Interpreting Language Models with Contrastive Explanations

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing | EMNLP 2022

Kayo Yin (University of California) **Graham Neubig** (Carnegie Mellon University)

TABLE OF CONTENTS

01

Introduction

1.1 OVERVIEW

Why Contrastive Explanations

- Traditional interpretability methods fall short in explaining language model predictions.
- LMs operate in a large, complex output space where subtle distinctions matter.
- Contrastive explanations identify why a model chose one output over another. In LMs, that would be tokens.

- **Goal:** Use contrastive explanations to provide more informative, human-intuitive explanations for LM behavior.

1.1 OVERVIEW

How It Works

- Table 1: *Explanations for the GPT-2 prediction. Input tokens that are measured to raise or lower the probability of "barking" are in red and blue respectively, and those with little influence are in white.*

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

1.2 QUESTIONS

Driving Questions (Q)

(RQ1) Are contrastive explanations better at identifying evidence that we believe, a-priori, to be useful to capture a variety of linguistic phenomena?

(RQ2) Do contrastive explanations allow human observers to better simulate language model behavior?

(RQ3) Are different types of evidence necessary to disambiguate different types of words, and does the evidence needed reflect (or uncover) coherent linguistic concepts?

02

Related Work

2.1 MODEL EXPLS.

Methods' Objective

- Explain why a model made a certain prediction by computing saliency scores over input features. Higher saliency score → greater contribution of token to the model's output.

Current Landscape

- Extensive research on input feature explanations in text classification.
- Studies focus on linguistic features like syntax in language models.

Gap in the Literature

- Few methods exist for explaining language modeling predictions directly.
- The large output space of LMs makes explanation harder.

2.2 CONTRASTIVE EXPLS.

Methods' Objective

- Contrastive explanations clarify why, for a given input x, the model predicts a target y_t instead of a foil y_f.

Related Concepts

- Counterfactual explanations modify input x to make y_f more likely than y_t.
- Use feature erasure in text classification to identify contrastive factors by projecting inputs into a space separating decisions.

Gap in the Literature

- Extend contrastive explanations to language models, where input and output complexity is much higher.

03

Method

3.1 GRADIENT NORM

Standard Gradient Norm

- Measures saliency using the L1 norm of the gradient of output w.r.t. input.
- Highlights which input tokens most influence the model's prediction of token y_t.

$$g(x_i) = \nabla_{x_i} q(y_t|\boldsymbol{x})$$

$$S_{GN}(x_i) = ||g(x_i)||_{L1}$$

Contrastive Extension: Contrastive Gradient Norm (CGN)

- Measures how input x_i shifts the model's preference for target y_t over foil y_f.
- Captures differential influence of x_i on selecting y_t instead of y_f.

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

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

3.2 GRADIENT × INPUT

Standard Gradient × Input

- Computes saliency as the dot product between the gradient and input.
- Captures how much each input token contributes to the prediction based on both sensitivity and magnitude.

$$S_{GI}(x_i) = g(x_i) \cdot x_i$$

Contrastive Extension: Contrastive Gradient × Input (CG×I)

- Measures differential contribution of token x_i in target vs. foil prediction.
- Highlights how the token affects preference for one output over another.

$$S^*_{GI}(x_i) = g^*(x_i) \cdot x_i$$
$$\text{where } g^*(x_i) = \nabla_{x_i}(q(y_t \mid x) - q(y_f \mid x))$$

3.3 INPUT ERASURE

Standard Input Erasure

- Measures saliency by removing a token and observing the change in output.
- Indicates how much token x_i contributes to the prediction of y_t.

$$S_E(x_i) = q(y_t \mid x) - q(y_t \mid x_{\neg i})$$

Contrastive Extension: Contrastive Input Erasure (CIE)

- Measures how erasing x_i affects the preference between target and foil tokens.
- High value means removing x_i hurts y_t more than y_f, suggesting it supports y_t.
- More accurate but computationally expensive (multiple forward passes).

$$S_E^*(x_i) = [q(y_t \mid x) - q(y_t \mid x_{\neg i})] - [q(y_f \mid x) - q(y_f \mid x_{\neg i})]$$

04

Experiments

4.1 RQ1: LINGUISTIC EVIDENCE

Motivation & Methodology

- **Goal:**
  - Assess whether contrastive explanations better identify linguistically meaningful evidence than non-contrastive ones.
- **Approach:**
  - Use BLiMP dataset (67 linguistic paradigms) of minimal sentence pairs differing in grammatical acceptability.
  - Define ground-truth evidence tokens via linguistic rules (e.g. anaphor agreement depends on antecedents).
  - Compare how explanation methods align with this known evidence.

