Evaluating Saliency Methods for Neural Language Models

Shuoyang Ding **Philipp Koehn**
Center for Language and Speech Processing
Johns Hopkins University
{dings, phi}@jhu.edu

Abstract

Saliency methods are widely used to interpret neural network predictions, but different variants of saliency methods often disagree even on the interpretations of the same prediction made by the same model. In these cases, how do we identify when are these interpretations trustworthy enough to be used in analyses? To address this question, we conduct a comprehensive and quantitative evaluation of saliency methods on a fundamental category of NLP models: neural language models. We evaluate the quality of prediction interpretations from two perspectives that each represents a desirable property of these interpretations: *plausibility* and *faithfulness*. Our evaluation is conducted on four different datasets constructed from the existing human annotation of syntactic and semantic agreements, on both sentence-level and document-level. Through our evaluation, we identified various ways saliency methods could yield interpretations of low quality. We recommend that future work deploying such methods to neural language models should carefully validate their interpretations before drawing insights.

1 Introduction

While neural network models for Natural Language Processing (NLP) have recently become popular, a general complaint is that their internal decision mechanisms are hard to understand. To alleviate this problem, recent work has deployed interpretation methods on top of the neural network models. Among them, there is a category of interpretation methods called saliency method that is especially widely adopted (Li et al., 2016a,b; Arras et al., 2016, 2017; Mudrakarta et al., 2018; Ding et al., 2019). At a very high level, these methods assign an importance score to each feature in the input feature set F, regarding a specific prediction y made by a neural network model M. Such feature importance scores can hopefully shed light on the neural network models' internal decision mechanism.

V	U.S.	companies	wanting	to	expand	in	Europe
SG	U.S.	companies	wanting	to	expand	in	Europe
IG	U.S.	companies	wanting	to	expand	in	Europe

Table 1: An example from our evaluation where different saliency methods assign different importance scores for the same model (Transformer language model) and the same next word prediction (*are*). V, SG and IG are different saliency methods (see Section 2). The tints of green and yellow mark the magnitude of positive and negative importance scores, respectively.

While analyzing saliency interpretations uncovers useful insights for their respective task of interest, different saliency methods often give different interpretations even when the internal decision mechanism remains the same (with F, y and M held constant), as exemplified in Table 1. Even so, most existing work that deploys these methods often makes an ungrounded assumption that a specific saliency method can reliably uncover the internal model decision mechanism or, at most, relies merely on *qualitative* inspection to determine their applicability. Such practice has been pointed out in Adebayo et al. (2018); Lipton (2018); Belinkov and Glass (2019) to be potentially problematic for model interpretation studies – it can lead to misleading conclusions about the deep learning model's reasoning process. On the other hand, in the context of NLP, the *quantitative* evaluation of saliency interpretations largely remains an open problem (Belinkov and Glass, 2019).

In this paper, we address this problem by building a comprehensive quantitative benchmark to evaluate saliency methods. Our benchmark focuses on a fundamental category of NLP models: neural language models. Following the concepts proposed by Jacovi and Goldberg (2020), our benchmark evaluates the credibility of saliency interpretations from two aspects: *plausibility* and *faithfulness*. In short, plausibility measures how much these interpretations align with basic human intuitions about

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5034–5052
June 6–11, 2021.

the model decision mechanism, while faithfulness measures how consistent the interpretations are regarding perturbations that are supposed to preserve the same model decision mechanism on either the input feature F or the model M.

With these concepts in mind, our main contribution is materializing these tests' procedure in the context of neural language modeling and building four test sets from existing linguistic annotations to conduct these tests. Our study covering SOTA-level models on three different network architectures reveals that saliency methods' applicability depends heavily on specific choices of saliency methods, model architectures, and model configurations. We suggest that future work deploying these methods to NLP models should carefully validate their interpretations before drawing conclusions.

This paper is organized as follows: Section 2 briefly introduces saliency methods; Section 3 describes the plausibility and faithfulness tests in our evaluation; Section 4 presents the datasets we built for the evaluation; Section 5 presents our experiment setup and results; Section 6 discusses some limitations and implications of the evaluation; Section 7 concludes the paper.

2 Saliency

The notion of *saliency* discussed in this paper is a category of neural network interpretation methods that interpret a specific prediction y made by a neural network model M, by assigning a distribution of importance $\Psi(F)$ over the input feature set F of the original neural network model.

The most basic and widely used method is to assign importance by the gradient (Simonyan et al., 2013), which we refer to as vanilla gradient method (V). For each $x \in F$, $\psi(x) = \frac{\partial p_y}{\partial x}$, while p_y is the score of prediction y generated by M. We also examine two improved version of gradient-based saliency: SmoothGrad (SG) (Smilkov et al., 2017) and Integrated Gradients (IG) (Sundararajan et al., 2017). SmoothGrad reduces the noise in vanilla gradient-based scores by constructing several corrupted instances of the original input by adding Gaussian noise, followed by averaging the scores. Integrated Gradients computes feature importance by computing a line integral of the vanilla saliency from a "baseline" point F_0 to the input F in the feature space. We refer the readers to the cited papers for details of these saliency methods.

There is a slight complication in the meaning of F when applying these methods in the context of NLP: all the methods above will generate one importance score for each dimension of the word embedding, but most applications of saliency to NLP want a word-level importance score. Hence, we need composition schemes to combine scores over word embedding dimensions into a single score for each word. In the rest of this paper, we assume the "features" in the feature set F are input words to the language model, and word-level importance scores are composed using the gradient · input scheme (Denil et al., 2014; Ding et al., 2019).[1]

3 Evaluation Paradigm

In this section, we first introduce the notion of plausibility and faithfulness in the context of neural network interpretations (following Jacovi and Goldberg (2020)), and then, respectively, introduce the test we adopt to evaluate them.

3.1 Plausibility

Concept An interpretation is plausible if it aligns with human intuitions about how a specific neural model makes decisions. For example, intuitively, an image classifier can identify the object in the image because it can capture some features of the main object in the image. Hence, a plausible interpretation would assign high importance to the area occupied by the main object. This idea of comparison with human-annotated ground-truth (often as "bounding-boxes" signaling the main object's area) is used by various early studies in computer vision to evaluate saliency methods' reliability (Jiang et al., 2013, *inter alia*). However, the critical challenge of such evaluations for neural language models is the lack of such ground-truth annotations.

Test To overcome this challenge, we follow Poerner et al. (2018) to construct ground-truth annotations from existing lexical agreement annotations. Consider, for example, the case of morphological number agreement. Intuitively, when the language model predicts a verb with a singular morphological number, the singular nouns in the prefix should be considered important features, and vice versa. Based on this intuition, we divide the nouns in the prefix into two different sets: the *cue* set $\mathcal{C}$, which shares the same morphological number as the verb in the sentence; and the *attractor* set $\mathcal{A}$, which has

[1] We also experimented with the vector norm Li et al. (2016a) scheme in our preliminary study, and we find it performing much worse. See details in Appendix B.1.

a different morphological number than the verb in the sentence.

