Interpreting Language Models with Contrastive Explanations

Kayo Yin[*]
University of California, Berkeley
kayoyin@berkeley.edu

Graham Neubig
Carnegie Mellon University
gneubig@cs.cmu.edu

Abstract

Model interpretability methods are often used to explain NLP model decisions on tasks such as text classification, where the output space is relatively small. However, when applied to language generation, where the output space often consists of tens of thousands of tokens, these methods are unable to provide informative explanations. Language models must consider various features to predict a token, such as its part of speech, number, tense, or semantics. Existing explanation methods conflate evidence for all these features into a single explanation, which is less interpretable for human understanding.

To disentangle the different decisions in language modeling, we focus on explaining language models *contrastively*: we look for salient input tokens that explain why the model predicted one token *instead of* another. We demonstrate that contrastive explanations are quantifiably better than non-contrastive explanations in verifying major grammatical phenomena, and that they significantly improve contrastive model simulatability for human observers. We also identify groups of contrastive decisions where the model uses similar evidence, and we are able to characterize what input tokens models use during various language generation decisions.[1]

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

Table 1: Explanations for the GPT-2 prediction given the input "Can you stop the dog from _____". 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. Non-contrastive explanations such as gradient × input (1) usually attribute the highest saliency to the token immediately preceding the prediction. Contrastive explanations (2, 3) give a more fine-grained and informative explanation on why the model predicted one token over another.

1 Introduction

Despite their success across a wide swath of natural language processing (NLP) tasks, neural language models (LMs) are often used as black boxes: *how* they make certain predictions remains obscure (Belinkov and Glass, 2019). This is in part due to the high complexity of the LM task itself, as well as that of the model architectures used to solve it.

We argue that this is also due to the fact that interpretability methods commonly used in NLP, such as gradient-based saliency maps (Li et al., 2016a; Sundararajan et al., 2017), are not as informative for LM predictions compared to other tasks like text classification. For example, to explain why an LM predicts *"barking"* given "Can you stop the dog from ____", we demonstrate in experiments that the input token preceding the prediction is often marked as the most influential token to the prediction (Table 1) by instance attribution methods. The preceding token is indeed highly important to determine *certain* features of the next token, ruling out words that would obviously violate syntax in that context (e.g. non "-ing" verbs in the given example). However, this does not explain why the model made other more subtle decisions, such as why it predicts *"barking"* instead of *"crying"* or *"walking"*, which are all plausible choices if we only look at the preceding token. In general, language modeling has a large output space and a high complexity compared to other NLP tasks; at each time step, the LM chooses one word out of all vocabulary items, and several linguistic distinctions come into play for each language model decision.

[*] Work done while at Carnegie Mellon University.

[1] Code and demo: https://github.com/kayoyin/interpret-lm.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 184–198
December 7-11, 2022 ©2022 Association for Computational Linguistics

To better explain LM decisions, we propose interpreting LMs with *contrastive explanations* (Lipton, 1990). Contrastive explanations aim to identify causal factors that lead the model to produce one output **instead of** another output. We believe that contrastive explanations are especially useful to handle the complexity and the large output space of language modeling. In Table 1, the second explanation suggests that the input word “dog” makes “barking” more likely than a verb not typical for dogs such as “crying”, and the third explanation suggests that the input word “stop” increases the likelihood of “barking” over a verb without negative connotations such as “walking”.

In this paper, we first extend three interpretability methods to compute contrastive explanations (§3). We then perform a battery of experiments aimed at examining to what extent these contrastive explanations are superior to their non-contrastive counterparts from various perspectives:

- RQ1: Are contrastive explanations better at identifying evidence that we believe, *a-priori*, to be useful to capture a variety of linguistic phenomena (§4)?
- RQ2: Do contrastive explanations allow human observers to better simulate language model behavior (§5)?
- RQ3: Are different types of evidence necessary to disambiguate different types of words, and does the evidence needed reflect (or uncover) coherent linguistic concepts (§6)?

2 Background

2.1 Model Explanation

Our work focuses on model explanations that communicate *why* a model made a certain prediction. Particularly, we focus on methods that compute **saliency scores** $S(x_i)$ over input features x_i to reveal which input tokens are most relevant for a prediction: the higher the saliency score, the more x_i supposedly contributed to the model output.

Despite a large body of literature examining input feature explanations for NLP models on tasks such as text classification (for a complete review see Belinkov and Glass (2019); Madsen et al. (2021)), or interpreting how language models use linguistic features such as syntax (Ravfogel et al., 2021; Finlayson et al., 2021), few works attempt to explain language modeling predictions (Wallace et al., 2019). Despite the importance of both language models and interpretability in the NLP literature, the relative paucity of work in this area may be somewhat surprising, and we posit that this may be due to the large output space of language models necessitating the use of techniques such as contrastive explanations, which we detail further below.

2.2 Contrastive Explanations

Contrastive explanations attempt to explain *why* given an input x the model predicts a **target** y_t *instead of* a **foil** y_f. Relatedly, **counterfactual** explanations explore *how* to modify the input x so that the model more likely predicts y_f instead of y_t (McGill and Klein, 1993).

While contrastive and counterfactual explanations have been explored to interpret model decisions (see Stepin et al. (2021) for a broad survey), they are relatively new to NLP and have not yet been studied to explain language models.

Recently, Jacovi et al. (2021) produce counterfactual explanations for text classification models by erasing certain features from the input and projecting the input representation to the “contrastive space” that minimally separates two decision classes. Then, they compare model probabilities before and after the intervention.

We, on the other hand, propose contrastive explanations for language modeling, where both the number of input factors and the output space are much larger. While we also use a counterfactual approach with erasure (§3.3), counterfactual methods may become intractable over long input sequences and a large foil space. We, therefore, also propose contrastive explanations using gradient-based methods (§3.1,§3.2) that measure the saliency of input tokens for a contrastive model decision.

3 Contrastive Explanations for Language Models

In this section, we describe how we extend three existing input saliency methods to the contrastive setting. These methods can also be easily adapted to tasks beyond language modeling, such as machine translation (Appendix A).

3.1 Gradient Norm

Simonyan et al. (2013); Li et al. (2016a) calculate saliency scores based on the norm of the gradient of the model prediction, such as the output logit, with respect to the input. Applying this method to LMs entails first calculating the gradient as follows:

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

where $\boldsymbol{x}$ is the input sequence embedding, y_t is the next token in the input sequence, $q(y_t|\boldsymbol{x})$ is the model output for the token y_t given the input $\boldsymbol{x}$.

