Interpreting Language Models with Contrastive Explanations

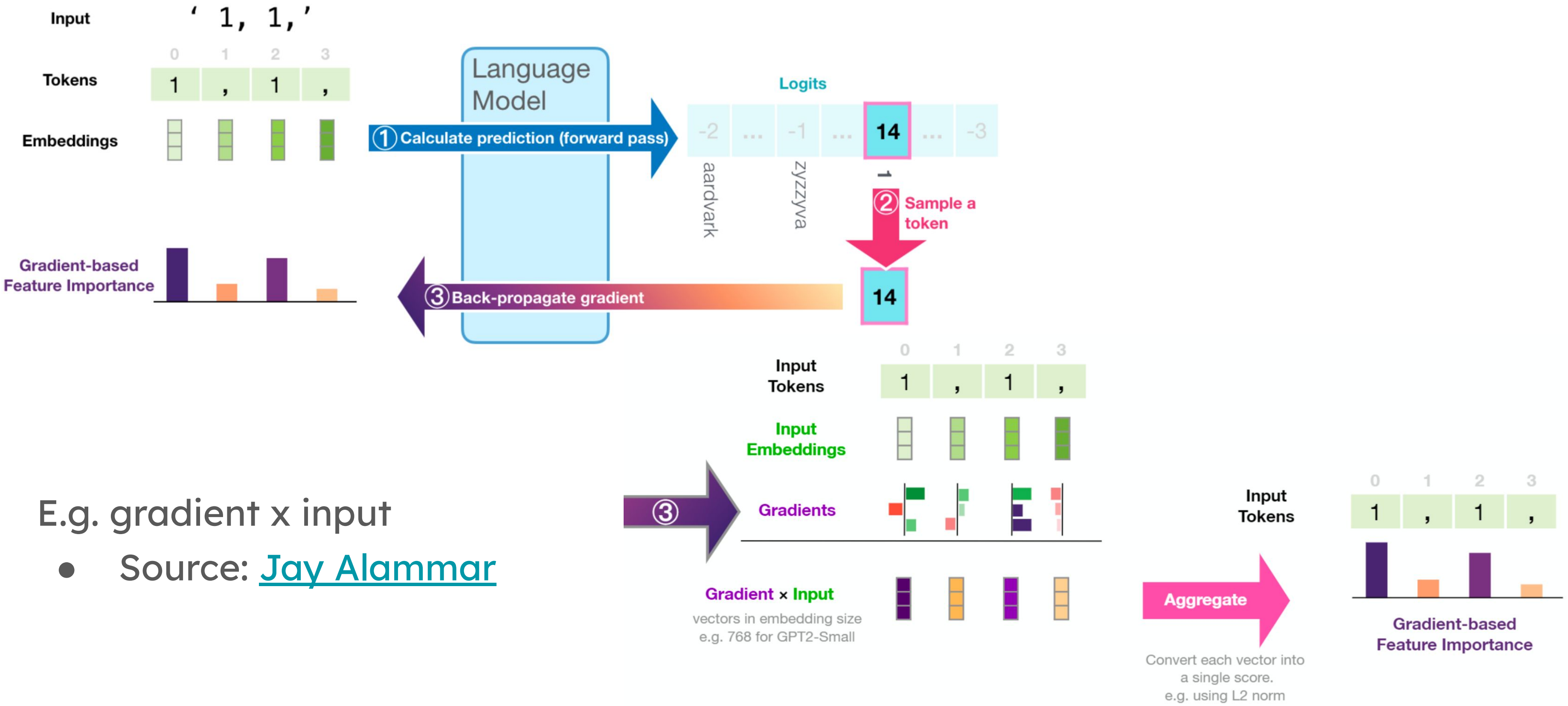
Kayo Yin & Graham Neubig

Presented by Charumathi Badrinath, Eric Shen, Leonard Tang, and Skyler Wu

Motivation + Example

- We've seen many interpretability and explanation strategies being applied to LMs, including transformer-based autoregressive LMs…
  - **Gradient-based/erasure-based feature attribution** methods provide a straightforward way to do this
  - Interpret each token of the input text as a feature
  - At each step, use gradients to calculate a **saliency score** to quantify the importance of each previous token to the model output (i.e. logit prediction for the next token)
- E.g. gradient x input

Motivation + Example



E.g. gradient x input

- Source: [Jay Alammar]

Motivation + Example

- **Problem:** For many LMs, using typical gradient- or erasure-based methods doesn't provide informative explanations
- Most of the time, the token with the highest saliency is the token immediately before the prediction
- How can we use these explanation ideas to create more meaningful attributions? By ***contrasting*** them with other token predictions.