Then, according to the prediction y made by the model M, the test will be conducted under one of the two following scenarios:

- *Expected*: when y is the verb with the correct number, the interpretation passes the test if $\max_{w \in \mathcal{C}} \psi(w) > \max_{w \in \mathcal{A}} \psi(w)$
- *Alternative*: when y is the verb with the incorrect number, the interpretation passes the test if $\max_{w \in \mathcal{C}} \psi(w) < \max_{w \in \mathcal{A}} \psi(w)$

However, this test has a flaw: while the evaluation criteria focus on a specific category of lexical agreement, the prediction of a word could depend on multiple lexical agreements simultaneously. To illustrate this point, consider the verb prediction following the prefix *"At the polling station people ..."*. Suppose the model M predicts the verb *vote*. One could argue that *people* is more important than *polling station* because it needs the subject to determine the morphological number of the verb. However, the semantic relation between *vote* and *polling station* is also important because that is what makes *vote* more likely than other random verbs, e.g. *sing*.

To minimize such discrepancy and constrain the scope of agreements used to make predictions, we draw inspiration from the previous work on representation probing and make adjustment to the model M we are evaluating on (Tenney et al., 2019a,b; Kim et al., 2019; Conneau et al., 2018; Adi et al., 2017; Shi et al., 2016). The idea is to take a language model that is trained to predict words (e.g., *vote* in the example above) and substitute the original final linear layer with a new linear layer (which we refer to as a *probe*) fine-tuned to predict a binary lexical agreement tag (e.g., `PLURAL`) corresponding to the word choice. By making this adjustment, the final layer extracts a subspace in the representation that is relevant to the prediction of particular lexical agreement during the forward computation, and reversely, filters out gradients that are irrelevant to the agreement prediction in the backward pass, creating an interpretation that is only subject to the same agreement constraints as to when the annotation for the test set is done.

Apart from the adjustment made on the model M above, we also extend Poerner et al. (2018) in the other two aspects: (1) we evaluate on one more lexical agreement: gender agreements between pronouns and referenced entities, and on both natural and synthetic datasets; (2) instead of evaluating on small models, we evaluate on large SOTA-level models for each architecture. We also show that evaluation results obtained on smaller models cannot be trivially extended to larger models.

3.2 Faithfulness

Concept An interpretation is faithful if the feature importance it assigns is consistent with the internal decision mechanism of a model. However, as Jacovi and Goldberg (2020) pointed out, the notion of "decision mechanism" lacks a standard definition and a practical way to make comparisons. Hence, as a proxy, we follow the working definition of faithfulness as proposed in their work, which states that an interpretation is faithful if the feature importance it assigns remains consistent with changes that should not change the internal model decision mechanism. Among the three relevant factors for saliency methods (prediction y, model M, and input feature set F), we focus on consistency upon changes in model M (model consistency) and input feature set F (input consistency).[2] Note that these two consistencies respectively correspond to assumptions 1 and 2 in the discussion of faithfulness evaluation in Jacovi and Goldberg (2020).

Model Consistency Test To measure model consistency, we propose to measure the consistency between feature importance $\Psi_M(F)$ and $\Psi_{M'}(F)$, which is respectively generated from the original model M and a smaller model M' that is trained by distilling knowledge from M. In this way, although M and M' have different architectures, M' is trained to mimic the behavior of M to the extent possible, and thus having similar underlying decision mechanisms.

Input Consistency Test To measure input consistency, we perform substitutions in the input and measure the consistency between feature importance $\Psi(F)$ and $\Psi(F')$, where F and F' are input features sets before/after the substitution. For example, the following prefix-prediction pairs should have the same feature importance distribution:

- *The **nun** bought the **son** a gift because (she...)*
- *The **woman** bought the **boy** a gift because (she...)*

We measure consistency by Pearson correlation between pairs of importance score over the input

[2] Although evaluating interpretation consistency over similar predictions y is also possible, it is not of interest as most applications expect different interpretations for different predictions.

feature set F for both tests. Also, note that although we can theoretically conduct faithfulness tests with any model M and any dataset, for the simplicity of analysis and data creation, we will use the same model M (with lexical agreement probes) and the same dataset as plausibility tests.

4 Data[3]

Following the formulation in Section 3, we constructed four novel datasets for our benchmark, as exemplified in Table 2. Two of the datasets are concerned with *number agreement* of a verb with its subject. The other two are concerned with *gender agreement* of a pronoun with its anteceding entity mentions. For each lexical agreement type, we have one *synthetic* dataset and one *natural* dataset. Both synthetic datasets ensure there is only one cue and one attractor for each test instance, while for natural datasets, there are often more than one.

For number agreement, our synthetic dataset is constructed from selected sections of **Syneval**, a targeted language model evaluation dataset from Marvin and Linzen (2018), where the verbs and the subjects could be easily induced with heuristics. We only use the most challenging sections where strongly interceding attractors are involved. Our natural dataset for this task is filtered from Penn Treebank (Marcus et al., 1993, **PTB**), including training, development, and test. We choose PTB because it offers not only human-annotated POS-tags necessary for benchmark construction but also dependent subjects of verbs for further analysis.

For gender agreement, our synthetic dataset comes from the unambiguous **Winobias** coreference resolution dataset used in Jumelet et al. (2019), and we only use the 1000-example subset where there is respectively one male and one female antecedent. Because this dataset is intentionally designed such that most humans will find pronouns of either gender equally likely to follow the prefix, no such pronoun gender is considered to be "correct". Hence, without loss of generality, we assign the female pronoun to be the expected case.[4] Our natural dataset for this task is filtered from **CoNLL**-2012 shared task dataset for coreference resolution (Pradhan et al., 2012, also including training, development, and test). The prefix of each test example covers a document-level context, which usually spans several hundred words.

Plausibility Test For number agreement, the cue set $\mathcal{C}$ is the set of all nouns that have the same morphological number as the verb. In contrast, the attractor set $\mathcal{A}$ is the set of all nouns with a different morphological number. For gender agreement, the cue set $\mathcal{C}$ is the set of all nouns with the same gender as the pronoun, while the attractor set $\mathcal{A}$ is the set of all nouns with a different gender.

Model Consistency Test No special treatment to data is needed for this test. We conduct model consistency tests on all datasets we built.

Input Consistency Test We recognize that generating interpretation-preserving input perturbations for natural datasets is quite tricky. Hence, unlike the model consistency test, we focus on the two synthetic datasets for faithfulness tests because they are generated from templates. As can be seen from the examples, when the nouns in the cue/attractor set are substituted while maintaining the lexical agreement, the underlying model decision mechanism should be left unchanged; hence they can be viewed as interpretation-preserving perturbations. We identified 24 and 254 such interpretation-preserving templates from our Syneval and Winobias dataset and generated perturbations pairs by combining the first example of each template with other examples generated from the same template.