Then, we obtain the saliency score for the input token x_i by taking the L1 norm:

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

We extend this method to the **Contrastive Gradient Norm** defined by:

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

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

where $q(y_f|\boldsymbol{x})$ is the model output for foil y_f given the input $\boldsymbol{x}$. This tells us how much an input token x_i influences the model to increase the probability of y_t while decreasing the probability of y_f.

3.2 Gradient × Input

For the gradient × input method (Shrikumar et al., 2016; Denil et al., 2014), instead of taking the L1 norm of the gradient, we take the dot product of the gradient with the input token embedding x_i:

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

We define the **Contrastive Gradient × Input**:

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

3.3 Input Erasure

Erasure-based methods measure how erasing certain parts of the input affects the output (Li et al., 2016b). This can be measured as the difference between the model output given the input $\boldsymbol{x}$ and given the input where x_i has been zeroed out, $\boldsymbol{x}_{\neg i}$:

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

We define the **Contrastive Input Erasure**:

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

This measures how much erasing x_i from the input makes the foil more likely and the target less likely.

Although erasure-based methods directly measure the change in the output due to a perturbation in the input, while gradient-based methods approximate this measurement, erasure is usually more computationally expensive due to having to run the model on all possible input perturbations.

4 Do Contrastive Explanations Identify Linguistically Appropriate Evidence?

First, we ask whether contrastive explanations are quantifiably better than non-contrastive explanations in identifying evidence that we believe *a priori* should be important to the LM decision. In order to do so, we develop a methodology in which we specify certain types of evidence that indicate how to make particular types of linguistic distinctions, and measure how well each variety of explanation method uncovers this specified evidence.

4.1 Linguistic Phenomena

As a source of linguistic phenomena to study, we use the BLiMP dataset (Warstadt et al., 2020). This dataset contains 67 sets of 1,000 pairs of minimally different English sentences that contrast in grammatical acceptability. An example of a linguistic paradigm may be *anaphor number agreement*, where an acceptable sentence is “Many teenagers were helping **themselves**.” and a minimally contrastive unacceptable sentence is “Many teenagers were helping **herself**.” because in the latter, the number of the reflexive pronoun does not agree with its antecedent.

From this dataset, we chose 12 paradigms belonging to 5 phenomena and created a set of rules to identify the input tokens that enforce grammatical acceptability. In the previous example, the anaphor agreement is enforced by the antecedent “teenagers”. We show examples for each linguistic phenomenon and its associated rule in Table 2.

Anaphor Agreement. The gender and number of a pronoun must agree with its antecedent. We implement the `coref` rule using spaCy (Honnibal and Montani, 2017) and NeuralCoref[2] to extract all input tokens coreferent with the target token.

Argument Structure. Certain arguments can only appear with certain verbs. For example, many action verbs must be used with animate objects. We implement the `main_verb` rule using spaCy to extract the main verb of the input sentence.

Determiner-Noun Agreement. Demonstrative determiners and the associated noun must agree. We implement the `det_noun` rule by generating the dependency tree using spaCy and extracting the determiner of the target noun.

[2] https://github.com/huggingface/neuralcoref

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.

NPI Licensing. Certain negative polarity items (NPI) are only allowed to appear in certain contexts, e.g. "never" appears on its own, while "ever" generally must be preceded by "not". In all of our examples with NPI licensing, the word "even" is an NPI that can appear in the acceptable example but not in the unacceptable example, so we create the `npi` rule that extracts this NPI.

Subject-Verb Agreement. The number of the subject and its verb must agree. We implement the `subj_verb` rule by generating the dependency tree using spaCy to extract the subject of the verb.

4.2 Alignment Metrics

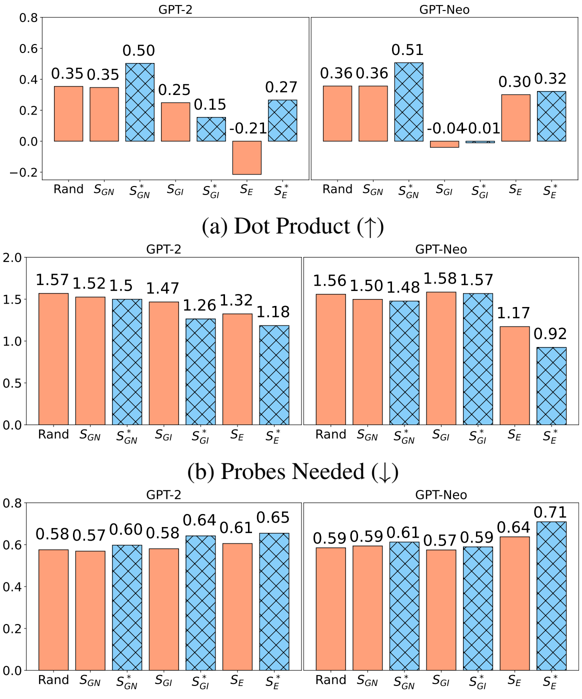
We use three metrics to quantify the alignment between an explanation and the known evidence enforcing a linguistic paradigm. The explanation is a vector $\boldsymbol{\mathcal{S}}$ of the same size as the input $\boldsymbol{x}$, where the i-th element $\mathcal{S}_i$ gives the saliency score of the input token x_i. The known evidence is represented with a binary vector $\boldsymbol{\mathcal{E}}$, also of same size as the input $\boldsymbol{x}$, where $\mathcal{E}_i = 1$ if the token x_i enforces a grammatical rule on the model decision.

Dot Product. The dot product $\boldsymbol{\mathcal{S}} \cdot \boldsymbol{\mathcal{E}}$ measures the sum of saliency scores of all input tokens that are part of the known evidence.

Probes Needed (Zhong et al., 2019; Yin et al., 2021b). We measure the number of tokens we need to probe, based on the explanation $\boldsymbol{\mathcal{S}}$, to find a token that is in the known evidence. This corresponds to the ranking of the first token x_i such that $G_i = 1$ after sorting tokens by descending saliency.

Mean Reciprocal Rank (MRR). We calculate the average of the inverse of the rank of the first token that is part of the known evidence if the tokens are sorted in descending saliency. This also corresponds to the average of the inverse of the probes needed for each sentence evaluated.

Dot Product and Probes Needed calculate alignment for each sentence, and we compute the average over all sentence-wise alignment scores for the alignment score over a linguistic paradigm. MRR calculates alignment over an entire paradigm.



