Interpreting Language Models with Contrastive Explanations

Yin and Neubig, 2022

GROUP 13

Cristina Racoviță Bogdan Bîndilă

Yann LeCun:

“LLMs are stochastic parrots”

Contrastive Explanations on LLM

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

1. Why did the model predict “barking”?

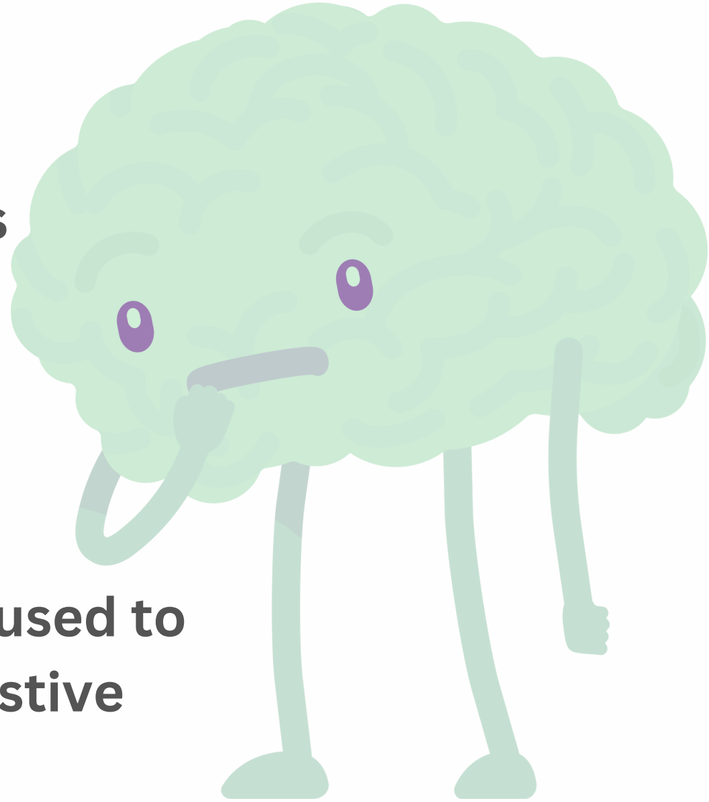
Can you stop the dog from

2. Why did the model predict “barking” instead of “crying”?

Can you stop the dog from

1. How similar are our results to those of the authors when redoing the experiment on identifying linguistically appropriate evidence?

2. To what extent do contrastive explanations help non-native English speakers predict LM behavior?



3. What is the relevance of the data sets used to demonstrate the effectiveness of contrastive methods for real-world use cases?

Data

BLiMP: The Benchmark of Linguistic Minimal Pairs

Automatically generated according to expert-crafted grammars

1000 minimal pairs / linguistic paradigm

{
sentence_good: “Most teenagers boasted about themselves.”,
sentence_bad: “Most teenagers boasted about himself.”
}

Methods

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

Contrastive Gradient Norm

$$S^*_{GN}(x_i) = ||g^*(x_i)||_{L1}$$

how much an input token increases the probability of y true and decreases the probability of y foil

Gradient x Input

$$S^*_{GI}(x_i) = g^*(x_i) \cdot x_i$$

how much each token contributes to the saliency score

Input Erasure

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

how much erasing a token from the input increases the likelihood of the foil and decreases the likelihood of the target in the model's output

Experiments

Do Contrastive Explanations Identify Linguistically Appropriate Evidence?

Focus on GPT2 and contrastive explanations only
9000 pairs from 4 phenomena

Argument Structure
The glove was noticed by some **woman / mouse**.