5 Experiments

5.1 Setup

Interpretation Methods For SmoothGrad (SG), we set sample size $N = 30$ and sample variance σ^2 to be 0.15 times the L2-norm of word embedding matrix; for Integrated Gradients (IG), we use step size $N = 100$. These choices are made empirically and verified on a small held-out development set.

Interpreted Model Our benchmark covers three different neural language model architectures, namely LSTM (Hochreiter and Schmidhuber, 1997), QRNN (Bradbury et al., 2017) and Transformer (Vaswani et al., 2017; Baevski and Auli, 2019; Dai et al., 2019). All language models are trained on WikiText-103 dataset (Merity et al., 2017). For the first two architectures, we use the implementation as in `awd-lstm-lm` toolkit (Merity et al., 2018). For Transformer, we use the imple-

[3] More details on data filtering are in Appendix A.

[4] Note that this assumption will not change the interpretations we generate or the benchmark test conducted for interpretations, as we always interpret the argmax decision of the model, which is not affected by this assumption. It will only affect the breakdown of the result we report.

PTB	U.S. Trade Representative Carla Hills said the first dispute-settlement panel set up under the U.S.-Canadian " free trade " agreement has ruled that Canada 's restrictions on exports of Pacific salmon and herring (PLURAL...)
Syneval	the consultant that loves the parents (SINGULAR...)
CoNLL	Israeli Prime Minister Ehud Barak says he is freezing tens of millions of dollars in tax payments to the Palestinian Authority . Mr. Barak says he is withholding the money until the Palestinians abide by cease - fire agreements . Earlier Thursday Mr. Barak ruled out an early resumption of peace talks , even with the United States acting as intermediary . Eve Conette reports from Jerusalem . Defending what (MASCULINE...)
Winobias	The bride examined the son for injuries because (FEMININE...)

Table 2: Examples prefixes from the four evaluation datasets, followed by the probing tag prediction under the expected scenario. The cue and attractor sets are marked with solid Green and yellow, respectively.

	Number Agreement						**Gender Agreement**					
	PTB			**Syneval**			**CoNLL**			**Winobias**		
	all	**exp.**	**alt.**	**all**	**exp.**	**alt.**	**all**	**exp.**	**alt.**	**all**	**exp.**	**alt.**
Random	-	0.546	0.454	-	0.500	0.500	-	0.519	0.481	-	0.500	0.500
Nearest	-	0.502	0.498	-	0.140	0.860	-	0.038	0.962	-	0.500	0.500
LSTM		(0.858)	(0.142)		(0.596)	(0.404)		(0.730)	(0.270)		(0.584)	(0.416)
V	0.452	0.484	0.259	0.304	0.371	0.206	0.288	0.266	0.348	0.403	0.440	0.351
SG	0.780	0.805	**0.629**	**0.950**	**0.951**	**0.949**	**0.799**	**0.767**	**0.880**	**0.984**	**0.981**	**0.988**
IG	**0.816**	**0.856**	0.571	0.888	0.941	0.811	0.585	0.561	0.652	0.881	0.853	0.921
QRNN		(0.818)	(0.182)		(0.558)	(0.442)		(0.712)	(0.288)		(0.715)	(0.285)
V	0.463	0.501	0.289	0.511	0.536	0.480	0.669	0.638	0.546	0.242	0.269	0.175
SG	0.575	0.599	0.468	0.707	0.692	0.726	0.503	0.436	0.669	**0.790**	**0.801**	0.761
IG	**0.697**	**0.728**	**0.555**	**0.797**	**0.764**	**0.838**	**0.737**	**0.700**	**0.828**	0.768	0.730	**0.863**
Transformer		(0.919)	(0.081)		(0.594)	(0.406)		(0.761)	(0.239)		(0.219)	(0.781)
V	0.551	0.551	0.551	0.723	0.785	0.632	0.674	0.693	0.614	0.781	0.799	0.766
SG	**0.842**	**0.851**	**0.737**	**0.895**	**0.879**	0.920	**0.956**	**0.951**	**0.971**	**0.994**	**1.00**	**0.992**
IG	0.734	0.741	0.652	0.849	0.786	**0.940**	0.829	0.843	0.786	0.806	0.865	0.775

Table 3: Plausibility benchmark result. Each number is the fraction of cases the interpretation passes the benchmark test, while the numbers in brackets for each architecture are the fraction of times these scenarios occur for predictions generated by the corresponding model. Results from the best interpretation method for each architecture are boldfaced. The *exp.* and *alt.* columns are breakdown of evaluation results into expected scenarios and alternative scenarios as defined in Section 3. V, SG, IG stands for the vanilla saliency, SmoothGrad, and Integrated Gradients, respectively.

mentation in `fairseq` tookit (Ott et al., 2019).

For all the task-specific "probes", the fine-tuning is performed on examples extracted from WikiText-2 training data. A tuning example consists of an input prefix and a gold tag for the lexical agreement in both cases. For number agreement, we first run Stanford POS Tagger (Toutanova et al., 2003) on the data, and an example is extracted for each present tense verb and each instance of *was* or *were*. For gender agreement, an example is extracted for each gendered pronoun. During fine-tuning, we fix all the other parameters except the final linear layer. The final layer is tuned to minimize cross-entropy, with Adam optimizer (Kingma and Ba, 2015) and initial learning rate of $1e{-}3$ with `ReduceLROnPlateau` scheduler.

We follow the setup for DistillBERT (Sanh et al., 2019) for the distillation process involved during the model consistency test, which reduces the depth of models but not the width. For our LSTM (3 layers) and QRNN model (4 layers), the M' we distill is one layer shallower than the original model M. For our transformer model (16 layers), we distill a 4-layer M' largely due to memory constraints.

5.2 Main Results

Plausibility According to our plausibility evaluation result, summarized in Table 3, both SG and IG consistently perform better than the vanilla saliency method regardless of different benchmark datasets and interpreted models. However, the comparison between SG and IG interpretations varies depend-

ing on the model architecture and test sets.

Across different architectures, Transformer language model achieves the best plausibility except on the Syneval dataset. LSTM closely follows Transformer for most benchmarks, while the plausibility of the interpretation from QRNN is much worse. Another trend worth noting is that the gap between Transformer and the other two architectures is much larger on the CoNLL benchmark, which is the only test that involves interpreting document-level contexts. However, these architectures' prediction accuracy is similar, meaning that there is no significant modeling power difference for gender agreements in this dataset. We hence conjecture that the recurrent structure of LSTM and QRNN might diminish gradient signals with increasing time steps, which causes the deterioration of interpretation quality for long-distance agreements – a problem that Transformer is exempt from, thanks to the self-attention structure.

