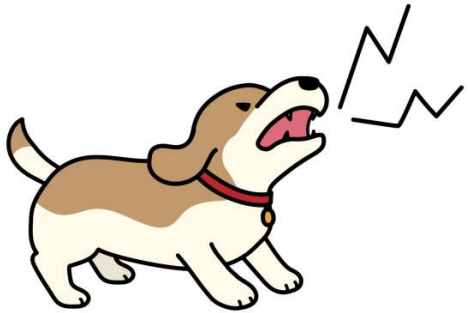
Output: barking



Why did the LM predict “barking”?

Input: Can you stop the dog from

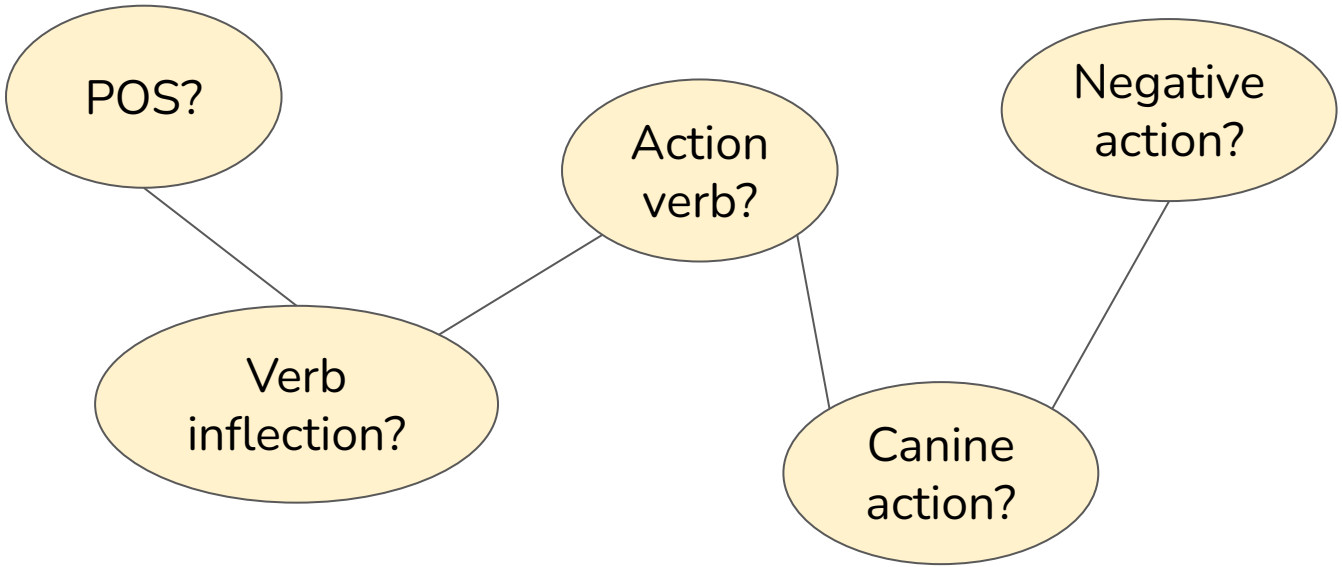
Output: barking



Language modeling is complex

Input: Can you stop the dog from

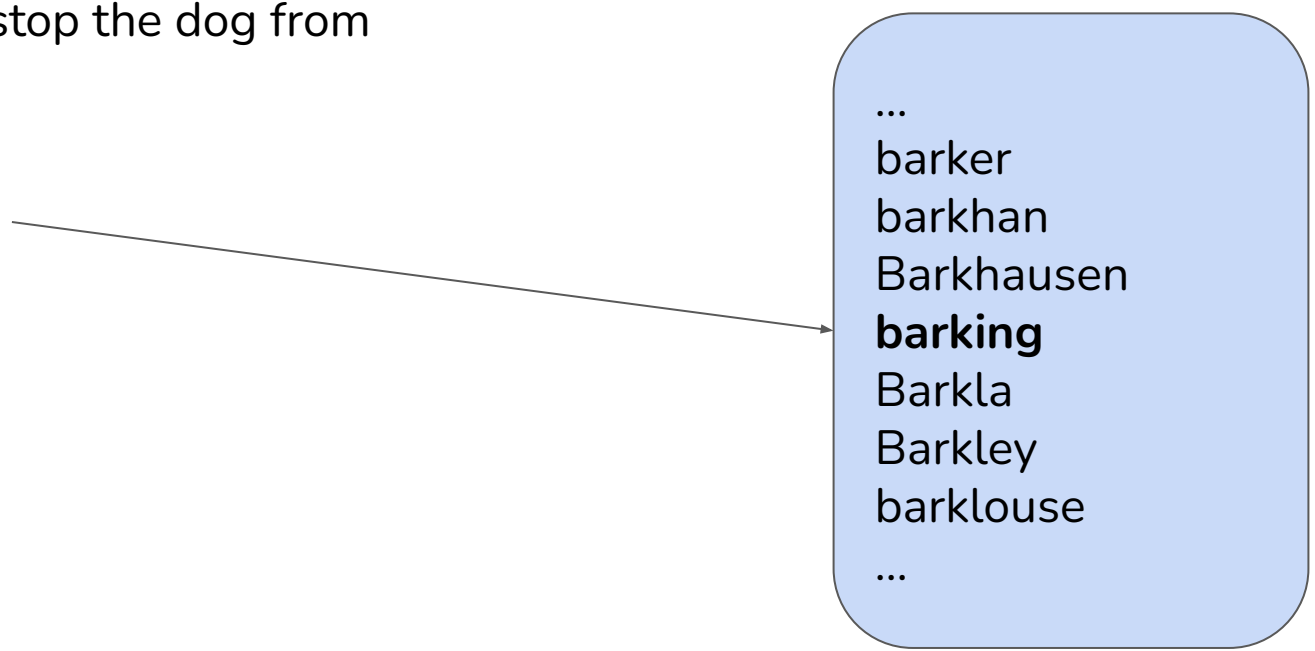
Output: barking



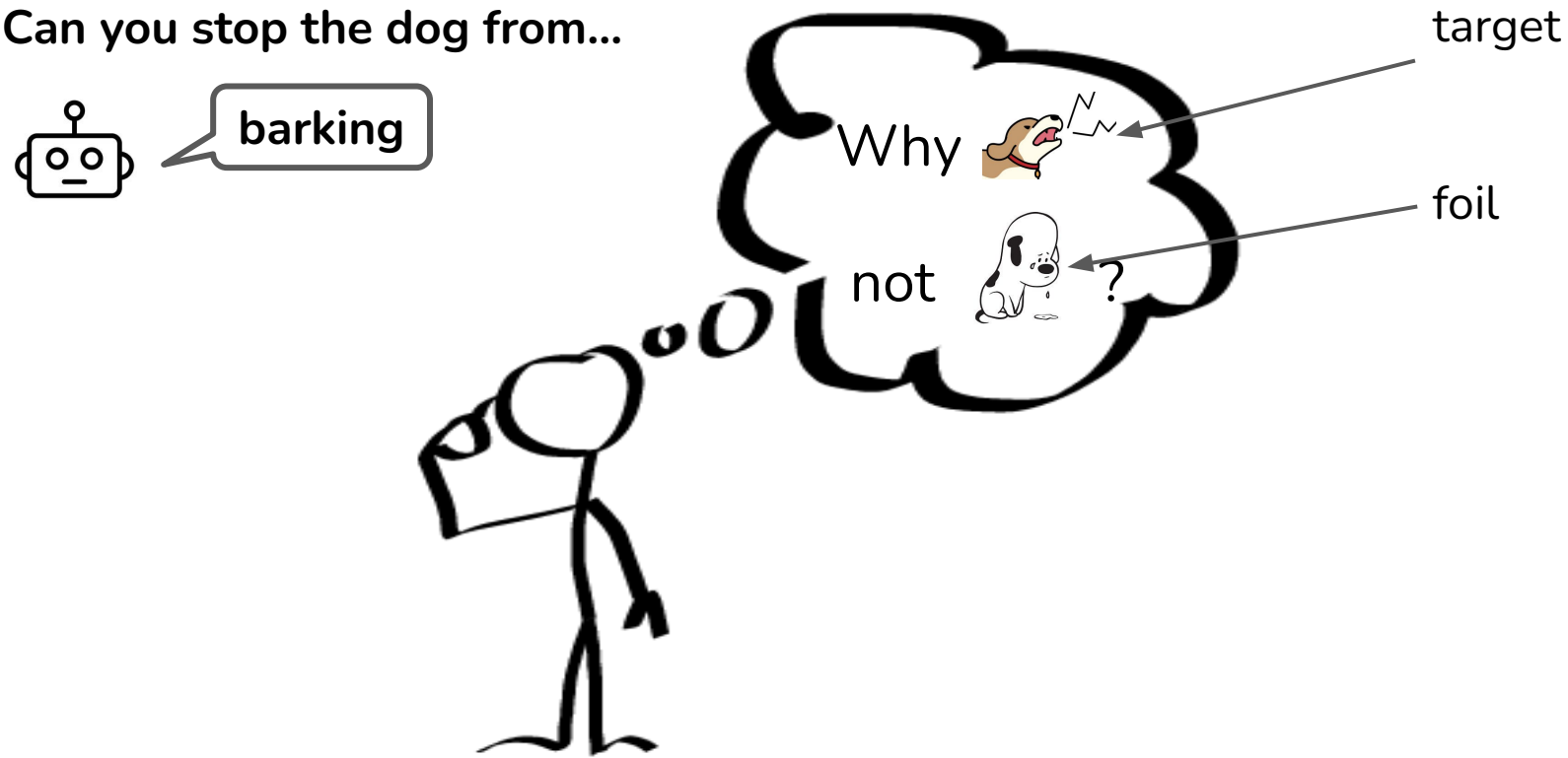