4.1 RQ1: LINGUISTIC EVIDENCE

Linguistic Phenomena

- Five key phenomena studied, and rule-based extraction done using spaCy.
- Rules identify input tokens that should control the prediction decision (e.g. "teenagers" → "themselves").

Phenomenon	Acceptable Example	Unacceptable Example	Rule
Anaphor Agreement	Katherine can't help **herself**. Many teenagers were helping **themselves**.	Katherine can't help **himself**. Many teenagers were helping **herself**.	`coref` `coref`
Argument Structure	Amanda was respected by some **waitresses**.	Amanda was respected by some **picture**.	`main_verb`
Determiner-Noun Agreement	Phillip was lifting this **mouse**. Tracy praises those lucky **guys**.	Phillip was lifting this **mice**. Tracy praises those lucky **guy**.	`det_noun` `det_noun`
NPI Licensing	Even these trucks have **often** slowed.	Even these trucks have **ever** slowed.	`npi`
Subject-Verb Agreement	A sketch of lights **doesn't** appear.	A sketch of lights **don't** appear.	`subj_verb`

Table 2: Examples of BLiMP minimal pairs. Contrastive tokens are **bolded**. Tokens extracted by our rules that enforce grammatical acceptability are underlined.

4.1 RQ1: LINGUISTIC EVIDENCE

Evaluation Metrics

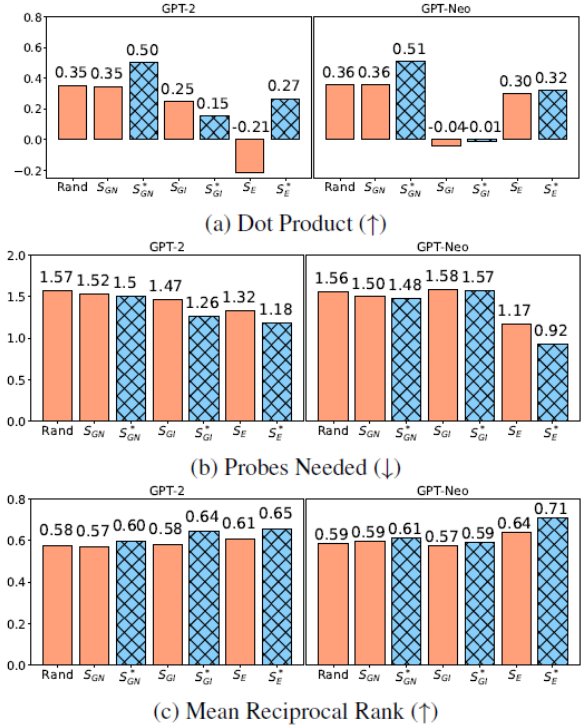
- **Models Used:**
  1. *Dot Product*: Sum of saliency on known evidence tokens.
  2. *Probes Needed*: How many top-ranked tokens before a relevant one appears.
  3. *Mean Reciprocal Rank (MRR)*: Inverse of the rank of the first relevant token.

- **Interpretation:**
  - Higher dot product and MRR = better alignment
  - Fewer probes needed = more efficient explanation

4.1 RQ1: LINGUISTIC EVIDENCE

Results Overview

- **Models Used:**
  - *GPT-2*: 1.5B parameters
  - *GPT-Neo*: 2.7B parameters
- **Key Findings:**
  - Contrastive outperforms non-contrastive across all metrics and both models.
  - Random baseline performs worse than contrastive, sometimes better than non-contrastive.



4.1 RQ1: LINGUISTIC EVIDENCE

Error Analysis & Insights

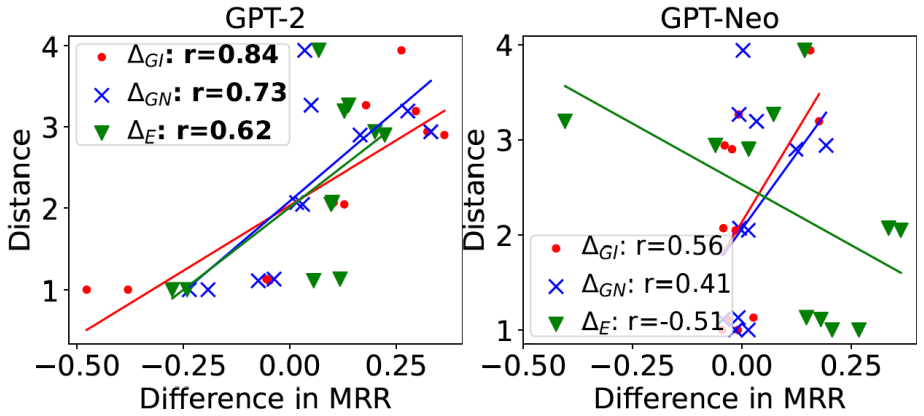
- On correct predictions, contrastive methods align significantly better.
- On incorrect predictions, performance varies; contrastive methods are still competitive.