Faithfulness Table 4a shows the input consistency benchmark result. Firstly, it can be seen that the interpretations of LSTM and Transformer are more resilient to input perturbations than that of QRNN. This is the same trend as we observed for plausibility benchmark on these datasets. When comparing different saliency methods, we see that SG consistently outperforms for Transformer, but fails for the other two architectures, especially for QRNN. Also, note that achieving higher plausibility does not necessarily imply higher faithfulness. For example, compared to the vanilla saliency method, SG and IG almost always significantly improve plausibility but do not always improve faithfulness. This lack of improvement is different from the findings in computer vision (Yeh et al., 2019), where they show both SG and IG improve input consistency. Also, for LSTM, although SG works slightly better than IG in terms of plausibility, IG outperforms SG in terms of input consistency by a large margin.

Table 4b shows the model consistency benchmark result. One should first notice that model consistency numbers are lower than input consistency across the board, and the drop is more significant for LSTM and QRNN even though their student model is not as different as the Transformer model (<20% parameter reduction vs. 61%). As a result, there is a significant performance gap in terms of best model consistency results between LSTM/QRNN and Transformer. Note that, like in plausibility results, such gap is most notable on the CoNLL dataset. On the other hand, when comparing between saliency methods, we again see that SG outperforms for Transformer while failing most of the times for QRNN and LSTM.

5.3 Analysis

Plausibility vs. Faithfulness A natural question for our evaluation is how the property of plausibility and faithfulness interact with each other. Table 5 illustrates such interaction with qualitative examples. Among them, 1 and 2 are two cases where the plausibility and input faithfulness evaluation results do not correlate. In general, the interpretations in both cases are of low quality, but they also fail in different ways. In case 1, the interpretation assigns the correct relative ranking for the cue words and attractor words, but the importance of the words outside the cue/attractor set varies upon perturbation. On the other hand, in case 2, the importance ranking among features is roughly maintained upon perturbation, but the importance score assigned for both examples do not agree with the prediction interpreted (`FEMININE` tag) and thus can hardly be understood by humans. It should be noted that these defects can only be revealed when both plausibility and faithfulness tests for interpretations are deployed.

Case 3 shows a scenario where the saliency method yields very different interpretations for the same input/prediction pair, indicating that interpretations from this architecture/saliency method combination are subject to changes upon changes in the architecture configurations. Finally, in case 4, we see that an architecture/saliency method combination performing well in all tests yields stable interpretations that humans can easily understand.

Sensitivity to Model Configurations Our model faithfulness evaluation shows that variations in the model configurations (number of layers) could drastically change the model interpretation in many cases. Hence, we want to answer two analysis questions: (1) are these interpretations changing for the better or worse quality-wise with the distilled smaller models? (2) are there any patterns for such changes? Due to space constraints, we only show some analysis results for question (1) in Table 6. Overall, compared to the corresponding results in Table 3 (for plausibility) and Table 4a (for input faithfulness), the saliency methods we evaluated perform better with the

	Syneval		Winobias	
	exp.	alt.	exp.	alt.
LSTM				
V	0.532	0.533	0.447	0.447
SG	0.481	0.491	0.560	0.404
IG	**0.736**	**0.695**	**0.735**	**0.795**
QRNN				
V	0.226	0.223	**0.566**	0.566
SG	0.166	0.239	0.184	0.239
IG	**0.448**	**0.387**	0.499	**0.622**
Transformer				
V	0.367	0.375	0.545	0.545
SG	**0.604**	**0.627**	**0.775**	**0.752**
IG	0.521	0.480	0.542	0.494

(a) Input Consistency

	Number Agreement				Gender Agreement			
	PTB		Syneval		CoNLL		Winobias	
	exp.	alt.	exp.	alt.	exp.	alt.	exp.	alt.
LSTM								
V	0.325	0.324	0.370	0.370	**0.301**	**0.301**	0.082	0.082
SG	0.242	0.294	**0.453**	0.394	0.190	0.235	0.071	0.138
IG	**0.548**	**0.487**	0.439	**0.513**	0.256	0.275	**0.435**	**0.252**
QRNN								
V	0.208	0.207	0.228	0.229	0.147	0.147	0.212	0.212
SG	0.043	0.044	0.144	0.131	0.010	0.016	0.063	0.070
IG	**0.259**	**0.387**	**0.316**	**0.350**	**0.305**	**0.375**	**0.303**	**0.285**
Transformer								
V	0.160	0.160	0.219	0.219	0.289	0.289	0.104	0.104
SG	**0.584**	**0.584**	**0.598**	**0.570**	**0.688**	**0.693**	**0.656**	**0.581**
IG	0.239	0.294	0.450	0.413	0.219	0.277	0.310	0.291

(b) Model Consistency

Table 4: Faithfulness Benchmark Result. Each number is the average Pearson correlation computed on the corresponding dataset. Results from the best interpretation method for each architecture are boldfaced. Refer to the caption of Table 3 for other notations.

1a	QRNN+SG	The [grandmother] examined the (grandson) for injuries because
1b	QRNN+SG	The [sister] examined the (groom) for injuries because
2a	QRNN+V	The [grandmother] examined the (grandson) for injuries because
2b	QRNN+V	The [aunt] examined the (groom) for injuries because
3a	QRNN+SG	The [grandmother] examined the (grandson) for injuries because
3b	QRNN_distilled+SG	The [grandmother] examined the (grandson) for injuries because
4a	Transformer+SG	The [grandmother] examined the (grandson) for injuries because
4b	Transformer+SG	The [aunt] examined the (groom) for injuries because
4c	Transformer_distilled+SG	The [grandmother] examined the (grandson) for injuries because

Table 5: Examples from Winobias dataset for qualitative analysis. Cue words are marked with [] while attractor words are marked with (). The tints of green and yellow mark the magnitude of positive and negative importance scores, respectively. For all examples, the prediction interpreted is the FEMININE tag. 1 is a case with high plausibility and low input faithfulness; 2 is a case with low plausibility and high input faithfulness; 3 is a case with low model faithfulness; 4 is a case with high plausibility and high input/model faithfulness.

smaller distilled models. Most remarkably, we see a drastic performance improvement for QRNN, both in plausibility and faithfulness. For LSTM and Transformer, we observe an improvement for input faithfulness on Winobias and roughly the same performance for other tests.

As for the second question, we build smaller Transformer language models with various depth, number of heads, embedding size, and feed-forward layer width settings, while keeping other hyperparameters unchanged. Unfortunately, the trends are quite noisy and also heavily depends on the chosen saliency methods.[5] Hence, it is highly recommended that evaluation of saliency methods be conducted on the specific model configurations of interest, and trends of interpretation quality on a specific model configuration should not be over-generalized to other configurations.

Saliency vs. Probing Our evaluation incorporates probing to focus only on specific lexical agreements of interest. It should be pointed out that in the literature of representation probing, the method has always been working under the following assumption: when the model makes an expected-scenario ("correct") prediction, it is always referring to a *grammatical* cue, for example, the subject of the verb in the number agreement case. However, in our evaluation, we also observe some interesting phenomena in the interpretation of saliency methods that breaks the assumption, which is exemplified in Table 7. This calls for future work that aims to better understand language model behaviors by examining other possible cues used for

[5]Detailed discussion of these analyses is in Appendix B.2.