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

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

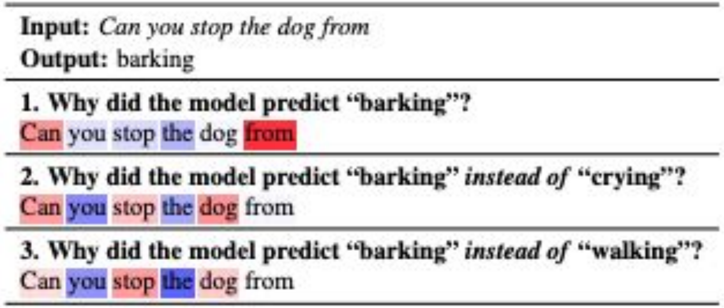
Knowing the previous word is certainly very important for figuring out the next word, but that's not very helpful!

Main Contributions and Key Ideas

1. **Contrastive explanations:** why did the model predict one token *instead* of another? Which tokens were most influential? Extended previous methods.
2. **Grammatical consistency:** contrastive explanations > non-contrastive explanations w.r.t. verifying linguistic / grammatical phenomena.
3. **Human simulatability:** contrastive explanations help users better predict LLM behavior, also found to be more useful explanations by humans.

Non-contrastive gradient x input

Contrastive explanations



Red = raise probability of "barking," Blue = decrease probability of "barking," White = little influence.

GPT-2 (Radford et al. 2019)

- Authors focus on GPT-2 (1.5B) and GPT-Neo (2.7B) → very similar to each other.
- **Training:** WebText dataset of 8 million web-pages.
  - No task-specific supervised training = "Multi-task training"
- **Objective:** predict the next word, given all previous words in input.
- **Behavior:** "chameleon-like," adapts to style + content of the input text.
- **Architecture:** Transformer-based architecture.
  - "Autoregressive": outputs tokens one at a time, *but* each token generated is appended to the input sequence → feed back to the model for next step.

Sources:

1. OpenAI
2. Jay Alammar
3. Radford et al. 2018

Gradient Norm Saliency Scores (Simonyan et al. 2013)

- **Originally for image classification:** compute gradient of class score w.r.t input image, take the norm. Main idea = big gradient, big influence.
- **For LLMs:** compute gradient of next token in input sequence w.r.t. current input.

Specific token in input (as embedding vector)

Entire input sequence (as embedding)

$$g(x_i) = \nabla_{x_i} \underbrace{q(y_t|\boldsymbol{x})}_{\text{Model output for the predicted token, given input (logit)}}$$

g outputs a vector!

$$\underbrace{S_{GN}(x_i)}_{\text{Saliency score}} = ||g(x_i)||_{L1}$$

Gradient x Input Saliency Scores (Shrikumar et al. 2016)

- **Method:** Similar gradient computation as Gradient Norm, simply replacing L1 norm with dot product with input itself.

$$g(x_i) = \overbrace{\nabla_{x_i} q(y_t|\boldsymbol{x})}^{\text{Exact same gradient computation}} \quad \Longrightarrow \quad S_{GI}\overbrace{(x_i) = g(x_i)}^{\text{Dot product, instead of L1 norm}} \cdot x_i$$

Input Erasure Saliency Scores (Li et al. 2016)

- **Intuition:** how does erasing different parts of the input affect the output?
- **Procedure:** compute difference in model outputs using full input vs. input with a specific token zeroed out. NOT gradient-based!

$$S_E(x_i) = \underbrace{q(y_t|\boldsymbol{x})}_{\text{Probability of output given entire input}} - \overbrace{q(y_t|\boldsymbol{x}_{\neg i})}^{\text{Probability of output given input \textit{without} token of interest}}$$

Saliency score of token of interest

Related Work + Relevant Limitations (Pt. 1)

- Non-contrastive saliency score methods:
  - Simonyan et al. 2013, Shrikumar et al. 2016, Li et al. 2016 (see previous slides). Not NLP specific though!
  - Not very developed in NLP use cases.
  - When applied to NLP (e.g., Wallace et al. 2019), authors found that methods often **returns last token before output as most influential.**
- Adversarial Methods on NLP:
  - Wallace et al. 2019: HotFlip on NLPs, replace words to change model's prediction. AllenNLP suite.