	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

4.1 RQ1: LINGUISTIC EVIDENCE

Error Analysis & Insights

- Contrastive methods are better at capturing long-distance dependencies.
- Strong positive correlation between evidence-target distance and contrastive advantage.



4.2 RQ2: USER PREDICTION

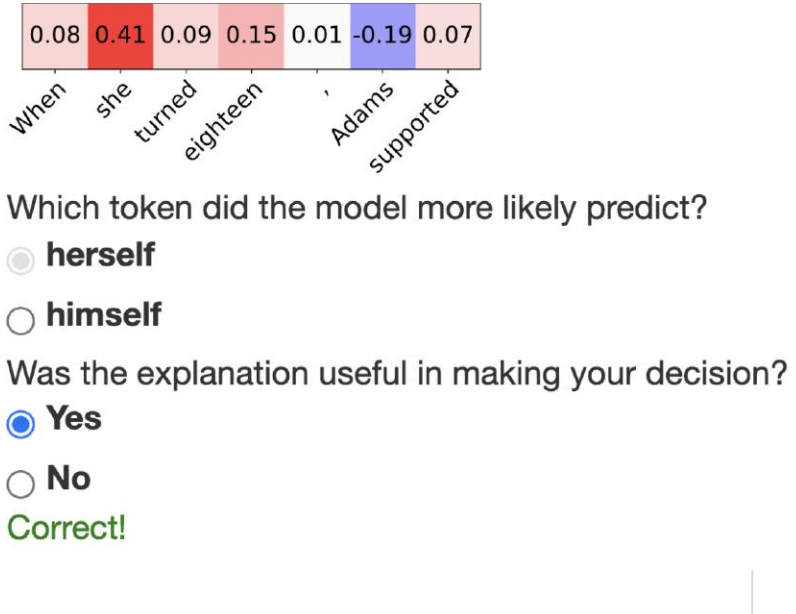
Motivation & Methodology

- **Goal:**
  - Evaluate whether explanations help users simulate model behavior—i.e., predict what token the model will choose.
- **Approach:**
  - Compare user performance with: no explanation; non-contrastive explanations; contrastive explanations (proposed method).
- **Key Metric:**
  - Simulation accuracy; how often users predict the model's choice.

4.2 RQ2: USER PREDICTION

Study Design

- **Explanation Types:**
  - Gradient × Input (G×I)
  - Contrastive G×I (CG×I)
  - Erasure
  - Contrastive Erasure (CE)
- **Participants:**
  - 10 ML grad students
  - Each viewed 40 sentence–word pair tasks



4.2 RQ2: USER PREDICTION

Results Overview

- All explanations improve simulation accuracy vs. no explanation.
- Contrastive > Non-contrastive across methods.
- Users performed best when contrastive explanations were provided.
- Users rated contrastive explanations as significantly more useful.

	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

4.3 RQ3: MODEL CONTEXTS

Motivation & Overview

- **Goal:**
  - Use contrastive explanations to uncover the types of context language models rely on for specific linguistic distinctions.
- **Hypothesis:**
  - Linguistically similar foils require similar context to disambiguate.
  - These contexts can be identified and grouped through clustering.
- **Approach:**
  - Represent each foil by its contrastive saliency vector, & cluster these vectors to infer linguistic distinctions.

4.3 RQ3: MODEL CONTEXTS

Methodology

- **Data:**
  - *Targets*: 10 most frequent words per POS in WikiText-103.
  - *Foils*: 10,000 most frequent vocabulary tokens.
  - *Sentences*: 500 randomly sampled per target.
- **Steps:**
  1. Generate contrastive explanations (Gradient Norm & G × Input only)
  2. Aggregate saliency maps: $e(y_t, y_f) = \cup\, e(x_i, y_t, y_f)$
  3. Apply k-means clustering to group foils for each target.
  4. Compare foil clusters to nearest neighbors in embedding space.