	Syneval			Winobias		
	all	**exp.**	**alt.**	**all**	**exp.**	**alt.**
best plausibility						
LSTM (SG)	0.945	0.922	0.973	0.948	0.950	0.904
QRNN (IG)	0.981	0.964	0.998	0.974	0.974	1.00
Transformer (SG)	0.917	0.908	0.929	0.997	1.00	0.996
best (input) faithfulness						
LSTM (IG)	–	0.628	0.739	–	0.820	0.769
QRNN (IG)	–	0.733	0.831	–	0.891	0.841
Transformer (SG)	–	0.569	0.581	–	0.932	0.912

Table 6: Plausibility & input faithfulness on synthetic datasets with distilled models. Only results for the interpretation method with best performance are shown. Refer to the caption of Table 3 for other notations.

V	"	The	[fact]	that	this	happened	two	(years)	ago	and	there	was	a	[recovery]
SG	"	The	[fact]	that	this	happened	two	(years)	ago	and	there	was	a	[recovery]
IG	"	The	[fact]	that	this	happened	two	(years)	ago	and	there	was	a	[recovery]

Table 7: A number agreement test case where the distilled Transformer model makes the correct prediction (singular) but all interpretation methods unanimously point to a singular noun that is not grammatical subject as the most salient cue for this prediction.

predictions made in representation probing under the validated cases where saliency methods could be reliably applied.

6 Discussion

Most existing work on evaluating saliency methods focuses only on computer vision models (Adebayo et al., 2020; Hooker et al., 2019; Adebayo et al., 2018; Heo et al., 2019; Ghorbani et al., 2019, *inter alia*). In the context of NLP, Poerner et al. (2018) is the first work to conduct such evaluations for NLP and the only prior work that conducts such evaluations for neural language models but has several limitations as we have already pointed out in Section 3. Arras et al. (2019); Atanasova et al. (2020); Hao (2020) conducted similar evaluations based on specifically designed diagnostic toy tasks and/or text classification, while Bastings and Filippova (2020) casted doubt on whether these conclusions could be generalized to sequence generation tasks. Li et al. (2020) evaluated various interpretation methods for neural machine translation models by building proxy models on only the top-k important input words as determined by the interpretation methods, but such evaluation requires generating interpretations for a large training set and hence is intractable for even mildly computationally-expensive methods such as SmoothGrad and Integrated Gradients. On a slightly different line, DeYoung et al. (2020) built a benchmark to evaluate a specific category of NLP models that generate rationales during predictions, which is a different path towards building explainable NLP models.

Our evaluation is not without its limitations. The first limitation, inherited from earlier work by Poerner et al. (2018), is that our plausibility test only concerns the words in cue/attractor sets rather than other words in the input prefix. Such limitation is inevitable because the annotations from which we build our ground-truth interpretations are only concerned with a specific lexical agreement. This limitation can be mitigated by combining plausibility tests with faithfulness tests, which concern all the input prefix words.

The second limitation is that the test sets used in these benchmarks need to be constructed in a case-to-case manner, according to the chosen lexical agreements and the input perturbations. While it is hard to create plausibility test sets without human interference, future work could explore automatic input consistency tests by utilizing adversarial input generation techniques in NLP (Alzantot et al., 2018; Cheng et al., 2019, 2020).

It should also be noted that while our work focuses on evaluating a specific category of interpretation methods for neural language models, our evaluation paradigm can be easily extended to evaluating other interpretation methods such as attention mechanism, and with other sequence models such as masked language models (e.g., BERT). We would also like to extend these evaluations beyond English datasets, especially to languages with richer morphological inflections.

7 Conclusion

We conduct a quantitative evaluation of saliency methods on neural language models based on the perspective of plausibility and faithfulness. Our evaluation shows that a model interpretation can either fail due to a lack of plausibility or faithfulness, and the interpretations are trustworthy only when they do well with both tests. We also noticed that the performance of saliency interpretations are generally sensitive to even minor model configuration changes. Hence, trends of interpretation quality on a specific model configuration should not be over-generalized to other configurations.

We want the community to be aware that saliency methods, like many other post-hoc interpretation methods, still do not generate trustworthy interpretations all the time. Hence, we recommend that adopting any model interpretation method as a source of knowledge about NLP models' reasoning process should only happen after similar quantitative checks as presented in this paper are performed. We also hope our proposed test paradigm and accompanied test sets provide useful guidance to future work on evaluations of interpretation methods. Our evaluation dataset and code to reproduce the analysis are available at https://github.com/shuoyangd/tarsius.

Acknowledgements

The authors would like to thank colleagues at CLSP and anonymous reviewers for feedback at various stages of the draft. This material is based upon work supported by the United States Air Force under Contract No. FA8750-19-C-0098. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force and DARPA.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 9525–9536.

Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. 2020. Debugging tests for model explanations. *CoRR*, abs/2011.05429.

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2016. Explaining predictions of non-linear classifiers in NLP. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 1–7, Berlin, Germany. Association for Computational Linguistics.

Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. Explaining recurrent neural network predictions in sentiment analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 159–168, Copenhagen, Denmark. Association for Computational Linguistics.

Leila Arras, Ahmed Osman, Klaus-Robert Müller, and Wojciech Samek. 2019. Evaluating recurrent neural network explanations. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 113–126, Florence, Italy. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Alexei Baevski and Michael Auli. 2019. Adaptive input representations for neural language modeling. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155, Online. Association for Computational Linguistics.

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

James Bradbury, Stephen Merity, Caiming Xiong, and Richard Socher. 2017. Quasi-recurrent neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4324–4333, Florence, Italy. Association for Computational Linguistics.

Yong Cheng, Lu Jiang, Wolfgang Macherey, and Jacob Eisenstein. 2020. AdvAug: Robust adversarial augmentation for neural machine translation. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5961–5970, Online. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Misha Denil, Alban Demiraj, and Nando de Freitas. 2014. Extraction of salient sentences from labelled documents. *CoRR*, abs/1412.6815.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Shuoyang Ding, Hainan Xu, and Philipp Koehn. 2019. Saliency-driven word alignment interpretation for neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 1–12, Florence, Italy. Association for Computational Linguistics.

Amirata Ghorbani, Abubakar Abid, and James Y. Zou. 2019. Interpretation of neural networks is fragile. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3681–3688. AAAI Press.

Yiding Hao. 2020. Evaluating attribution methods using white-box LSTMs. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 300–313, Online. Association for Computational Linguistics.

Juyeon Heo, Sunghwan Joo, and Taesup Moon. 2019. Fooling neural network interpretations via adversarial model manipulation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 2921–2932.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 9734–9745.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. 2013. Salient object detection: A discriminative regional feature integration approach. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 2083–2090.