Related Work + Relevant Limitations (Pt. 2)

- Counterfactual explanations in *text classification:*
    - Jacovi et al. 2021: erase features from input, project input representation into “contrastive space” → measure importance of erased feature by comparing class probabilities before / after erasure.
    - But, unsure how to extend into *language modeling* space, with much bigger input + output spaces.
- Contrastive methods are not new, just not used for NLP very much (Stepin et al. 2021, survey).

Method

- Simple modification to formulation of existing gradient-based explanations.
- **Contrastive** setup:

"Given input token (embedding) sequence **x**, why did the model predict this next token y?"

Given **x**, why did the model predict this next **target token** y_t **instead of** a **foil token** y_f?

Calculate gradients of the logit of y with respect to input tokens

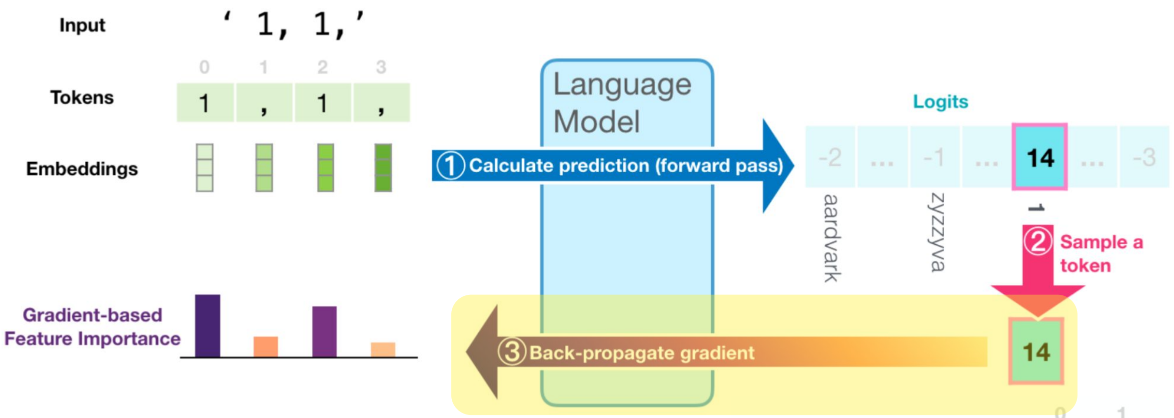
Calculate gradients of the **difference in logits** between y_t and y_f with respect to input tokens

Method

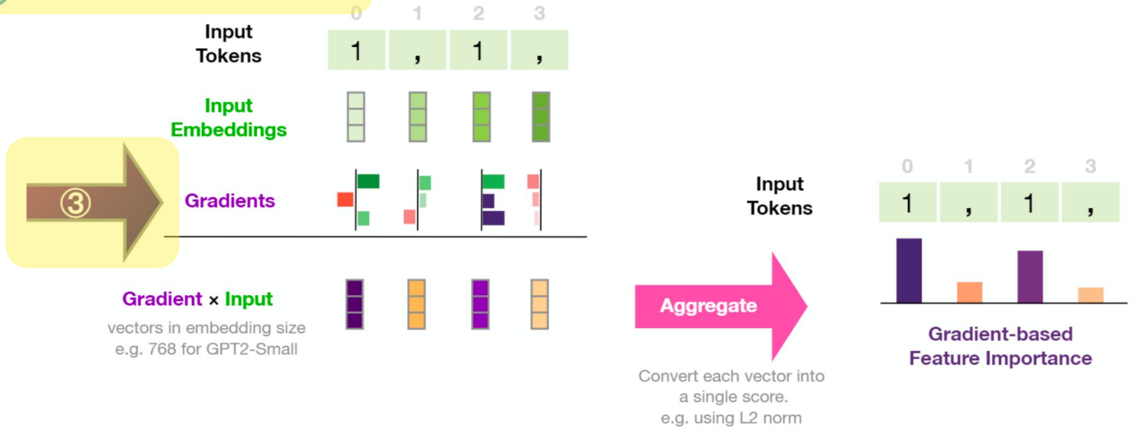
Let $q(y_t | \mathbf{x})$ be the model output for token y_t given the input $\mathbf{x}$.
Let $S(x_i)$ be the saliency score for token x_i in input $\mathbf{x}$.
Let $\mathbf{x}_{\neg i}$ be the input $\mathbf{x}$ where x_i is zeroed out.