Language modeling is complex

Input: Can you stop the dog from

Output: barking



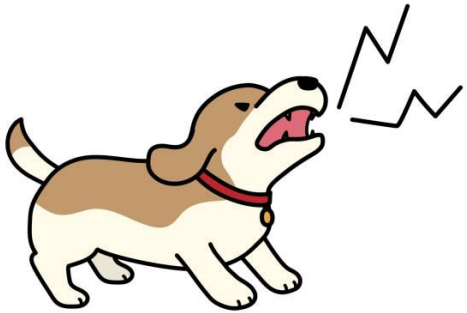
Contrastive explanations are more intuitive



Why did the LM predict “barking” instead of “crying”?

Input: Can you stop the dog from

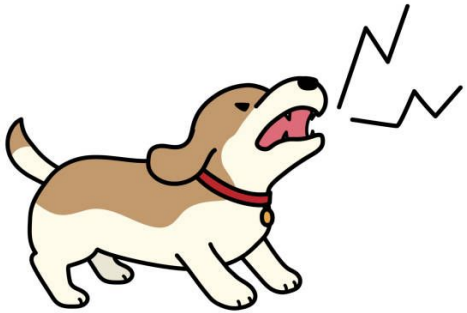
Output: barking



Why did the LM predict “barking” instead of “crying”?