Jaap Jumelet, Willem Zuidema, and Dieuwke Hupkes. 2019. Analysing neural language models: Contextual decomposition reveals default reasoning in number and gender assignment. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 1–11, Hong Kong, China. Association for Computational Linguistics.

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jierui Li, Lemao Liu, Huayang Li, Guanlin Li, Guoping Huang, and Shuming Shi. 2020. Evaluating explanation methods for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 365–375, Online. 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. *CoRR*, abs/1612.08220.

Zachary C. Lipton. 2018. The mythos of model interpretability. *Commun. ACM*, 61(10):36–43.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Comput. Linguistics*, 19(2):313–330.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2018. Regularizing and optimizing LSTM language models. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. Did the model understand the question? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1896–1906, Melbourne, Australia. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Nina Poerner, Hinrich Schütze, and Benjamin Roth. 2018. Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 340–350, Melbourne, Australia. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning - Proceedings of the Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes, EMNLP-CoNLL 2012, July 13, 2012, Jeju Island, Korea*, pages 1–40.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas. Association for Computational Linguistics.

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

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 3319–3328.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4593–4601.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 252–259.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. 2019. On the (in) fidelity and sensitivity of explanations. In *Advances in Neural Information Processing Systems*, pages 10965–10976.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

A Data Filtering Details

A.1 Penn Treebank (PTB)

A potential candidate for a test case is extracted every time a word with POS tag VBZ (Verb, 3rd person singular present) or VBP (Verb, non-3rd person singular present), or a copula that is among *is, are, was, were*, shows up. The candidate will then be filtered subjecting to the following criteria:

1. The prefix has at least one attractor word (a noun that has a different morphological number as the verb that is predicted). This is to ensure that evaluation could be conducted in the alternative scenario.

2. The verb cannot immediately follow its grammatical subject (note: it may still immediately follow a cue word that is not a grammatical subject). This is to ensure that the signal of the subject is not overwhelmingly strong compared to the attractors.

3. Not all attractors occur earlier 10 words than the grammatical subject. Same reason as the previous criteria.

Overall, we obtained 1448 test cases out of 49168 sentences in PTB (including train, dev, and test set). We lose a vast majority of sentences mostly because of the last two criteria.

A.2 Syneval

We use the following sections of the original data (followed by their names in the data dump, Marvin and Linzen, 2018):

- Agreement in a sentential complemenet: `sent_comp`
- Agreement across a prepositional phrase: `prep_anim` and `prep_inanim`
- Agreement across a subject relative clause: `subj_rel`
- Agreement across an object relative clause: `obj_rel_across_anim`, `obj_rel_across_inanim`, `obj_rel_no_comp_across_anim`, `obj_rel_no_comp_across_inanim`
- Agreement within an object relative clause: `obj_rel_within_anim`, `obj_rel_within_inanim`, `obj_rel_no_comp_within_anim`, `obj_rel_no_comp_within_inanim`

We select these sections because they all have strong interfering attractors or have cues that may potentially be mistaken as attractors. We obtained much fewer examples (6280) than the original data (249760) because lots of examples only differ in the verb or the object they use, which become duplicates when we extract prefix before the verb.

The original dataset does not come with cue/attractor annotations, but it can be easily inferred because they are generated by simple heuristics.

Note that most of these sections have only around 50% prediction accuracy with RNNs in the original paper. Our results on large-scale language models corroborate the findings in the original paper.

A.3 CoNLL

We use the dataset (Pradhan et al., 2012) with gold parses, entities mentions, and mention boundaries. A potential candidate for a test case is extracted every time a pronoun shows up. The male pronouns are *he, him, himself, his*, while the female pronouns are *she, her, herself, hers*. We don't include cases of epicene pronouns like *it, they*, etc. because they often involve tricky cases like entity mentions covering a whole clause. We break prefixes according to the document boundaries as provided in the original dataset unless the prefix is longer than 512 words, in which case we instead break at the nearest sentence boundary.

The annotation for this dataset does not cover the gender of entities. We are aware that the original shared task provides gender annotation, but to this day, the documentation for the data is missing and hence we cannot make use of this annotation. Hence, we instead used several heuristics to infer the gender of an entity mention, in descending order:

- If an entity mention and a pronoun have a coreference relationship, they should share the same gender.
- If an entity mention starts with "Mr." or "Mrs." or "Ms.", we assign the corresponding gender.
- If the entity mention has a length of two tokens, we assume it's a name and use gender inference tools[6] to guess its gender. Note that the gender

[6] `https://github.com/lead-ratings/gender-guesser`

guesser may also indicate that it's not able to infer the gender, in that case, we do not assign a gender.

- If a mention is co-referenced with another mention that is not a pronoun, they should also have the same gender.

Manual inspection of the resulting data indicates that the scheme above covers the gender of most entity mentions correctly. We hope that our dataset could be further perfected by utilizing higher quality annotation on entity genders.

Since each entity mention could span more than one word, we add all words within the span into their corresponding cue/attractor set. A tricky case is where two entity mention spans are nested or intersected. For the first case, we exclude a smaller span from the larger one to create two unintersected spans as the new span for the cue/attractor set. For the second case, we exclude the intersecting parts from both spans.

Finally, all candidates are filtered subject to the following two criteria:

1. The prefix should include one attractor entity.

2. The entity mention that is closest to the verb should be of different gender (either the opposite or epicene).

We obtained 586 document segments from the 2280 documents in the original data. As pointed out in Zhao et al. (2018), the CoNLL dataset is significantly biased towards male entity mentions. Nevertheless, our filtering scheme generated a relatively balanced test set: among 586 test cases, 258 are male pronouns, while 328 are female pronouns.

A.4 Winobias

We used the same data as the unambiguous coreference resolution dataset in Jumelet et al. (2019), which is in turn generated by a script from Zhao et al. (2018), except that we excluded cases where both nouns in the sentence are of the same gender. Similar to **Syneval** dataset, the cue and attractors could easily be inferred with heuristics.

B Additional Results

We leave some results that we cannot fit into the main paper here.

B.1 Vector Norm (VN) Composition Scheme

In this section, we explain why we chose not to cover the vector norm composition scheme (mentioned in 2) in our main evaluation results.

We would like to argue first that even mathematically, VN is not a good fit for our evaluation paradigm. Vector norm composition scheme will only indicate the importance of a feature, but will not indicate the polarity of the importance because it cannot generate a negative word importance score, which is important for our evaluation. The reason why it is important is that our plausibility evaluation does distinguish between input words that should have positive/negative importance scores by placing them in cue and attractor sets, respectively. For example, in Table 1, the singular proper noun *U.S.* and *Europe* are important input words because they could potentially lead the model to make the alternative prediction *is* instead of the expected prediction *are*. Hence, they are placed into the attractor set, and when interpreting the next word prediction *are*, our plausibility test expects that they should have large negative importance scores.