Saliency Score Explanation Method	**Non-Contrastive (Standard) Saliency Calculation**	**Contrastive Saliency Calculation**
Gradient Norm	$S_{GN}(x_i) = \|\nabla_{x_i} q(y_t\|\mathbf{x})\|_{L_1}$	$S^*_{GN}(x_i) = \|\nabla_{x_i}(q(y_t\|\mathbf{x}) - q(y_f\|\mathbf{x}))\|_{L_1}$
Gradient x Input	$S_{GI}(x_i) = \nabla_{x_i} q(y_t\|\mathbf{x}) \cdot x_i$	$S_{GI}(x_i) = \nabla_{x_i}(q(y_t\|\mathbf{x}) - q(y_f\|\mathbf{x})) \cdot x_i$
Input Erasure	$S_E(x_i) = q(y_t\|\mathbf{x}) - q(y_t\|\mathbf{x}_{\neg i})$	$S^*_E(x_i) = (q(y_t\|\mathbf{x}) - q(y_t\|\mathbf{x}_{\neg i})) - (q(y_t\|\mathbf{x}_{\neg i}) - q(y_f\|\mathbf{x}_{\neg i}))$

Method



Step 3 is changed;
we calculate and
back-propagate
a different quantity



Do Contrastive Explanations Identify Linguistically Appropriate Evidence?

Q: Are contrastive explanations >> non-contrastive explanations in identifying words that we think should influence the output token?

Experimental Setup:

- BLiMP dataset: pairs of minimally different English sentences that contrast in grammatical acceptability under some linguistic paradigm
- 5 linguistic phenomena with 12 paradigms + used spaCy NLP library to extract grammatically relevant parts of each sentence

Do Contrastive Explanations Identify Linguistically Appropriate Evidence?

gender + number of pronoun must agree with antecedent

action verbs used with animate objects

some negative polarity words only appear in some contexts

number of subject and verb in present tense must agree

Phenomenon	UID[3]	Acceptable Example	Unacceptable Example
Anaphor Agreement	aga	Katherine can't help **herself**.	Katherine can't help **himself**.
	ana	Many teenagers were helping **themselves**.	Many teenagers were helping **herself**.
Argument Structure	asp	Amanda was respected by some **waitresses**.	Amanda was respected by some **picture**.
Determiner-Noun Agreement	dna	Craig explored that grocery **store**.	Craig explored that grocery **stores**.
	dnai	Phillip was lifting this **mouse**.	Phillip was lifting this **mice**.
	dnaa	Tracy praises those lucky **guys**.	Tracy praises those lucky **guy**.
	dnaai	This person shouldn't criticize this upset **child**.	This person shouldn't criticize this upset **children**.
NPI Licensing	npi	Even these trucks have **often** slowed.	Even these trucks have **ever** slowed.
Subject-Verb Agreement	darn	A sketch of lights **doesn't** appear.	A sketch of lights **don't** appear.
	ipsv	This goose **isn't** bothering Edward.	This goose **weren't** bothering Edward.
	rpsv	Jeffrey **hasn't** criticized Donald.	Jeffrey **haven't** criticized Donald.

Do Contrastive Explanations Identify Linguistically Appropriate Evidence?

Evaluation Metrics:

S = explanation vector; S_i = saliency of x_i

E = known evidence; $E_i = 1(x_i$ grammatically influences output token)

- $S \cdot E$ → sum of saliency scores of all input tokens that are part of evidence
- Probes needed → ranking of first token x_i where $E_i = 1$ when sorted by decreasing saliency
- MRR (mean reciprocal rank) → average (over all sentences) of inverse rank of first x_i where $E_i = 1$ when sorted in descending saliency

Findings: Linguistic Agreement

- Contrastive (cf. non-contrastive) explanations are more aligned with linguistic paradigms
- Contrastive explanations have a better alignment with BLiMP than random vectors baseline
  - Note high baseline
- Non-contrastive explanations do not outperform random baseline