(a) Dot Product (↑)

(b) Probes Needed (↓)

(c) Mean Reciprocal Rank (↑)

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

4.3 Results

We use GPT-2 (Radford et al., 2019) and GPT-Neo (Black et al., 2021) to extract explanations. GPT-2 is a large autoregressive transformer-based LM with 1.5 billion parameters and trained on 8 million web pages. GPT-Neo is a similar LM with 2.7 billion parameters and trained on The Pile (Gao et al., 2020) containing 825.18GB of largely English text. In addition to the explanation methods described above, we also set up a random baseline as a comparison, where we create a vector of the same size as explanations with values randomly sampled from a uniform distribution over $[0, 1)$.

In Figure 1, we can see that overall, contrastive explanations have a higher alignment with linguistic paradigms than their non-contrastive counterparts for both GPT-2 and GPT-Neo across the different metrics. Although non-contrastive explanations do not always outperform the random baseline, contrastive explanations have a better alignment with BLiMP than random vectors for most cases.

	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

Table 3: Alignment of GPT-2 explanations to known evidence on examples where the model makes a correct (left) and incorrect (right) prediction, according to dot product (DP), probes needed (PN), and mean reciprocal rank (MRR). Alignment scores that are better than the score for the analogous explanation method with the different contrastive setting are bolded.

In Table 3, we further examined alignment between model explanations and known evidence on instances where the model correctly allocates more probability to the acceptable token, or incorrectly selects the other token. On examples where the model makes an incorrect prediction, it is not clear whether non-contrastive or contrastive methods have better alignment. On examples where the model predicts correctly, contrastive explanations obtain better alignment than their non-contrastive counterparts for each explanation method and alignment metric.

In Figure 2, we see that for most explanation methods, the larger the distance between the known evidence and the target token, the larger the increase in alignment of contrastive explanations over non-contrastive explanations. This suggests that contrastive explanations particularly outperform non-contrastive ones when the known evidence is relatively further away from the target token, that is, contrastive explanations can better capture model decisions requiring longer-range context.

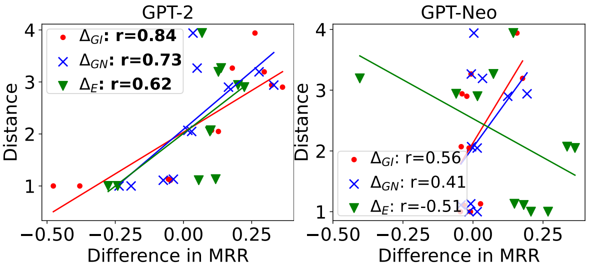


Figure 2: Scatter plot of the average distance of the known evidence to the target token across each linguistic paradigm against the difference in MRR scores between the contrastive and non-contrastive versions of each explanation method, with the Pearson correlation for each explanation method. Statistically 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.

In Appendix B, we also provide a table with the full alignment scores for each paradigm, explanation method, metric and model.

5 Do Contrastive Explanations Help Users Predict LM Behavior?

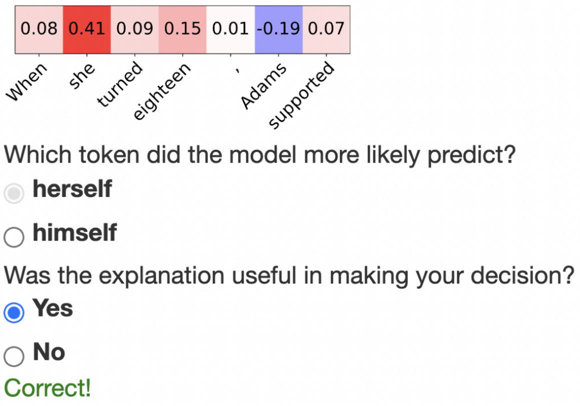
To further evaluate the quality of different explanation methods, we next describe methodology and experiments to measure to what extent explanations can improve the ability of users to predict the output of the model, namely model *simulatability* (Lipton, 2018; Doshi-Velez and Kim, 2017).

5.1 Study Setup

Our user study is similar in principle to previous works that measure model simulatability given different explanations (Chandrasekaran et al., 2018; Hase and Bansal, 2020; Pruthi et al., 2020). In our study (Figure 3), users are given the input of a GPT-2 model, two choices for the next token, and an explanation for the model output. They are asked to select which of the two choices is more likely the model output, then answer whether the explanation was useful in making their decision[3].

We compare the effect of having no explanation, explanations with Gradient × Input, Contrastive Gradient × Input, Erasure and Contrastive Erasure. We do not include Gradient Norm and Contrastive Gradient Norm because these methods do

[3] Although Hase and Bansal (2020) suggest not showing explanations for certain methods at test time due to potential for directly revealing the model output, this is less of a concern for saliency-based methods as their design makes it non-trivial to leak information in this way. We opt to show explanations to measure whether they sufficiently help the user make a prediction similar to the model on an individual example.



Which token did the model more likely predict?

○ herself

○ himself

Was the explanation useful in making your decision?

● Yes

○ No

Correct!

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

not provide information on directionality. For non-contrastive methods, we provide the explanation for why the model predicted a token. For contrastive methods, we provide the explanation for why the model predicted one token instead of another.

We include 20 pairs of highly confusable words for our study (Appendix C). 10 of these pairs are selected from BLiMP to reflect certain linguistic phenomena, and the other 10 word pairs are selected from pairs with the highest "confusion score" on WikiText-103 test split (Merity et al., 2016). We define confusion using the joint probability of a confusion from token a to b given a corpus X:

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

where x is a sentence in X, $\text{pos}(x)$ is the set of positions of tokens in x, N is the size of the corpus. The confusion from a to b is the sum of the probabilities assigned by the model to token b where token a is the ground truth, normalized by the number of sentences in the corpus.

The confusion score for word pair (a, b) is the minimum of confusion from a to b and vice-versa, to ensure that words are mutually confusable:

$$\begin{aligned} \mathcal{C}(a, b) = \min(&P(x_{true} = a, x_{model} = b), \\ &P(x_{true} = b, x_{model} = a)). \end{aligned}$$

We recruited 10 graduate students in machine learning (not authors of this paper) to perform the study. Each participant is given 10 different word pairs. For each word pair, one explanation method was chosen at random to generate the accompanying explanations, and the participant is given 40 sentences in a row. We balance the data so that there were an equal number of examples where the true output $x_t = a$ and $x_t = b$, and also by model correctness so that the model chooses the correct output 50% of the time, preventing users from guessing model behavior by selecting a certain token or the true token. In total, we obtain 4000 data points for model simulatability.