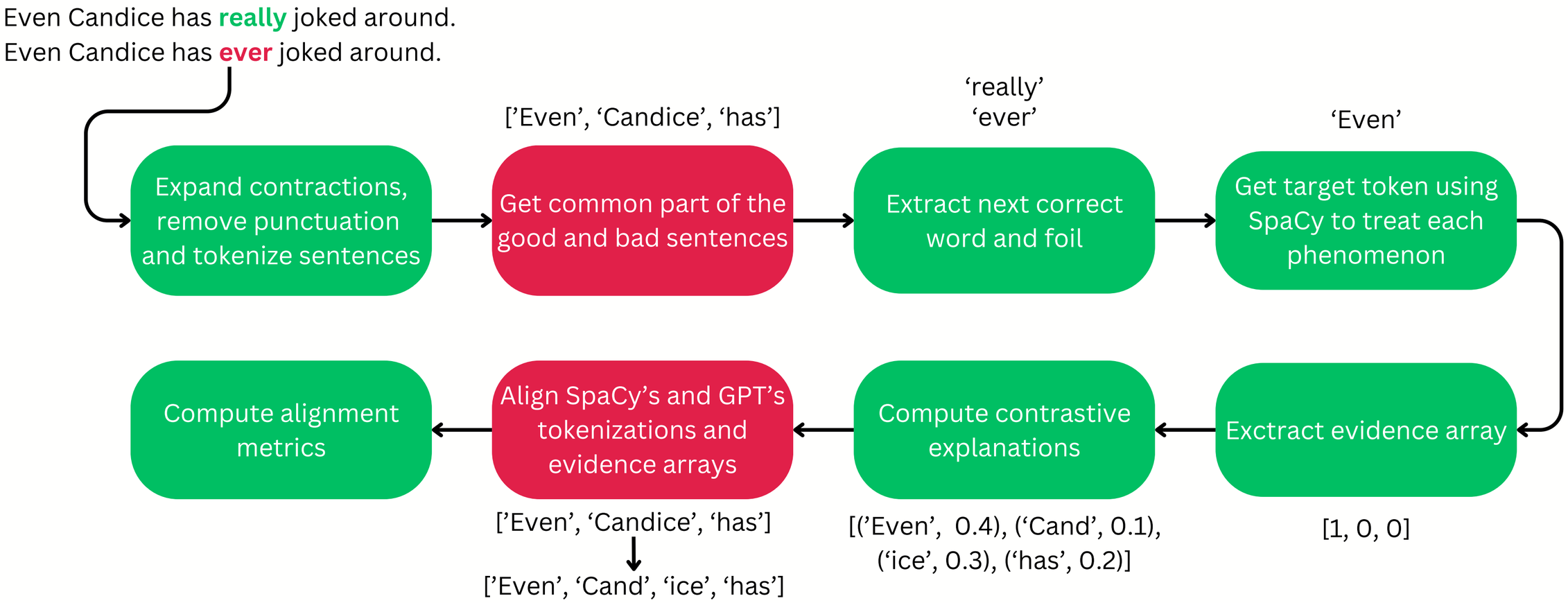
Determiner-Noun Agreement
Craig explored that grocery **store / stores.**

Negative Polarity Items (NPI) Licensing
Even Candice has **really** / **ever** joked around.

Subject-Verb Agreement
Nina **hasn't / haven’t** cleaned the skirt.

Experiments

Do Contrastive Explanations Identify Linguistically Appropriate Evidence?