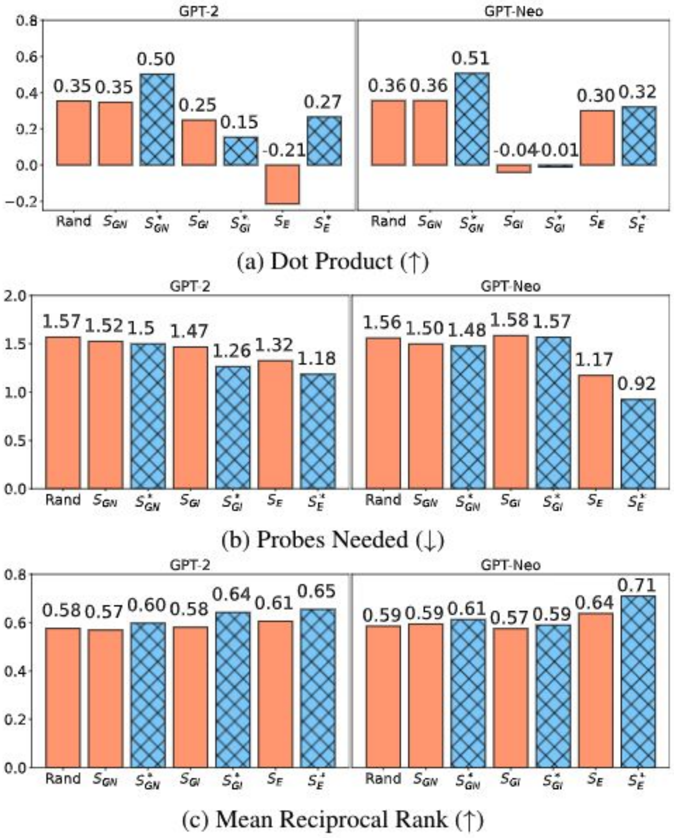


Figure 1: Alignment of different GPT-2 (left) and GPT-Neo (right) explanations with known evidence in BLiMP according to dot product (top), probes needed (middle), mean reciprocal rank (bottom) averaged over linguistic paradigms.

Findings: Linguistic Agreement

- Further apart *known evidence* token is from *target token* → larger increase in alignment (MRR) of contrastive cf. non-contrastive
- Contrastive explanations can particularly capture model decisions requiring *longer-range context*

Katherine can't help **herself**.

Phillip was lifting this **mouse**.

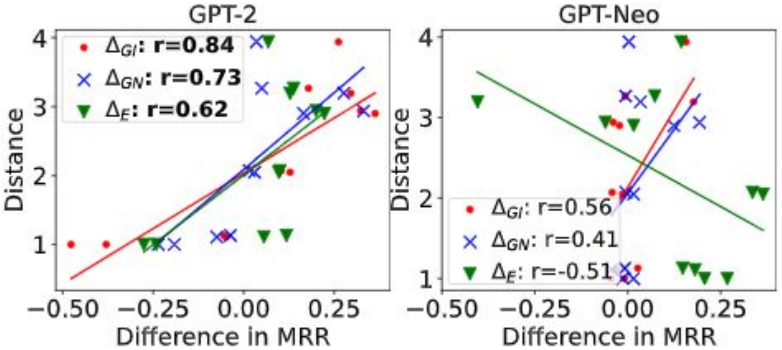


Figure 2: Scatter plot of the average distance of the known evidence to the target token across each linguistic paradigm against the difference in MRR scores between the contrastive and non-contrastive versions of each explanation method, with the Pearson correlation for each explanation method. Statistically significant Pearson's r values ($p < 0.05$) are in **bold**. In most cases, there is a positive correlation between the increase in MRR and the distance of the evidence.

Do Contrastive Explanations Help Users Predict LM Behavior?

Q: Do contrastive explanations increase users' ability to predict a model's output token (i.e. "simulate" model behavior)?

** normalized for model accuracy, # of each condition

Experimental Setup:

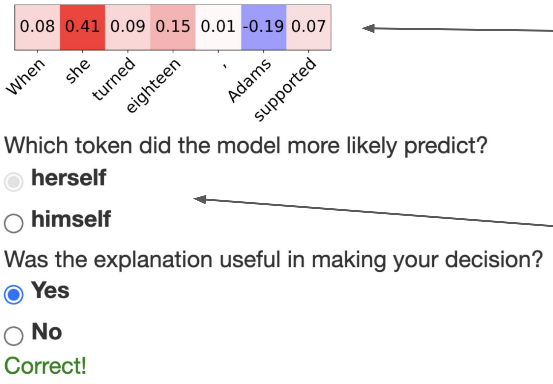


Figure 3: Example of a prompt in our human study.

model = GPT-2

explanation = {no explanation, gradient x input, contrastive gradient x input, erasure, contrastive erasure}