4.3 RQ3: MODEL CONTEXTS

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- _____
Subject-Verb Agreement	go	doesn, causes, looks, needs, makes, isn, says, seems, seeks, displays, gives, wants, takes, uses, fav, contains, keeps, sees, tries, sounds	**doesn**, **isn**, didn, does, hasn, wasn, don, wouldn, **makes**, gets, has, is, aren, **gives**, Doesn, couldn, **seems**, **takes**, **keeps**,doesn	Mala and the Eskimos _____
ADJ	black	Black, white, black, White, red, BLACK, green, brown, dark, orange, African, blue, yellow, pink, purple, gray, grey, whites, Brown, silver	**Black**,Black, **black**,black, **White**, **BLACK**, **white**, Blue, Red,White, In, B, The,The, It, **red**, Dark, 7, Green, **African**	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	**Asian**,Asian, Asia, Asians, **Chinese**, **African**, **Japanese**, Korean, China, **European**, **Indian**, ethnic,Chinese, Japan, **American**, Caucasian, **Australian**, Hispanic, white, Arab	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	**to**, **in**, **for**, **on**, **and**, **as**, **with**, of, a, **at**, that, the, **from**, **by**, an, (, To, is, it, or	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	**the**, **a**, an, **it**, **this**, **that**, **in**, The, **to**,The, **all**, **and**, their, as, **for**, **on**, his, **at**, some, what	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	**the**, a, an, **it**, **this**, **that**, in, The, to,The, **all**, and, their, as, for, on, **his**, at, **some**, **what**	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	**Guy**,Guy, guy,guy, Gu, Dave, Man, dude, Girl, Guys, John, Steve, \x00, \xef \xbf \xbd, \xef \xbf \xbd, \x1b, \xef\xbf \xbd, \x12, \x1c, \x16	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	**the**, **a**, an, **it**, **this**, **that**, **in**, The, **to**,The, all, **and**, their, as, **for**, **on**, his, **at**, some, what	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	**got**, gets, get, had, **went**, **gave**, **took**, **came**, **didn**, did, getting, been, became, has, was, made, **started**, have, gotten, showed	Truman had dreamed of _____

4.3 RQ3: MODEL CONTEXTS

Observations

1. **Anaphor Agreement:**
   - Female pronouns cluster together when target is a male pronoun.
   - Embedding neighbors of "she" include mixed genders—clusters do not.
2. **Animacy:**
   - Animate noun targets → clusters of inanimate noun foils.
   - "Fruit" clusters with inanimate nouns, not with embedding neighbors.
3. **Plurality:**
   - Plural and singular noun clusters align with grammatical rules.
   - Embedding neighbors mix singular/plural forms.
4. **Subject-Verb Agreement:**
   - Plural verbs cluster with singular foils.
   - Embedding neighbors again lack this linguistic coherence.

4.3 RQ3: MODEL CONTEXTS

Explanation Insights

- **How GPT-2 Makes Linguistic Decisions:**
  - *Adjectives*: Relies on semantic context (e.g. to disambiguate “black”).
  - *Adpositions/Adverbs*: Sensitive to associated verbs (e.g. “back” → “dating”, “traced”).
  - *Gender Determiners/Pronouns*: Looks at gendered proper nouns in input.
  - *Numbers vs. Words*: Uses contextual cues like “age”, “least”, “consists”.
- **Error Analysis:**
  - Gender errors often due to conflicting gender cues in input.
  - Clustering helps explain these errors and model’s misalignment.

05

Discussion

5.1 IMPLICATIONS

Interpretability Advancements

- Contrastive explanations provide fine-grained insights into why language models choose one token over another.
- Better aligned with linguistic evidence than non-contrastive methods.
- Enable human users to better simulate and understand model decisions.

Model Analysis at Scale

- Aggregated contrastive explanations can reveal general model behavior (e.g., grammatical rules, context use).
- Clustering in explanation space uncovers linguistic distinctions that are not obvious in word embeddings.

5.2 PROS & CONS

Further Analysis of Method

- **Pros:**
  - Post-hoc methods; no need to retrain model or train a new one.
  - Intuitive explanations for why one output is preferred over another.
  - Scalable to many evaluation types: alignment, user studies, clustering.
  - Generalizes across different explanation methods (gradients, erasure).
- **Cons:**
  - Computationally intensive, especially erasure methods.
  - Requires strong NLP infrastructure (e.g., POS taggers, coref systems).

06

Conclusion

6.1 LIMITATIONS

Current Drawbacks in the Research

- **Methodological Constraints**
  - Hard to adapt to non-English languages due to lack of resources, tools, and annotated data.
  - Automatic extraction of evidence (e.g., grammatical cues) depends on NLP tools with non-perfect accuracy.
- **Generalizability**
  - Not all explanation techniques easily adapt to the contrastive setting.
  - Focused only on GPT-2/Neo and English in experiments; the broader applicability is not yet proven.

6.1 FUTURE WORK

Directions for Future Research

- **Scaling & Efficiency**
  - Optimize computation to make methods more efficient for real-time interpretability tools.
- **Broader Applications**
  - Apply contrastive explanations to other ML models and tasks and extend to non-English LMs and multilingual interpretability.
- **Theoretical Development**
  - Explore new contrastive formulations and investigate how contrastive signals can inform fairness, bias, and safety in LMs.

THANKS!

Any questions?

Presentation by: Maryam Rezaee

TGML Lab | Spring 1404

Sharif University of Technology

Under the supervision of

Dr. Fatemeh SeyyedSalehi