Besides, we did run the plausibility evaluation with vector norm composition scheme under some settings, as shown in Table 8. For the vanilla gradient saliency method, the VN composition scheme performs on-par with the gradient · input (GI) scheme (which is used for our main results). However, with SmoothGrad, the plausibility result does not significantly change like the case with the gradient · input (GI) scheme. This corroborates with the results in (Ding et al., 2019), where they also show that SmoothGrad does not improve the interpretation quality with VN composition scheme.

With these theoretical and empirical evidence, we decided to drop vector norm composition scheme for our evaluation.

B.2 Patterns for Changes of Interpretation Quality with Varying Model Configurations

As mentioned in Section 5.3, we would like to know if there are any predictable patterns in how interpretation quality changes with varying model configurations. To answer this question, we build smaller Transformer language models with various depth, number of heads, embedding size, and feed-forward layer width settings, while keeping other hyperparameters unchanged.

	Number Agreement						Gender Agreement					
	PTB			Syneval			CoNLL			Winobias		
	all	exp.	alt.	all	exp.	alt.	all	exp.	alt.	all	exp.	alt.
LSTM		(0.858)	(0.142)		(0.596)	(0.404)		(0.730)	(0.270)		(0.584)	(0.416)
V+VN	0.683	0.719	0.463	0.643	0.466	0.903	0.459	0.393	0.639	0.680	0.807	0.502
SG+VN	0.543	0.540	0.561	0.549	0.271	0.959	0.394	0.234	0.829	0.625	0.587	0.678
QRNN		(0.818)	(0.182)		(0.558)	(0.442)		(0.712)	(0.288)		(0.715)	(0.285)
V+VN	0.630	0.673	0.437	0.579	0.456	0.735	0.427	0.309	0.716	0.526	0.650	0.214
SG+VN	0.559	0.567	0.521	0.556	0.352	0.813	0.398	0.230	0.811	0.539	0.538	0.540
Transformer		(0.919)	(0.081)		(0.594)	(0.406)		(0.761)	(0.239)		(0.219)	(0.781)
V+VN	0.604	0.620	0.424	0.671	0.525	0.885	0.507	0.511	0.493	0.481	0.840	0.380
SG+VN	0.592	0.596	0.542	0.654	0.504	0.872	0.529	0.538	0.500	0.437	0.836	0.325

Table 8: Plausibility benchmark result for Vector Norm (VN) composition scheme. Refer to the caption of Table 3 for notations.

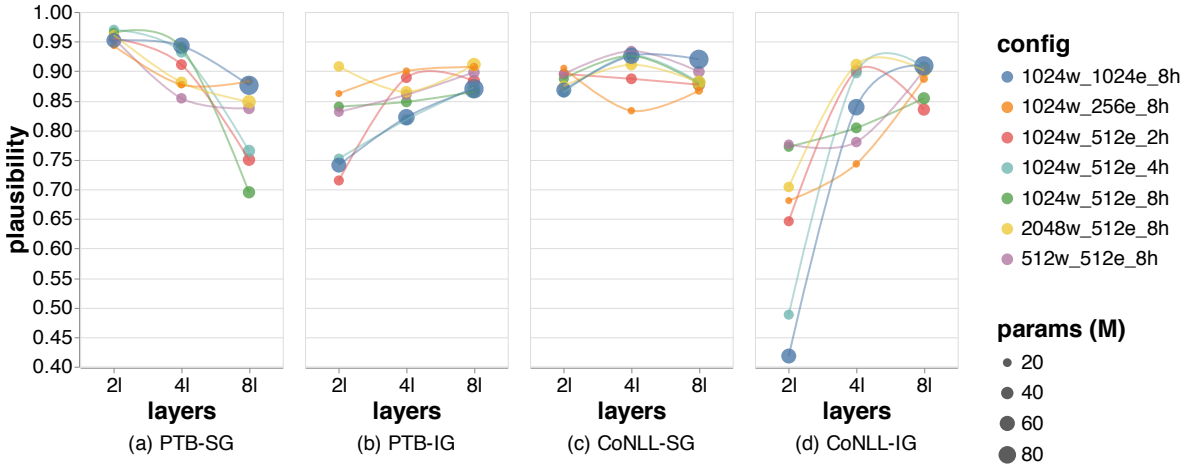
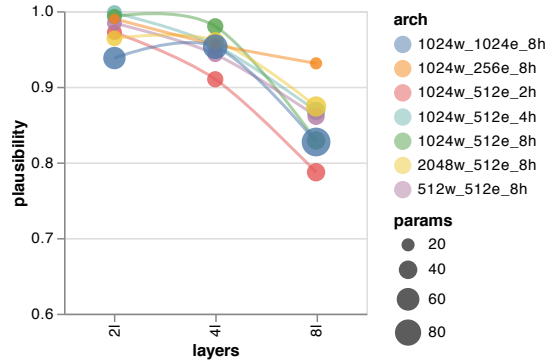


Figure 1: Analysis of model configuration vs. plausibility on PTB and CoNLL benchmark. Each model configuration is color-coded, while the parameter size (in millions) is shown with circle size. `l`, `w`, `e`, `h` stands for model depth, width of feed-forward layers after self-attention, embedding size, and the number of heads.

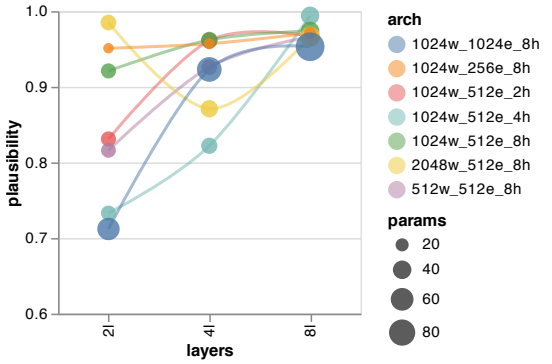
We show two different groups of comparison here. Figure 1 shows our investigation on the interaction between model configuration and interpretation plausibility on PTB and CoNLL test sets. In general, Integrated Gradients method works better for deeper models, while SG works better for shallower models on the PTB test set, but remains roughly the same performance for all architectures on the CoNLL test set. This indicates the noisiness of the trend we are investigating, as both interpretability methods and evaluation dataset choice can influence the trend. As for the other factors of the model configurations, the trend is even noisier (note how much rankings of different configurations change moving from shallow to deep models) and do not show any clear patterns.

Figure 2, on the other hand, focuses on one specific dataset and investigates the trend on both the plausibility and input faithfulness with varying model configurations. For plausibility results, we largely see the same trend as on PTB dataset. For faithfulness results, the trend for SG is largely the same as plausibility. For IG, the variance across other factors of configurations tends to be different on shallower models vs. deeper models, but overall still shows higher numbers for deeper models like plausibility.