10 word pairs from BLiMP, 10 word pairs selected to maximize confusion score on WikiText-103 test split

$$P(x_{true} = a, x_{model} = b) = \frac{1}{N} \sum_{x \in X} \sum_{t \in \text{pos}(x) | x_t = a} P_{model}(\hat{x}_t = b | x_{<t})$$

$$\mathcal{C}(a, b) = \min(P(x_{true} = a, x_{model} = b), P(x_{true} = b, x_{model} = a))$$

Findings: User Alignment

- All four types of explanations help users simulate model behavior
- *Contrastive explanations* lead to *more accurate simulations*
- *Contrastive explanations* are considered *more useful*
- Takeaway: contrastive explanations help human observers accurately simulate model predictions the most

	Acc.	Acc. Correct	Acc. Incorrect	Useful	Acc. Useful	Acc. Not Useful
None	61.38	74.50	48.25	–	–	–
S_{GI}	64.00	78.25	49.75	62.12	67.20	58.75
S^*_{GI}	**65.62**	**79.00**	52.25	**63.88**	69.67	58.48
S_E	63.12	**79.00**	47.25	46.50	65.86	60.75
S^*_E	**64.62**	77.00	52.25	**64.88**	70.52	53.74

Table 3: Simulation accuracy (%) in predicting GPT-2 outputs and subjective usefulness of explanations for various explanation methods. For each explanation method, scores that are statistically significantly higher ($p \leq 0.05$) than the analogous method with a different contrastive setting are bolded. Overall, users achieve higher simulation accuracy with contrastive explanations.

What Context Do Models Use for Certain Decisions?

Q: How do language models achieve various linguistic distinctions? Is similar evidence necessary to disambiguate foils that are similar linguistically?

Experimental Setup:

Targets = 10 most frequent words in each major part of speech

Foils = 10000 most frequent vocab items

- For each target y_t select 500 sentences from WikiText-103 → "sentence set" X
- For each foil y_f and each sentence x_i in X generate contrastive explanation $e(x_i, y_t, y_f)$ + concatenate them
- For each target y_t apply k-means on $e(x_i, y_t, \text{all foils})$

Idea: Explanation vectors represent *type of context* needed to disambiguate foil from target; want to see if clusters associated with salient linguistic distinctions

Findings: What Context Do Models Use For Decisions?

- (Paradigmatically) linguistically similar foils cluster together
  - Foil clusters reflect linguistic distinctions unique from word embeddings
- Model use similar types of input features to make certain decisions
  - Animacy Ex) Target is animate, foils in cluster are all inanimate
- Examining cluster explanations (maps) yields insights into GPT-2 (“BERTology”)
  - Pronoun Ex) GPT-2 influenced by unrelated pronouns → produces incorrect gender

Phenomenon / POS	Target	Foil Cluster	Embd Nearest Neighbors	Example
Anaphor Agreement	he	she, her, She, Her, herself, hers	**she**,She, **her**, **She**, he, they, **Her**, we, it,she, I, that,Her, you, was, there,He, is, as, in’	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	**fruit**, fruits, Fruit, meat, flower,fruit, tomato, vegetables, **fish**, apple, berries, food, citrus, banana, vegetable, strawberry, fru, delicious, juice, foods	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	**tabs**, tab, Tab, apps, files, bags, tags, websites, sections, browsers, browser, icons, buttons, **pages**, keeps, clips, updates, 28, insists, 14	Immediately after "Heavy Competition" first aired, NBC created a sub- _____

Strengths/Weaknesses

A cursory list:

- Simple yet effective modification applicable to a variety of feature attribution methods using saliency scores
- Easy to compute, extensible to general LMs (including for NMT)
- Evaluated with interesting foil clustering analysis
- Empirically shown to help with human observers (empirically good for interpretability)

- Only applied to three feature attribution methods in the paper
- Only GPT-2 and GPT-Neo used as LM examples in main papers
- Human study very limited in scope
- Does not attempt to look at model internals; saliency scores are arguably a crude approximation of interpretation

Questions for the Audience

- Given that LLMs (and other ML models) often exhibit “phase shifts” at different sizes to what extent do you expect the results from this paper to generalize to cutting-edge models like GPT-4?
- In practice, how would one create foils for free-response questions and/or general conversational use? How generalizable are these contrastive tools?
- How effective do you think saliency scores (through gradient/erasure-base explanation methods, etc.) are for achieving interpretability?
- How much do we trust the GPT-2 embeddings (this is the primary workhorse for most of their methods) and the generalizability of the authors’ results?