Input: Can you stop the dog from

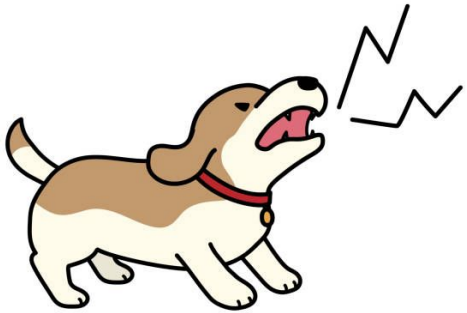
Output: barking



Why did the LM predict “barking” instead of “walking”?

Input: Can you stop the dog from

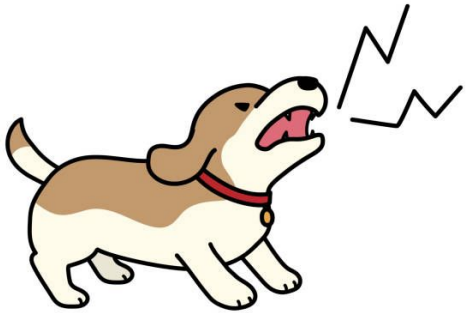
Output: barking



Why did the LM predict “barking” instead of “walking”?

Input: Can you stop the dog from

Output: barking

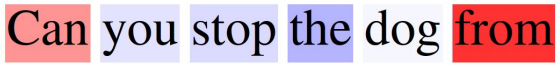


Contrastive explanations for language models

- Gradient $g(x_i) = \nabla_{x_i} q(y_t|\boldsymbol{x})$
- Input x gradient $g(x_i) \cdot x_i$

Contrastive explanations for language models

- Gradient $g(x_i) = \nabla_{x_i} q(y_t|\boldsymbol{x})$
- Input x gradient $g(x_i) \cdot x_i$



Contrastive explanations for language models

- Gradient $g(x_i) = \nabla_{x_i} q(y_t|\boldsymbol{x})$
- Input x gradient $g(x_i) \cdot x_i$

Can you stop the dog from

- Contrastive gradient $g^*(x_i) = \nabla_{x_i} \left(q(y_t|\boldsymbol{x}) - q(y_f|\boldsymbol{x})\right)$

Contrastive explanations for language models

- Gradient $g(x_i) = \nabla_{x_i} q(y_t|\boldsymbol{x})$
- Input x gradient $g(x_i) \cdot x_i$

Can you stop the dog from

- Contrastive gradient $g^*(x_i) = \nabla_{x_i} \left(q(y_t|\boldsymbol{x}) - q(y_f|\boldsymbol{x})\right)$
- Contrastive input x gradient $g^*(x_i) \cdot x_i$

Contrastive explanations for language models

- Gradient $g(x_i) = \nabla_{x_i} q(y_t|\boldsymbol{x})$
- Input x gradient $g(x_i) \cdot x_i$

Can you stop the dog from

- Contrastive gradient $g^*(x_i) = \nabla_{x_i} \left(q(y_t|\boldsymbol{x}) - q(y_f|\boldsymbol{x})\right)$
- Contrastive input x gradient $g^*(x_i) \cdot x_i$

Can you stop the dog from

Contrastive explanations for language models

- Contrastive input x gradient $g(x_i) \cdot x_i \longrightarrow$
- Contrastive gradient norm $||g(x_i)||_{L1} \longrightarrow$
- Contrastive erasure

$q(y_t|\boldsymbol{x}) - q(y_t|\boldsymbol{x}_{\neg i}) \longrightarrow$

Contrastive explanations for language models

- Contrastive input x gradient $g(x_i) \cdot x_i \longrightarrow g^*(x_i) \cdot x_i$
- Contrastive gradient norm $||g(x_i)||_{L1} \longrightarrow$
- Contrastive erasure

$q(y_t|\boldsymbol{x}) - q(y_t|\boldsymbol{x}_{\neg i}) \longrightarrow$

Contrastive explanations for language models

- Contrastive input x gradient $g(x_i) \cdot x_i \longrightarrow g^*(x_i) \cdot x_i$
- Contrastive gradient norm $||g(x_i)||_{L1} \longrightarrow ||g^*(x_i)||_{L1}$
- Contrastive erasure

$q(y_t|\boldsymbol{x}) - q(y_t|\boldsymbol{x}_{\neg i}) \longrightarrow$

Contrastive explanations for language models

- Contrastive input x gradient $g(x_i) \cdot x_i \longrightarrow g^*(x_i) \cdot x_i$
- Contrastive gradient norm $||g(x_i)||_{L1} \longrightarrow ||g^*(x_i)||_{L1}$
- Contrastive erasure

$$q(y_t|\boldsymbol{x}) - q(y_t|\boldsymbol{x}_{\neg i}) \longrightarrow (q(y_t|\boldsymbol{x}) - q(y_t|\boldsymbol{x}_{\neg i})) - (q(y_f|\boldsymbol{x}) - q(y_f|\boldsymbol{x}_{\neg i}))$$

Contrastive explanations for language models

- Contrastive input x gradient $g(x_i) \cdot x_i \longrightarrow g^*(x_i) \cdot x_i$
- Contrastive gradient norm $||g(x_i)||_{L1} \longrightarrow ||g^*(x_i)||_{L1}$
- Contrastive erasure

$$q(y_t|\boldsymbol{x}) - q(y_t|\boldsymbol{x}_{\neg i}) \longrightarrow (q(y_t|\boldsymbol{x}) - q(y_t|\boldsymbol{x}_{\neg i})) - (q(y_f|\boldsymbol{x}) - q(y_f|\boldsymbol{x}_{\neg i}))$$

```
(output.logits[-1][target]).backward()

(output.logits[-1][target]-output.logits[-1][foil]).backward()
```

Result 1:
Contrastive explanations can better identify
linguistically appropriate evidence
for LM decisions

Defining evidence for grammatical acceptability

BLiMP dataset (Warstadt et al., 2020): minimal pairs of grammatical acceptability

Many teenagers were helping themselves.
Many teenagers were helping themself.