5.2 Results

In Table 4, we provide the results of our user study. For each explanation method evaluated, we computed the simulation accuracy over all samples (Acc.) as well as accuracy over samples where the model output is equal to the ground truth (Acc. Correct) and different from the ground truth (Acc. Incorrect). We also computed the percentage of explanations that users reported useful, as well as the simulation accuracy over samples where the user found the given explanation useful (Acc. Useful) and not useful (Acc. Not Useful).

To test our results for statistical significance and account for variance in annotator skill and word pair difficulty, we fitted linear mixed-effects models using Statsmodels (Seabold and Perktold, 2010) with the annotator and word pair as random effects, the explanation method as fixed effect, and the answer accuracy or usefulness as the dependent variable. In Appendix D we provide the results of the mixed-effects models we fitted.

		Acc.	Acc.		Acc.	Acc.
	Acc.	Correct	Incorrect	Useful	Useful	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 4: Simulation accuracy (%) in predicting GPT-2 outputs and subjective usefulness of explanations for various explanation methods. For each 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.

Accuracy First of all, users have the lowest accuracy in predicting LM outputs when no explanation is given, which suggests that all four types of explanations help users simulate model behavior. For both explanation methods, the contrastive setting leads to a significantly higher contrastive simulation accuracy than the non-contrastive setting.

We also examined examples where annotators incorrectly predict the model output, and for all types of explanations given, the most human errors are made in examples where there are no words in the input sentence that makes one word more likely than the other. Notably, the three word pairs with the lowest user accuracy are “son/brother”, “fast/super”, and “black/green”, which are often interchangeable.

Usefulness Contrastive explanations were also considered useful to users for model simulation significantly more often than non-contrastive explanations, with a particularly large gain in the erasure-based setting. Answer accuracy on samples where the users found the explanation useful is higher than the accuracy over all samples for each explanation method, which suggests that users can also identify useful explanations to some extent.

These results, on the whole, provide evidence that contrastive explanations help human observers simulate model predictions more accurately.

6 What Context Do Models Use for Certain Decisions?

Finally, we use contrastive explanations to discover how language models achieve various linguistic distinctions. We hypothesize that *similar evidence is necessary to disambiguate foils that are similar linguistically*. To test this hypothesis, we propose a methodology where we first represent each token by a vector representing its saliency map when the token is used as a foil in contrastive explanation of a particular target word. Conceptually, this vector represents the *type of context* that is necessary to disambiguate the particular token from the target. Next, we use a clustering algorithm on these vectors, generating clusters of foils where similar types of context are useful to disambiguate. We then verify whether we find clusters associated with *salient linguistic distinctions* defined a-priori. Finally, we inspect the mean vectors of explanations associated with foils in the cluster to investigate how models perform these linguistic distinctions.

6.1 Methodology

We generate contrastive explanations for the 10 most frequent words in WikiText-103 for each major part of speech as the target token, and use the 10,000 most frequent vocabulary items as foils. For each target y_t, we randomly select 500 sentences from WikiText-103 and obtain a sentence set X. For each foil y_f and each sentence $x_i \in X$, we generate a single contrastive explanation $e(x_i, y_t, y_f)$. Then, for each target y_t and foil y_f, we generate an aggregate explanation vector $e(y_t, y_f) = \bigoplus_{x_i \in X} e(x_i, y_t, y_f)$ by concatenating the single explanation vectors for each sentence in the corpus.

Then, for a given target y_t, we apply k-means clustering on the concatenated contrastive explanations across different foils y_f to cluster foils by explanation similarity. We use GPT-2 to extract all the contrastive explanations due to its better alignment with linguistic phenomena than GPT-Neo (§4). We only extract contrastive explanations with gradient norm and gradient×input due to the computational complexity of input erasure (§3.3).

In Table 5, we show examples of the obtained clusters. Foils in each cluster are sorted in descending frequency in training data. For the first foil in each cluster, we also retrieve its 20 nearest neighbors in the word embedding space according to Euclidean distance for comparison, to disentangle the effect of word embeddings from the effect of linguistic distinctions on foil clusters.

6.2 Foil Clusters

First, we verify that linguistically similar foils are indeed clustered together: we discover clusters relating to a variety of previously studied linguistic phenomena, a few of which we detail below and give examples in Table 5. Moreover, foil clusters reflect linguistic distinctions that are not found in the nearest neighbors of word embeddings. This suggests that the model use similar types of input features to make certain decisions.

Anaphor agreement: To predict anaphor agreement, models must contrast pronouns from other pronouns with different gender or number. We find that indeed, when the target is a pronoun, other pronouns of a different gender or number are often clustered together: when the target is a male pronoun, we find a cluster of female pronouns. The foil cluster containing “she” includes several types of pronouns that are all of the female gender. On the other hand, the nearest neighbors of “she” are mostly limited to subject and object pronouns, and they are of various genders and numbers.

Animacy: In certain verb phrases, the main verb enforces that the subject is animate. Reflecting this, when the target is an animate noun, inanimate

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 _____

Table 5: Examples of foil clusters obtained by clustering contrastive explanations of GPT-2. For each cluster, the 20 most frequent foils are shown, as well as the 20 nearest neighbors in the word embedding space of the first foil, and an example is included for the contrastive explanation of the target token vs. the underlined foil in the cluster. In each explanation, the two most salient input tokens are highlighted in decreasing intensity of red.

nouns form a cluster. While the foil cluster in Table 5 contains a variety of singular inanimate nouns, the nearest neighbors of “fruit” are mostly both singular and plural nouns related to produce.

Plurality: For determiner-noun agreement, singular nouns are contrasted with clusters of plural noun foils, and vice-versa. We find examples of clusters of plural nouns when the target is a singular noun, whereas the nearest neighbors of “tabs” are both singular and plural nouns. To verify subject-verb agreement, when the target is a plural verb, singular verbs are clustered together, but the nearest neighbors of “doesn” contain both singular and plural verbs, especially negative contractions.

6.3 Explanation Analysis Results

By analyzing the explanations associated with different clusters, we are also able to learn interesting properties of how GPT-2 makes certain predictions. We provide our full analysis results in Appendix E.

To distinguish between adjectives, the model often relies on input words that are semantically similar to the target (e.g. “relativity” to distinguish “black” from other colors). To contrast adpositions and adverbs from other words with the same POS, verbs in the input that are associated with the target word are useful: for example, the verbs “dating” and “traced” are useful when the target is “back”. To choose the correct gender for determiners, nouns and pronouns, the model often uses gendered proper nouns and pronouns in the input. To disambiguate numbers from non-number words, input words related to enumeration or measurement (e.g. “age”, “consists”, “least”) are useful.