Metrics

Average Dot Product

average dot product value of saliency scores and known evidence

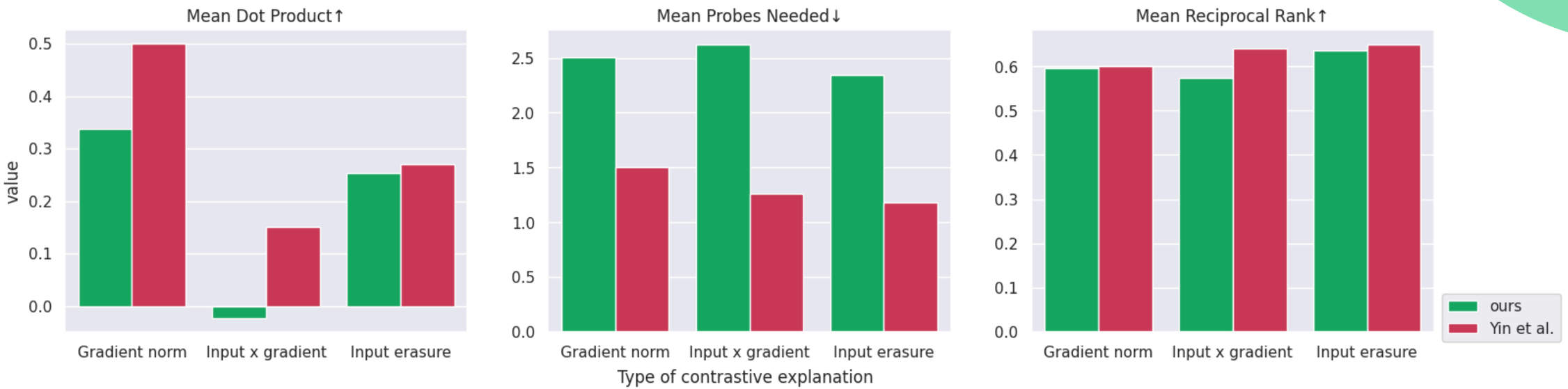
Average Probes Needed

mean number of tokens we need to probe until we find a token of the known evidence

Mean Reciprocal Rank (MRR)

average of inverse of the rank of the first token that is part of the known evidence

Results - Comparison



These low values indicate that:

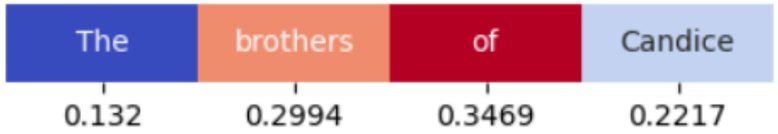
- the evidence token appears at the beginning of the sentence
- sentences are exceptionally short

Differences in the results because the first phenomenon has not been included yet

Experiments

Do Contrastive Explanations Help Users Predict LM Behavior?

The brothers of Candice ... *



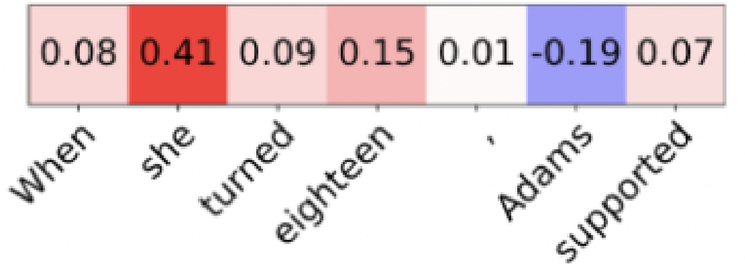
- have
- has

32 questions:

- 8 / phenomenon
- each with one of the explanations

10 non-native English speakers

vs



Which token did the model more likely predict?

- **herself**
- **himself**

10 machine learning students

400 questions

For each sentence, there is a question about the usefulness of the explanation.

Challenges and conclusions

Different tokenizers (**spaCy** and **GPT2**) complicate the implementation

Explanations are given at the **token level**

What happens when you have **only one word** in your sentence?

The dataset is based only on **grammar rules** and the sentences are too short to assess the real-world usefulness of the method

Next Steps

Send the form & analyse the responses

Include the first phenomenon (anaphor agreement)

Research the current state-of-the-art for contrastive explanations

Try the explanations for something more challenging like reasoning questions

Thank You

Questions?

Paradigm	Good Sentence	Bad Sentence
Anaphor Agreement	Katherine can't help **herself**.	Katherine can't help **himself**.
Argument Structure	Amanda was respected by some **waitresses**.	Amanda was respected by some **picture**.
Determiner-Noun Agreement	Phillip was lifting this **mouse**.	Phillip was lifting this **mice**.
NPI Licensing	Even these trucks have **often** slowed.	Even these trucks have **ever** slowed.
Subject-Verb Agreement	This goose **isn't** bothering Edward.	This goose **weren't** bothering Edward.