Anaphor number agreement

Defining evidence for grammatical acceptability

- *Anaphor agreement*: extract input tokens **coreferent** with the target

Many **teenagers** were helping themselves.
Many **teenagers** were helping themself.

Defining evidence for grammatical acceptability

- *Anaphor agreement*: extract input tokens **coreferent** with the target
- *Argument structure*: extract the **main verb**

Amanda was **respected** by some waitresses.
Amanda was **respected** by some picture.

Defining evidence for grammatical acceptability

- *Anaphor agreement*: extract input tokens **coreferent** with the target
- *Argument structure*: extract the **main verb**
- *Determiner-noun agreement*: extract the **determiner** of the target noun

Phillip was lifting **this** mouse.
Phillip was lifting **this** mice.

Defining evidence for grammatical acceptability

- *Anaphor agreement*: extract input tokens **coreferent** with the target
- *Argument structure*: extract the **main verb**
- *Determiner-noun agreement*: extract the **determiner** of the target noun
- *NPI licensing*: extract the **NPI**

Even these trucks have often slowed.
Even these trucks have ever slowed.

Defining evidence for grammatical acceptability

- *Anaphor agreement*: extract input tokens **coreferent** with the target
- *Argument structure*: extract the **main verb**
- *Determiner-noun agreement*: extract the **determiner** of the target noun
- *NPI licensing*: extract the **NPI**
- *Subject-verb agreement*: extract **subject** of the target verb

A **sketch of lights** doesn’t appear.
A **sketch of lights** don’t appear.

Quantifying alignment between linguistic rules and explanations

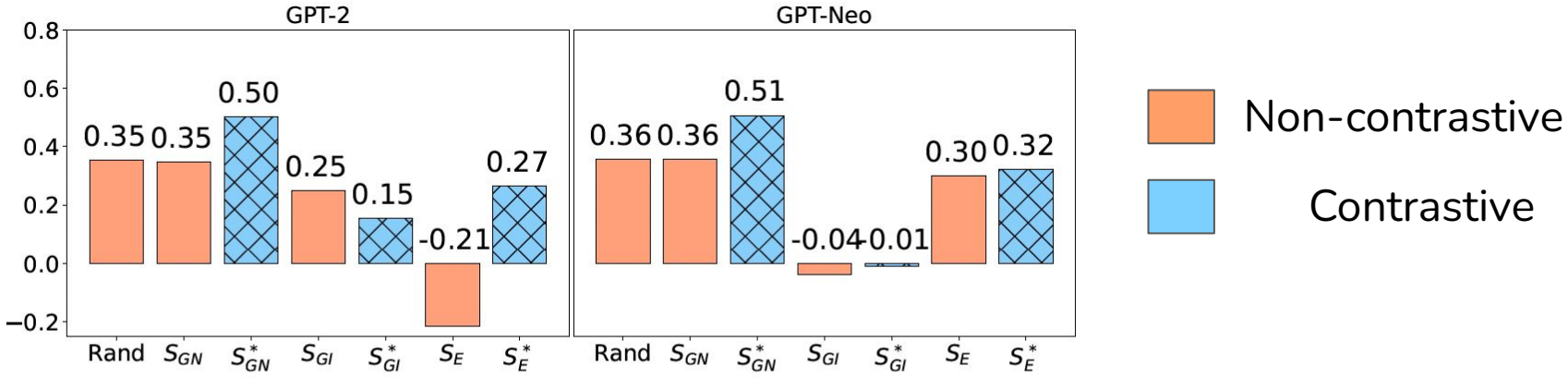
Dot product

Quantifying alignment between linguistic rules and explanations

Dot product

More alignment metrics in paper:
Probes needed
Mean reciprocal rank

Contrastive explanations can better identify linguistically appropriate evidence for LM decisions



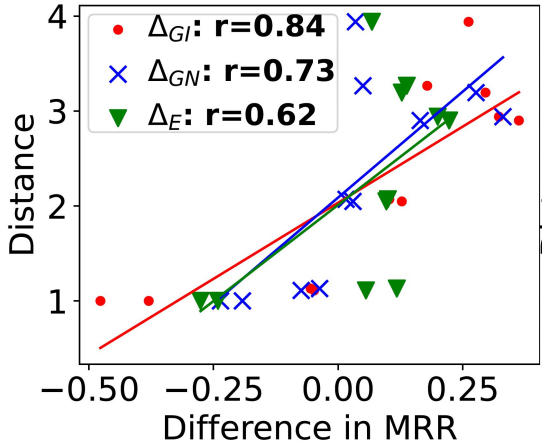
Contrastive explanations have better alignment than non-contrastive explanations

Contrastive explanations can better identify linguistically appropriate evidence for LM decisions

	Correct			Incorrect		
	DP (↑)	PN (↓)	MRR (↑)	DP (↑)	PN (↓)	MRR (↑)
Rand	0.34	1.66	0.57	0.27	2.05	0.50
S_{GN}	0.36	1.45	0.58	0.37	**1.60**	0.56
S^*_{GN}	**0.50**	**1.33**	**0.61**	**0.48**	1.71	**0.57**
S_{GI}	0.26	1.44	0.59	**0.24**	1.72	0.55
S^*_{GI}	**0.36**	**1.25**	**0.64**	-0.05	**1.27**	**0.64**
S_E	-0.51	1.34	0.64	**0.44**	**1.30**	0.55
S^*_E	**0.29**	**1.13**	**0.68**	0.18	1.71	0.55