Our analysis also reveals why the model may have made certain mistakes. For example, when the model generates a pronoun of the incorrect gender, it was often influenced by proper nouns and pronouns of a different gender in the input.

Overall, our methodology for clustering contrastive explanations provides an aggregate analysis of linguistic distinctions to understand general properties of language model decisions.

7 Conclusion and Future Work

In this work, we interpreted language model decisions using contrastive explanations by extending three existing input saliency methods to the contrastive setting. We also proposed three new methods to evaluate and explore the quality of contrastive explanations: an alignment evaluation to verify whether explanations capture linguistically appropriate evidence, a user evaluation to measure model simulatability of explanations, and a clustering-based aggregate analysis to investigate model properties using contrastive explanations.

We find that contrastive explanations are better aligned to known evidence related to major grammatical phenomena than their non-contrastive counterparts. Moreover, contrastive explanations allow better contrastive simulatability of models for users. From there, we studied what kinds of decisions require similar evidence and we used contrastive explanations to characterize how models make certain linguistic distinctions. Overall, contrastive explanations give a more intuitive and fine-grained interpretation of language models.

Future work could explore the application of these contrastive explanations to other machine learning models and tasks, extending other interpretability methods to the contrastive setting, as well as using what we learn about models through contrastive explanations to improve them.

8 Limitations

The experiments and methodology described in this paper have some limitations, notably their extensions to other explanation methods, other languages, and their resource requirements.

First, the applicability of the contrastive setting to other explanation methods may be limited. While extending gradient-based explanation methods to the contrastive setting is relatively straightforward (we can simply perform the same operations on the gradient over the difference between model probabilities for the target and foil tokens), it is nontrivial to design contrastive explanations based on other explanation methods such as attention-based input saliency.

Second, many of our experiments would not be easily reproduced in languages other than English that lack sufficient linguistic resources. All the experiments in our paper aimed at exploring the capabilities of contrastive explanations are performed using GPT-2 and GPT-Neo language models, that have been trained on large amounts of English data. To reproduce experiments in other languages, we would need a language model in the other language of sufficient power, which is not available for most languages. The experiments in Section 4 address only a subset of types of grammatical acceptability, and require a dataset of minimal pairs along different types of grammatical acceptability, which may not be available for most languages. Moreover, to automatically extract the expected evidence, we rely on core NLP tools such as coreference resolution, POS tagger and dependency parsers. Again, these tools are not available for most languages. Furthermore, the accuracy of the extracted evidence depends of the aforementioned tools, which have fairly high but not perfect accuracy. While our experiments are not easily extendable to languages other than English, our method itself of contrastive explanations is language agnostic and can be readily applied to models of any language.

In Section 5, we perform a human study to evaluate explanation methods. This evaluation method require human annotators and is therefore more resource intensive than automatic evaluation methods. We were motivated to perform this study nevertheless as model simulatability for human users is one important aspect of interpretability.

The experiments in Section 6 are also resource intensive. In total, we computed: 2 explanation methods × 8 parts of speech × 10 target words × 10,000 foils × 500 input sentences = 800,000,000 contrastive explanations. For this reason, we omitted the slower contrastive erasure explanation from this experiment, but generating all the contrastive explanations using the relatively faster explanation methods, then clustering them required about 48 hours of computation on 8 *RTX 8000* GPUs.

References

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. 2018. Do explanations make VQA models more predictable to a human? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1036–1042, Brussels, Belgium. Association for Computational Linguistics.

Misha Denil, Alban Demiraj, and Nando De Freitas. 2014. Extraction of salient sentences from labelled documents. *arXiv preprint arXiv:1412.6815*.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. 2021. Causal analysis of syntactic agreement mechanisms in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1828–1843, Online. Association for Computational Linguistics.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Peter Hase and Mohit Bansal. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, Online. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. 2021. Contrastive explanations for model interpretability. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1597–1611, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016a. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016b. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Peter Lipton. 1990. Contrastive explanation. *Royal Institute of Philosophy Supplement*, 27:247–266.

Zachary C. Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2021. Post-hoc interpretability for neural nlp: A survey. *arXiv preprint arXiv:2108.04840*.

Ann L McGill and Jill G Klein. 1993. Contrastive and counterfactual reasoning in causal judgment. *Journal of Personality and Social Psychology*, 64(6):897.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Danish Pruthi, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins, Zachary C Lipton, Graham Neubig, and William W Cohen. 2020. Evaluating explanations: How much do explanations from the teacher aid students? *arXiv preprint arXiv:2012.00893*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Shauli Ravfogel, Grusha Prasad, Tal Linzen, and Yoav Goldberg. 2021. Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. *arXiv preprint arXiv:2105.06965*.

Skipper Seabold and Josef Perktold. 2010. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, volume 57, page 61. Austin, TX.

Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. 2016. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Ilia Stepin, Jose M Alonso, Alejandro Catala, and Martín Pereira-Fariña. 2021. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9:11974–12001.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR.

Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. 2019. AllenNLP Interpret: A framework for explaining predictions of NLP models. In *Empirical Methods in Natural Language Processing*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Kayo Yin, Patrick Fernandes, André F. T. Martins, and Graham Neubig. 2021a. When does translation require context? a data-driven, multilingual exploration. *arXiv preprint arXiv:2109.07446*.

Kayo Yin, Patrick Fernandes, Danish Pruthi, Aditi Chaudhary, André F. T. Martins, and Graham Neubig. 2021b. Do context-aware translation models pay the right attention? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 788–801, Online. Association for Computational Linguistics.

Ruiqi Zhong, Steven Shao, and Kathleen R. McKeown. 2019. Fine-grained sentiment analysis with faithful attention. *CoRR*, abs/1908.06870.

A Contrastive Explanations for Neural Machine Translation (NMT) Models

A.1 Extending Contrastive Explanations to NMT

Machine translation can be thought of as a specific type of language models where the model is conditioned on both the source sentence and the partial translation. It has similar complexities as monolingual language modeling that make interpreting neural machine translation (NMT) models difficult. We therefore also extend contrastive explanations to NMT models.

We compute the contrastive gradient norm saliency for an NMT model by first calculating the gradient over the encoder input (the source sentence) and over the decoder input (the partial translation) as:

$$g^*(x_i^e) = \nabla_{x_i^e} \Big(q(y_t|\boldsymbol{x^e}, \boldsymbol{x^d}) - q(y_f|\boldsymbol{x^e}, \boldsymbol{x^d}) \Big)$$

$$g^*(x_i^d) = \nabla_{x_i^d} \Big(q(y_t|\boldsymbol{x^e}, \boldsymbol{x^d}) - q(y_f|\boldsymbol{x^e}, \boldsymbol{x^d}) \Big)$$

where $\boldsymbol{x^e}$ is the encoder input, $\boldsymbol{x^d}$ is the decoder input, and the other notations follow the ones in §3.1.