Contrastive explanations have better alignment than non-contrastive explanations, especially when
-> the model predicts **correctly**

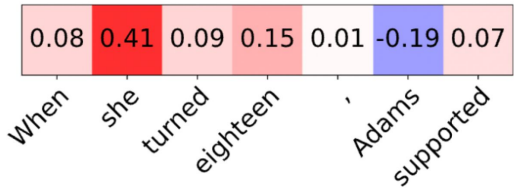
Contrastive explanations can better identify linguistically appropriate evidence for LM decisions



Contrastive explanations have better alignment than non-contrastive explanations, especially when
-> the model predicts **correctly**
-> the target and appropriate evidence are **distant**

Result 2:
Contrastive explanations improve model simulatability

Model simulatability test



Which token did the model more likely predict?

- herself
- himself

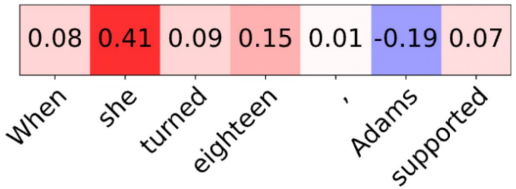
Was the explanation useful in making your decision?

- Yes
- No

Correct!

Model simulatability test

- 10 participants
  - ML graduate students



Which token did the model more likely predict?

- herself
- himself

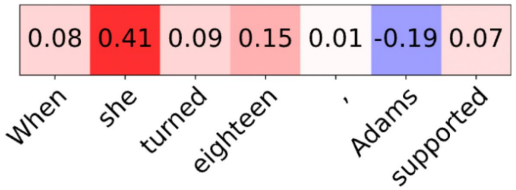
Was the explanation useful in making your decision?

- Yes
- No

Correct!

Model simulatability test

- 10 participants
  - ML graduate students
- 20 pairs of highly confusable words
  - Corpus-driven confusion metric



Which token did the model more likely predict?

- herself
- himself

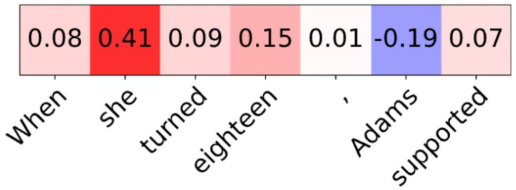
Was the explanation useful in making your decision?

- Yes
- No

Correct!

Model simulatability test

- 10 participants
  - ML graduate students
- 20 pairs of highly confusable words
  - Corpus-driven confusion metric
- Balanced data



Which token did the model more likely predict?

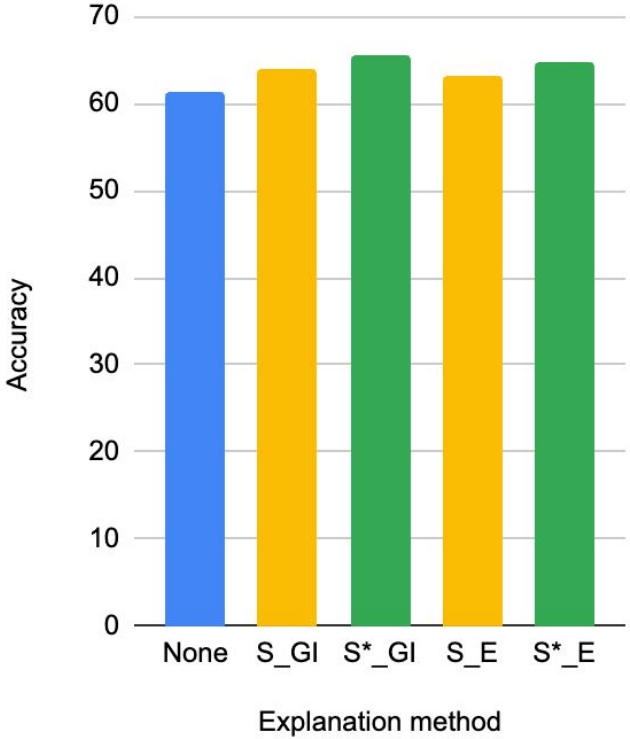
- herself
- himself

Was the explanation useful in making your decision?

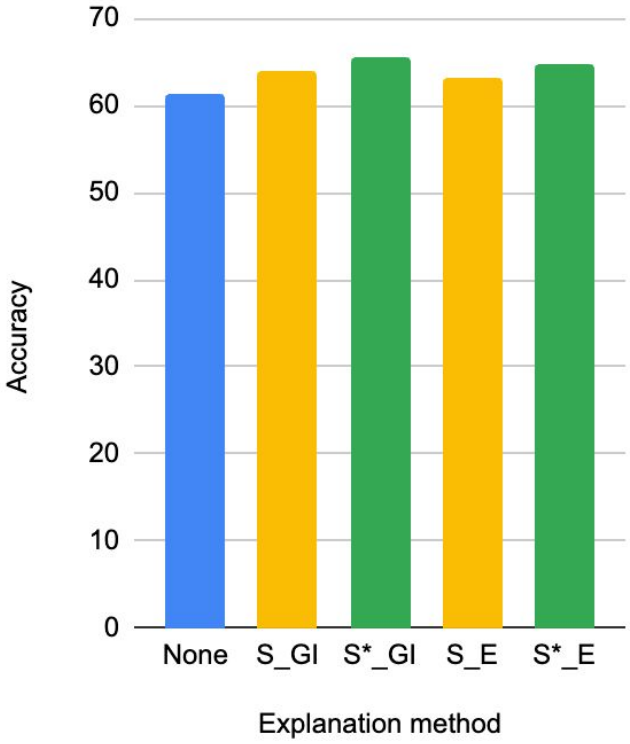
- Yes
- No

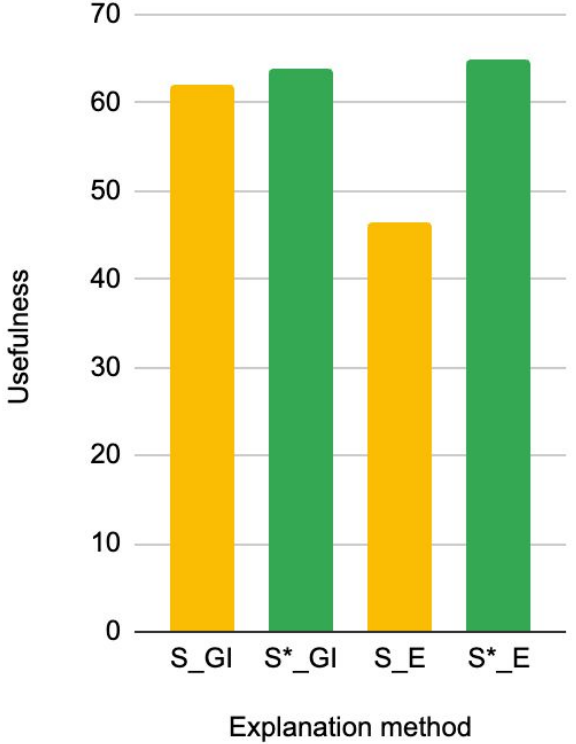
Correct!

Contrastive explanations improve model simulatability



Contrastive explanations improve model simulatability





Result 3:
Contrastive explanations help us characterize how LMs make decisions

What context do LMs use for certain decisions?

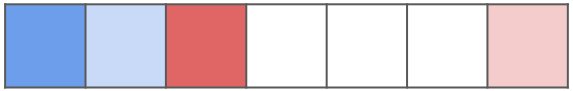
Hypothesis: linguistically similar decisions have similar contrastive explanations

Clustering foils by explanation

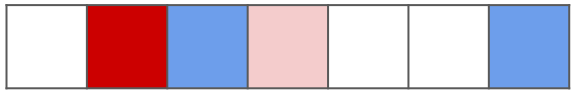
Input: General relativity predicts the existence of