Then, the contrastive gradient norm for each x_i^e and x_i^d are:

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

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

Similarly, the contrastive gradient × input are:

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

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

We define the input erasure for each x_i^e and x_i^d as:

$$S^*_E(x_i^e) = \Big(q(y_t|\boldsymbol{x^e}, \boldsymbol{x^d}) - q(y_t|\boldsymbol{x^e_{\neg i}}, \boldsymbol{x^d}) \Big) - \Big(q(y_f|\boldsymbol{x^e}, \boldsymbol{x^d}) - q(y_f|\boldsymbol{x^e_{\neg i}}, \boldsymbol{x^d}) \Big)$$

$$S^*_E(x_i^d) = \Big(q(y_t|\boldsymbol{x^e}, \boldsymbol{x^d}) - q(y_t|\boldsymbol{x^e}, \boldsymbol{x^d_{\neg i}}) \Big) - \Big(q(y_f|\boldsymbol{x^e}, \boldsymbol{x^d}) - q(y_f|\boldsymbol{x^e}, \boldsymbol{x^d_{\neg i}}) \Big)$$

A.2 Qualitative Results

In Table 6, we provide examples of non-contrastive and contrastive explanations for NMT decisions. We use MarianMT (Junczys-Dowmunt et al., 2018) with pre-trained weights from the model trained to translate from English to Romance languages[4] to extract explanations. Each example reflects a decision associated with one of the five types of linguistic ambiguities during translation identified in Yin et al. (2021a).

[4] https://github.com/Helsinki-NLP/Tatoeba-Challenge/blob/master/models/eng-roa/README.md

Why did the model predict "il" ?
en: I ordered a new vase and it arrived today
fr: J ’ ai commandé un nouveau vase et
Why did the model predict "il" instead of "elle" ?
en: I ordered a new vase and it arrived today
fr: J ’ ai commandé un nouveau vase et

2. Why did the model predict "votre" ?
en: You cannot bring your dog here .
fr: Vous ne pouvez pas amener
Why did the model predict "votre" instead of "ton" ?
en: You cannot bring your dog here .
fr: Vous ne pouvez pas amener

3. Why did the model predict "apprenais" ?
en: I liked school because I learned a lot there .
fr: J ’ aimais l ’ école parce que j ’
Why did the model predict "apprenais" instead of "ai" ?
en: I liked school because I learned a lot there .
fr: J ’ aimais l ’ école parce que j ’

4. Why did the model predict "sais" ?
en: They know what to do , I don ’ t .
fr: Ils savent quoi faire , je ne
Why did the model predict "sais" instead of "veux" ?
en: They know what to do , I don ’ t .
fr: Ils savent quoi faire , je ne

5. 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

Table 6: Examples of non-contrastive and contrastive explanations for NMT models translating from English to French using input × gradient. Input tokens that are measured to raise or lower the probability of each decision are in red and blue respectively, and those with little influence are in white.

In the first example, the model must translate the gender neutral English pronoun “it” into the masculine French pronoun “il”. In both non-contrastive and contrastive explanations, the English antecedent “vase” influences the model to predict “il”, however to disambiguate “il” from the feminine pronoun “elle”, the model also relies on the french antecedent and its masculine adjective “nouveau vase”.

In the second example, the model must translate “your” with the formality level consistent with the partial translation. While in the non-contrastive explanation, only tokens in the source sentence are salient which do not explain the model’s choice of formality level, in the contrastive explanation, other French words in the polite formality level such as “Vous” and “pouvez” are salient.

In the third example, the model must translate “learned” using the verb form that is consistent with the partial translation. Similarly to the previous example, only the contrastive explanation contains salient tokens in the same verb from as the target token such as “aimais”.

In the fourth example, the model needs to resolve the elided verb in “I don’t ~~know~~” to translate into French. The contrastive explanation with a different verb as a foil shows that the elided verb in the target side makes the correct verb more likely than another verb.

In the fifth example, the model must choose the translation that is lexically cohesive with the partial translation, where “carnet” refers to a book with paper pages and “ordinateur” refers to a computer notebook. In the non-contrastive explanation, the word “notebook” and the target token preceding the prediction are the most salient. In the contrastive explanation, the word “carnet” in the partial translation also becomes salient.

B Alignment of Contrastive Explanations to Linguistic Paradigms

In Table 7, we present the full alignment scores of contrastive explanations from GPT-2 and GPT-Neo models with the known evidence to disambiguate linguistic paradigms in the BLiMP dataset.

C Highly Confusable Word Pairs

In Table 8, we provide the list of contrastive word pairs used in our human study for model simulatability (§5). The first 10 pairs are taken from BLiMP linguistic paradigms and we provide the associated