Output: black

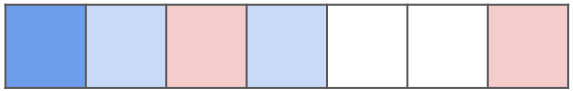
red



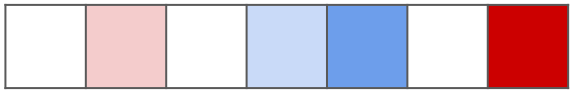
English



yellow



Asian

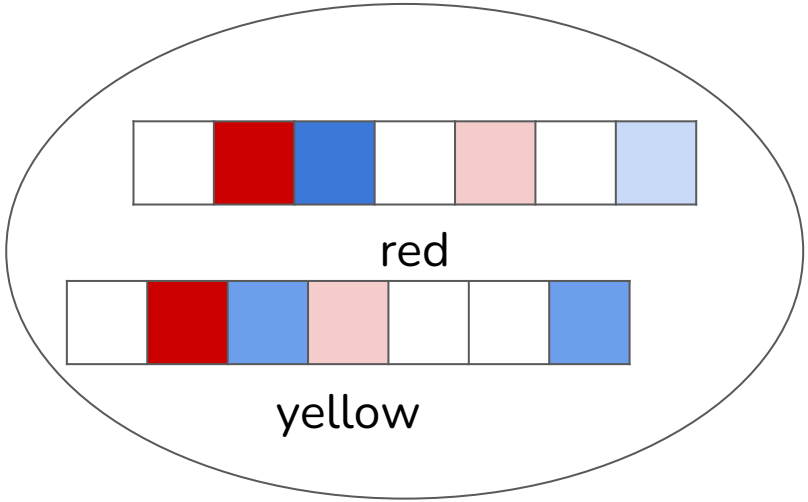


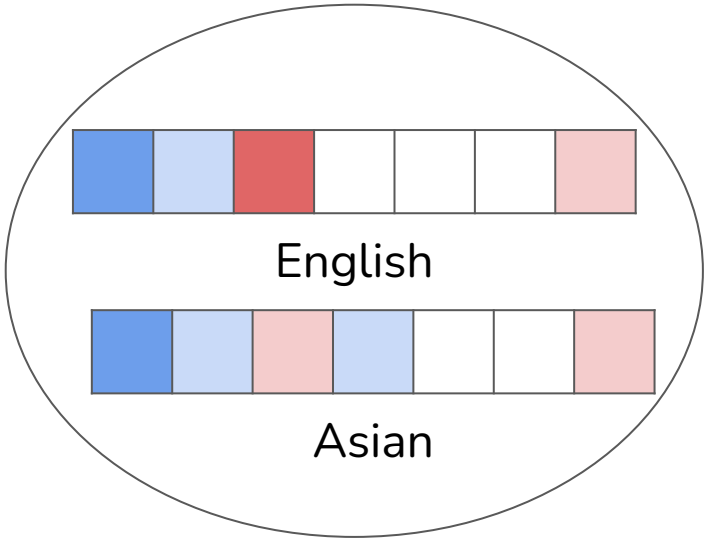
dog

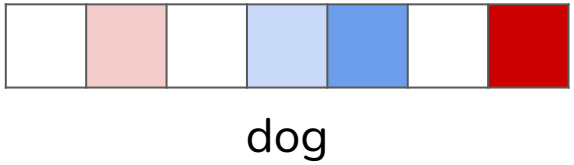
Clustering foils by explanation

Input: General relativity predicts the existence of

Output: black



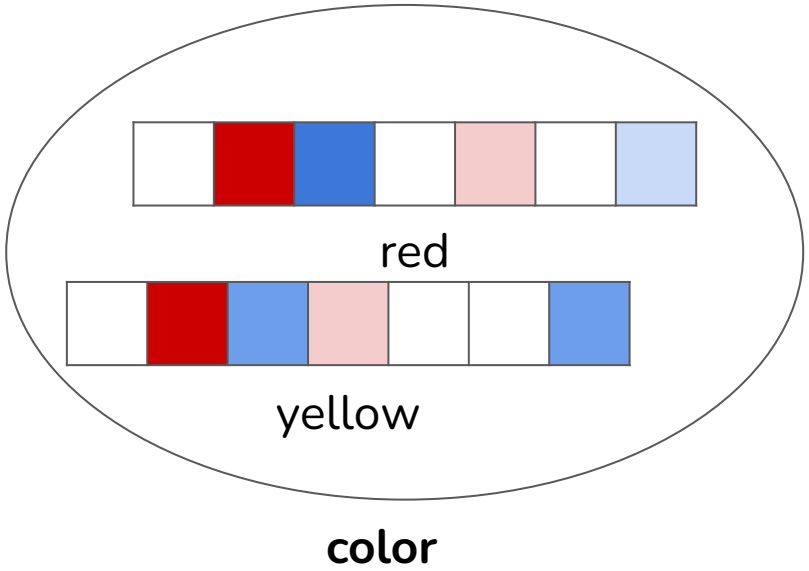




Clustering foils by explanation

Input: General relativity predicts the existence of

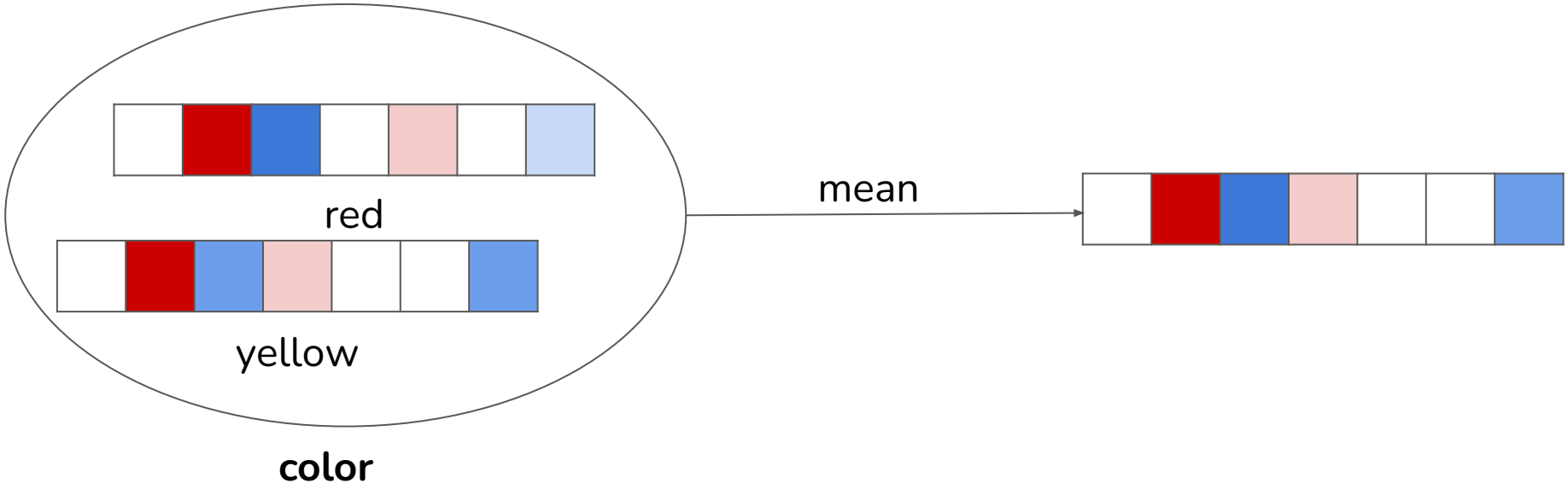
Output: black



Clustering foils by explanation

Input: General relativity predicts the existence of

Output: black



Scaling up the cluster analysis

- Target: 10 most frequent words for each POS

Scaling up the cluster analysis

- Target: 10 most frequent words for each POS
- Foil: 10,000 most frequent vocabulary items

Scaling up the cluster analysis

- Target: 10 most frequent words for each POS
- Foil: 10,000 most frequent vocabulary items
- Input sentence: 500 randomly selected

Foils related to a linguistic phenomenon are clustered together

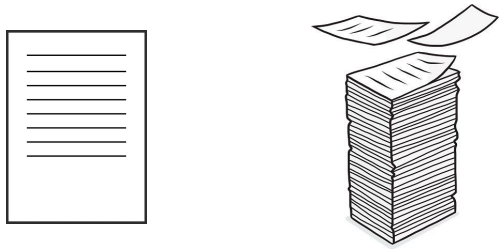
Phenomenon	Target	Foil Cluster	Example
Anaphor Agreement	he	she, her, She, Her, herself, hers	That night , Ilsa confronts Rick in the deserted café . When he refuses to give her the letters , _____