			GPT-2			GPT-Neo		
Paradigm	Dist	Explanation	Dot Product (↑)	Probes Needed (↓)	MRR (↑)	Dot Product (↑)	Probes Needed (↓)	MRR (↑)
anaphor_gender_agreement	2.94	Random	0.528	0.706	0.718	0.548	0.618	0.762
		S_{GN}	0.429	1.384	0.478	0.480	0.828	0.622
		S^*_{GN}	**0.834**	**0.472**	**0.809**	**0.785**	**0.432**	**0.815**
		S_{GI}	**0.078**	1.402	0.468	**-0.054**	**0.526**	**0.786**
		S^*_{GI}	-0.019	**0.502**	**0.791**	-0.133	0.684	0.747
		S_E	-0.350	0.564	0.764	**0.645**	**0.078**	**0.963**
		S^*_E	**0.603**	**0.090**	**0.964**	0.637	0.156	0.903
anaphor_number_agreement	2.90	Random	0.554	0.666	0.741	0.568	0.598	0.756
		S_{GN}	0.463	1.268	0.512	0.508	0.784	0.639
		S^*_{GN}	**0.841**	**0.702**	**0.677**	**0.816**	**0.524**	**0.763**
		S_{GI}	0.084	1.346	0.497	-0.095	**0.510**	**0.797**
		S^*_{GI}	0.084	**0.408**	**0.860**	**-0.068**	0.636	0.775
		S_E	-0.349	0.704	0.728	0.618	0.128	0.940
		S^*_E	**0.604**	**0.136**	**0.951**	**0.666**	**0.106**	**0.956**
animate_subject_passive	3.27	Random	0.155	2.940	0.378	0.150	2.976	0.379
		S_{GN}	0.211	1.080	0.699	0.236	**0.828**	**0.727**
		S^*_{GN}	**0.463**	**0.754**	**0.749**	**0.452**	0.862	0.721
		S_{GI}	0.016	4.004	0.233	**0.020**	**2.780**	**0.416**
		S^*_{GI}	**0.069**	**2.782**	**0.412**	0.016	2.844	0.409
		S_E	-0.036	3.214	0.362	**0.168**	**2.024**	0.444
		S^*_E	**0.125**	**2.122**	**0.500**	0.123	2.120	**0.517**
determiner_noun_agreement_1	1.00	Random	0.208	2.202	0.449	0.207	2.142	0.461
		S_{GN}	0.239	**1.320**	**0.598**	0.150	2.954	0.287
		S^*_{GN}	**0.275**	2.680	0.406	**0.258**	**2.906**	**0.302**
		S_{GI}	**0.560**	**0.038**	**0.983**	**-0.042**	**2.384**	**0.380**
		S^*_{GI}	0.162	1.558	0.603	-0.056	2.554	0.371
		S_E	0.022	**1.150**	**0.604**	0.234	1.290	0.543
		S^*_E	**0.031**	2.598	0.363	**0.362**	**0.612**	**0.811**
determiner_noun_agreement_irregular_1	1.00	Random	0.198	2.248	0.437	0.202	2.110	0.456
		S_{GN}	0.236	**1.228**	**0.616**	0.160	**2.716**	**0.324**
		S^*_{GN}	**0.286**	2.578	0.380	**0.266**	2.826	0.310
		S_{GI}	**0.559**	**0.034**	**0.984**	**-0.035**	**2.160**	**0.419**
		S^*_{GI}	0.046	2.038	0.507	-0.046	2.428	0.374
		S_E	0.020	**1.082**	**0.628**	0.205	1.360	0.548
		S^*_E	**0.026**	2.502	0.352	**0.306**	**0.784**	**0.755**
determiner_noun_agreement_with_adjective_1	2.05	Random	0.167	2.672	0.406	0.168	2.672	0.405
		S_{GN}	0.118	3.914	0.237	0.120	3.902	0.230
		S^*_{GN}	**0.210**	**3.532**	**0.267**	**0.228**	**3.814**	**0.245**
		S_{GI}	0.118	2.426	0.354	**-0.010**	**2.736**	**0.356**
		S^*_{GI}	**0.141**	**2.012**	**0.482**	-0.051	2.950	0.342
		S_E	0.042	1.730	0.583	0.092	2.748	0.333
		S^*_E	**0.305**	**1.084**	**0.680**	**0.260**	**1.176**	**0.697**
determiner_noun_agreement_with_adj_irregular_1	2.07	Random	0.167	2.620	0.401	0.158	2.820	0.392
		S_{GN}	0.116	3.920	0.240	0.125	**3.620**	**0.248**
		S^*_{GN}	**0.205**	**3.664**	**0.256**	**0.228**	3.718	0.243
		S_{GI}	0.106	2.620	0.345	**-0.007**	**2.754**	**0.358**
		S^*_{GI}	**0.111**	**2.244**	**0.448**	-0.047	3.126	0.316
		S_E	0.048	1.688	0.586	0.103	2.644	0.347
		S^*_E	**0.313**	**1.024**	**0.686**	**0.263**	**1.066**	**0.683**
npi_present_1	3.19	Random	0.336	1.080	0.604	0.350	0.984	0.632
		S_{GN}	0.294	1.160	0.510	0.376	0.454	0.778
		S^*_{GN}	**0.456**	**0.450**	**0.787**	**0.449**	**0.382**	**0.812**
		S_{GI}	0.100	1.374	0.463	-0.160	1.288	0.575
		S^*_{GI}	**0.144**	**0.570**	**0.759**	**0.202**	**0.766**	**0.752**
		S_E	-0.336	1.514	0.556	**0.624**	**0.086**	**0.960**
		S^*_E	**0.160**	**0.902**	**0.684**	0.062	1.204	0.556
distractor_agreement_relational_noun	3.94	Random	0.230	1.936	0.494	0.227	2.106	0.463
		S_{GN}	0.266	1.199	0.584	0.269	**0.965**	0.646
		S^*_{GN}	**0.408**	**1.092**	**0.619**	**0.392**	1.000	**0.649**
		S_{GI}	0.044	2.291	0.369	-0.066	2.326	0.434
		S^*_{GI}	**0.223**	**1.057**	**0.631**	**0.051**	**1.383**	**0.591**
		S_E	-0.023	1.922	0.434	0.120	2.007	0.400
		S^*_E	**0.190**	**1.709**	**0.502**	**0.186**	**1.617**	**0.544**
irregular_plural_subject_verb_agreement_1	1.11	Random	0.561	0.539	0.760	0.545	0.494	0.769
		S_{GN}	0.652	**0.242**	**0.917**	0.610	**0.348**	**0.860**
		S^*_{GN}	**0.676**	0.315	0.843	**0.644**	0.376	0.817
		S_{GI}	**0.590**	**0.253**	**0.912**	**0.067**	**0.472**	**0.783**
		S^*_{GI}	0.348	0.298	0.864	0.021	0.489	0.750
		S_E	-0.570	0.787	0.617	-0.021	0.893	0.553
		S^*_E	**0.264**	**0.635**	**0.673**	**0.267**	**0.584**	**0.734**
regular_plural_subject_verb_agreement_1	1.13	Random	0.694	0.316	0.853	0.693	0.336	0.849
		S_{GN}	0.740	**0.194**	**0.946**	0.724	**0.268**	**0.906**
		S^*_{GN}	**0.756**	0.251	0.909	**0.747**	0.274	0.898
		S_{GI}	**0.748**	**0.202**	**0.944**	-0.039	0.333	0.852
		S^*_{GI}	0.371	0.242	0.889	**0.039**	**0.262**	**0.879**
		S_E	-0.614	0.610	0.718	0.303	0.632	0.694
		S^*_E	**0.584**	**0.353**	**0.836**	**0.568**	**0.313**	**0.842**

Table 7: Alignment of GPT-2 and GPT-Neo explanations with BLiMP. Scores better than their (non-)contrastive counterparts are bolded. “Dist” gives the average distance from the target to the important context token.

unique identifier for each pair. The last 10 pairs are chosen from word pairs with the highest confusion score.

Word 1	Word 2	BLiMP UID
actor	actress	anaphor_gender_agreement
herself	himself	anaphor_gender_agreement
themselves	herself	anaphor_number_agreement
women	pictures	animate_subject_passive
boy	dog	animate_subject_passive
cat	cats	determiner_noun_agreement_1
is	are	irregular_plural_subject_verb_agreement_1
has	have	regular_plural_subject_verb_agreement_1
him	himself	principle_A_domain_1
he	who	wh_island
Word 1	Word 2	Confusion Score
black	green	0.0008
Bruce	Beth	0.0021
fast	super	0.0011
health	hospital	0.0012
red	bright	0.0007
snow	winter	0.0005
son	brother	0.0027
summer	winter	0.0003
white	blue	0.0034
wine	grape	0.0106

Table 8: List of highly confusable words pairs chosen for our user study.

Dependent Variable	Intercept	Effect	P-Value
Accuracy	0.624	0.015	0.050
Acc. Correct	0.744	0.026	0.005
Acc. Incorrect	0.530	-0.010	0.460
Useful	0.570	0.063	0.000
Acc. Useful	0.677	-0.020	0.513
Acc. Useful	0.450	-0.009	0.444

Table 9: The dependent variables, intercepts, the effect of the explanation method on the dependent variable and its p-value in the linear mixed-effects models fitted to model simulatability results.

D Mixed Effects Models Results

In Table 9, we show the results of fitting linear mixed-effects models to the results of our user study for model simulatability (§5).

E Analysis of Foil Clusters

In Figure 5, we give a few examples of clusters and explanations we obtain for each part of speech. For each part of speech, we describe our findings in more detail in the following.

Adjectives. When the target word is an adjective, other foil adjectives that are semantically similar to the target are often clustered together. For example, when the target is *"black"*, we find one cluster with various color adjectives, and we also find a different cluster with various adjectives relating to the race or nationality of a person.

We find that to distinguish between different adjectives, input words that are semantically close to the correct adjective are salient. For example to disambiguate the adjective *"black"* from other colors, words such as *"venom"* and *"relativity"* are important.

Adpositions. When the target is an adposition, other adpositions are often in the same cluster.

To distinguish between different adpositions, the verb associated with the adposition is often useful to the LM. For example, when the target word is *"from"*, verbs such as *"garnered"* and *"released"* helps the model distinguish the target from other adpositions that are less commonly paired with these verbs (e.g. *"for"*, *"of"*). As another example, for the target word *"for"*, verbs that indicate a long-lasting action such as *"continue"* and *"lived"* help the model disambiguate.

Adverbs. When the target is an adverb, other adverbs are often clustered together. Sometimes, when the target is a specific type of adverb, such as an adverb of place, we can find a cluster with other adverbs of the same type.

Similarly to adpositions, LMs often use the verb associated with the target adverb to contrast it from other adverbs. For example, the verbs *"dating"* and *"traced"* are useful when the target is *"back"*, and the verbs *"torn"* and *"lower"* are useful when the target is *"down"*.

Determiners. Other determiners are often clustered together when the target is a determiner. Particularly, when the target is a possessive determiner, we find clusters with other possessive determiners, and when the target is a demonstrative determiner, we find clusters with demonstrative determiners.

When the determiner is a gendered possessive determiner such as *"his"*, proper nouns of the same gender, such as *"John"* and *"George"*, are often useful. For demonstrative determiners, such as *"this"*, verbs that are usually associated with a targeted object, such as *"achieve"* and *"angered"* are useful.

Nouns. When the target noun refers to a person, for example, *"girl"*, foil nouns that also refer to a person form one cluster (e.g. *"woman", "manager", "friend"*), commonly male proper nouns form another (e.g. *"Jack", "Robin", "James"*), commonly female proper nouns form another (e.g. *"Sarah", "Elizabeth", "Susan"*), and inanimate objects form a fourth (e.g. *"window", "fruit", "box"*).

When the target noun is an inanimate object, there are often two notable clusters: a cluster with singular inanimate nouns and a cluster with plural inanimate nouns. This suggests how clustering foils by explanations confirm that certain grammatical phenomena require similar evidence for disambiguation; in this case, determiner-noun agreement.

To predict a target animate noun such as *"girl"* instead of foil nouns that refer to a non-female or older person, input words that are female names (e.g. *"Meredith"*) or that refer to youth (e.g. *"young"*) are useful. To disambiguate from male proper nouns, input words that refer to female people (e.g. *"Veronica", "she"*) or adjectives related to the target (e.g. *"tall"*) influence the model to generate a female common noun. To disambiguate from female proper nouns, adjectives and determiners are useful. To disambiguate from inanimate objects, words that describe a human or a human action (e.g. *"delegate", "invented"*) are useful.

To predict a target inanimate noun such as *"page"* instead of nouns that are also singular, input words with similar semantics are important such as *"sheet" and "clicking"* are important. For plural noun foils, the determiner (e.g. "a") is important.

Numbers. When the target is a number, non-number words often form one cluster and other numbers form another cluster.

To disambiguate numbers from non-number words, input words related to enumeration or measurement are useful (e.g. "age", "consists", "least"). To disambiguate words like *"hundred"* and *"thousand"* from other numbers such as *"20"* or *"five"*, input words used for counting (e.g. *"two", "several"*) are useful, because *"hundred"*s are countable in English (i.e. "*two* hundreds", "*several* hundreds").

Pronouns. When the target word is a gendered pronoun, foil pronouns of a different gender from the target form one cluster, foils with proper nouns of a different gender form a second cluster, and foils with proper nouns of the same gender as the target form a third cluster. This shows that the model uses similar evidence to make decisions to verify anaphor gender agreement. We also did not find foil clusters associated with distinguishing the number of the pronoun: often, these decisions follow directly from deciding between a pronoun and a proper noun, or deciding between a male and female pronoun.

To disambiguate a gendered pronoun such as such as *"he"*, from pronouns or proper nouns with different genders (e.g. *"she"* or *"Anna"*), proper nouns of the same gender as the target (e.g. *"James"*) and other gendered pronouns or determiners (e.g. *"his"*) are useful. To disambiguate from proper nouns of the same gender as the target, interestingly, the same proper noun as the foil appearing in the input is positively salient; GPT-2 is often influenced by previously appearing proper nouns to generate a pronoun instead.

Verbs. When the target word is a verb, foil verbs that have a different verb form are often clustered together. This suggests that the model uses similar input features to verify subject-verb agreement.

When the target verb is in present participle form, auxiliary verbs in the input are useful (e.g. *"is", "been"*) to distinguish from verbs in other forms. Similarly, when the target verb is in infinitive form, verbs in the same compound as the target verb are important.