Foils related to a linguistic phenomenon are clustered together

Phenomenon	Target	Foil Cluster	Example
Anaphor Agreement	he	she, her, She, Her, herself, hers	That night , Ilsa confronts Rick in the deserted café . When he refuses to give her the letters , _____
Animate Subject	man	fruit, mouse, ship, acid, glass, water, tree, honey, sea, ice, smoke, wood, rock, sugar, sand, cherry, dirt, fish, wind, snow	You may not be surprised to learn that Kelly Pool was neither invented by a _____

Foils related to a linguistic phenomenon are clustered together

Phenomenon	Target	Foil Cluster	Example
Anaphor Agreement	he	she, her, She, Her, herself, hers	That night , Ilsa confronts Rick in the deserted café . When he refuses to give her the letters , _____
Animate Subject	man	fruit, mouse, ship, acid, glass, water, tree, honey, sea, ice, smoke, wood, rock, sugar, sand, cherry, dirt, fish, wind, snow	You may not be surprised to learn that Kelly Pool was neither invented by a _____
Determiner-Noun Agreement	page	tabs, pages, icons, stops, boxes, doors, shortcuts, bags, flavours, locks, teeth, ears, tastes, permissions, stairs, tickets, touches, cages, saves, suburbs	Immediately after "Heavy Competition" first aired, NBC created a sub- _____



Foils related to a linguistic phenomenon are clustered together

Phenomenon	Target	Foil Cluster	Example
Anaphor Agreement	he	she, her, She, Her, herself, hers	That night , Ilsa confronts Rick in the deserted café . When he refuses to give her the letters , _____
Animate Subject	man	fruit, mouse, ship, acid, glass, water, tree, honey, sea, ice, smoke, wood, rock, sugar, sand, cherry, dirt, fish, wind, snow	You may not be surprised to learn that Kelly Pool was neither invented by a _____
Determiner-Noun Agreement	page	tabs, pages, icons, stops, boxes, doors, shortcuts, bags, flavours, locks, teeth, ears, tastes, permissions, stairs, tickets, touches, cages, saves, suburbs	Immediately after "Heavy Competition" first aired, NBC created a sub- _____
Subject-Verb Agreement	go	doesn, causes, looks, needs, makes, isn, says, seems, seeks, displays, gives, wants, takes, uses, fav, contains, keeps, sees, tries, sounds	Mala and the Eskimos _____

Contrastive explanations help us characterize how LMs make decisions

See paper for aggregate analysis of linguistic distinctions and results

Contrastive explanations help us characterize how LMs make decisions

-

Phenomenon / POS	Target	Foil Cluster	Example
ADJ	black	Black, white, black, White, red, BLACK, green, brown, dark, orange, African, blue, yellow, pink, purple, gray, grey, whites, Brown, silver	Although general relativity can be used to perform a semi @-@ classical calculation of _____
ADJ	black	Asian, Chinese, English, Italian, American, Indian, East, South, British, Japanese, European, African, Eastern, North, Washington, US, West, Australian, California, London	While taking part in the American Negro Academy (ANA) in 1897 , Du Bois presented a paper in which he rejected Frederick Douglass 's plea for _____
ADP	for	to, in, and, on, with, for, when, from, at, (, if, as, after, by, over, because, while, without, before, through	The war of words would continue _____
ADV	back	the, to, a, in, and, on, of, it, ", not, that, with, for, this, from, up, just, at, (, all	One would have thought that claims dating _____
DET	his	the, you, it, not, that, my, [, this, your, he, all, so, what, there, her, some, his, time, him, He	A preview screening of Sweet Smell of Success was poorly received , as Tony Curtis fans were expecting him to play one of _____
NOUN	girl	Guy, Jack, Jones, Robin, James, David, Tom, Todd, Frank, Mike, Jimmy, Michael, Peter, George, William, Bill, Smith, Tony, Harry, Jackson	Veronica talks to to Sean Friedrich and tells him about the _____
NUM	five	the, to, a, in, and, on, of, is, it, ", not, that, 1, with, for, 2, this, up, just, at	From the age of _____
VERB	going	got, didn, won, opened, told, went, heard, saw, wanted, lost, came, started, took, gave, happened, tried, couldn, died, turned, looked	Truman had dreamed of _____

Bonus: interpreting NMT models with contrastive explanations

Why did the model predict “carnet”?

En: I like my old notebook better than my new notebook.

Fr: J’aime mieux mon ancien carnet que mon nouveau

Bonus: interpreting NMT models with contrastive explanations

Why did the model predict “carnet”?

En: I like my old notebook better than my new notebook.

Fr: J’aime mieux mon ancien carnet que mon nouveau

Bonus: interpreting NMT models with contrastive explanations

Why did the model predict “carnet”?

En: I like my old notebook better than my new notebook.

Fr: J’aime mieux mon ancien carnet que mon nouveau

Why did the model predict “carnet” instead of “ordinateur”?

En: I like my old notebook better than my new notebook.

Fr: J’aime mieux mon ancien carnet que mon nouveau

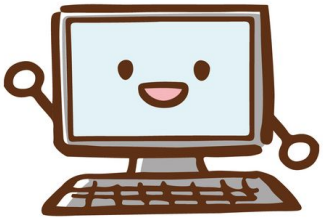
Bonus: interpreting NMT models with contrastive explanations

Why did the model predict “carnet”?

En: I like my old notebook better than my new notebook.

Fr: J’aime mieux mon ancien carnet que mon nouveau

Why did the model predict “carnet” instead of “ordinateur”?



En: I like my old notebook better than my new notebook.

Fr: J’aime mieux mon ancien carnet que mon nouveau

Summary

Contrastive explanations...

1. can better identify linguistically appropriate evidence
2. improve model simulatability
3. help us characterize how LMs make decisions