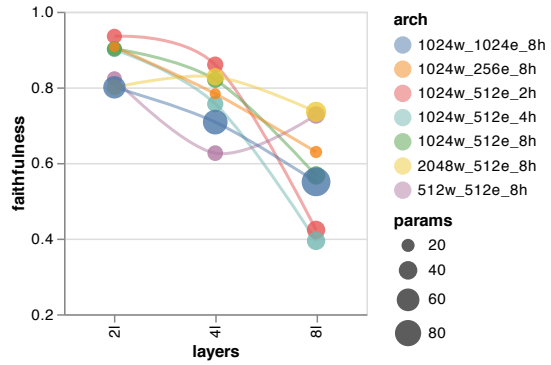
Overall, these analyses further support our conclusion in the main paper, that interpretation qualities are sensitive to model configuration changes, and we reiterate that evaluations of saliency methods should be conducted on the specific model configurations of interest, and trends of interpretation quality on a specific model configuration should not be over-generalized to other configurations.



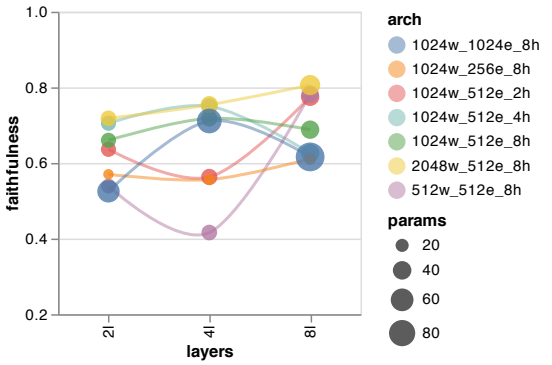
(a) Plausibility benchmark with SmoothGrad



(b) Plausibility benchmark with Integrated Gradients



(c) Faithfulness benchmark with SmoothGrad



(d) Faithfulness benchmark with Integrated Gradients

Figure 2: Analysis of model configuration vs. plausibility and faithfulness on Syneval benchmark. Each model configuration is color-coded, while the parameter size (in millions) is shown with circle size. `l`, `w`, `e`, `h` stands for model depth, width of feed-forward layers after self-attention, embedding size, and the number of heads. Note that the faithfulness numbers plotted here are the ones interpreted with expected scenario predictions.

C Language Model Perplexities

Parameter size and perplexity on WikiText-103 dev set for all language models are shown in Table 9 for reference.

Below are the respective commands to reproduce these results.

- LSTM:
```
python -u main.py -epochs 50 -nlayers 3 -emsize 400 -nhid 2000 -dropoute 0 -dropouth 0.01 -dropouti 0.01 -dropout 0.4 -wdrop 0.2 -bptt 140 -batch_size 60 -optimizer adam -lr 1e-3 -data data/wikitext-103 -save save -when 25 35 -model LSTM
```

- QRNN:
```
python -u main.py -epochs 14 -nlayers 4 -emsize 400 -nhid 2500 -alpha 0 -beta 0 -dropoute 0 -dropouth 0.1 -dropouti 0.1 -dropout 0.1 -wdrop 0 -wdecay 0 -bptt 140 -batch_size 40 -optimizer adam -lr 1e-3 -data data/wikitext-103 -save save -when 12 -model QRNN
```

- Transformer:
```
python train.py -task language_modeling data-bin/wikitext-103 -save-dir checkpoints -arch transformer_lm_wiki103 -decoder-layers $layers -decoder-attention-heads $num_heads -decoder-embed-dim $emb -decoder-ffn-embed-dim $width -max-update 286000 -max-lr 1.0 -t-mult 2 -lr-period-updates 270000 -lr-scheduler cosine -lr-shrink 0.75 -warmup-updates 16000 -warmup-init-lr 1e-07 -min-lr 1e-09 -optimizer nag -lr 0.0001 -clip-norm 0.1 -criterion adaptive_loss -max-tokens 3072 -update-freq 3 -tokens-per-sample 3072 -seed 1 -sample-break-mode none -skip-invalid-size-inputs-valid-test -ddp-backend=no_c10d
```

Architectures	Layers	Config	Params (M)	dev ppl
LSTM	3	-	162	37.65
	2	-	130	41.97
QRNN	4	-	154	32.12
	3	-	135	36.54
Transformer				
	16	4096w_1024e_8h	247	17.97
	4	4096w_1024e_8h (Distill Student)	96.1	24.92
	4	4096w_1024e_8h	96.1	28.96
	8	1024w_512e_2h	39.3	32.63
	8	1024w_512e_4h	39.3	32.09
	8	1024w_512e_8h	39.3	31.38
	8	512w_512e_8h	35.1	33.99
	8	2048w_512e_8h	47.7	30.19
	8	1024w_256e_8h	17.4	41.56
	8	1024w_1024e_8h	96.1	27.01
	4	1024w_512e_2h	30.8	37.03
	4	1024w_512e_4h	30.8	35.67
	4	1024w_512e_8h	30.8	35.82
	4	512w_512e_8h	28.7	38.34
	4	2048w_512e_8h	35.0	33.70
	4	1024w_256e_8h	14.3	48.47
	4	1024w_1024e_8h	70.9	30.46
	2	1024w_512e_2h	26.6	44.45
	2	1024w_512e_4h	26.6	42.23
	2	1024w_512e_8h	26.6	41.86
	2	512w_512e_8h	25.6	44.97
	2	2048w_512e_8h	28.7	38.99
	2	1024w_256e_8h	12.7	59.16
	2	1024w_1024e_8h	58.3	36.06

Table 9: Parameter size (in millions) and perplexity on WikiText-103 dev set for all language models we trained.

D Additional Interpretation Examples

We show some additional interpretations generated by the state-of-the-art LSTM (Table 10), QRNN (Table 11) and Transformer (Table 12) models on PTB and CoNLL dataset, with their respective best-performing interpretation method.

PTB

1- (U.S.) (Trade) (Representative) (Carla) (Hills) said the first dispute-settlement (panel) set up under the U.S.-Canadian " free (trade) " (agreement) has ruled that (Canada) 's [restrictions] on [exports] of (Pacific) (salmon) and (herring) | PLURAL

2- Individual [investors] , (investment) [firms] and [arbitragers] who speculate in the [stocks] of (takeover) [candidates] can suffer (liquidity) and (payment) [problems] when [stocks] dive ; those [investors] often | PLURAL

3- (U.S.) [companies] wanting to expand in (Europe) | PLURAL

4- CURBING [WAGE] (BOOSTS) will get high [priority] again in 1990 collective [bargaining] , a [Bureau] of [National] [Affairs] [survey] of 250 (companies) with (pacts) expiring next [year] | PLURAL

5- TEMPORARY (WORKERS) have good (educations) , the [National] [Association] of [Temporary] [Services] | SINGULAR

CoNLL

1- [Israeli] [Prime] [Minister] [Ehud] [Barak] says [he] is freezing tens of millions of dollars in tax payments to the Palestinian Authority . [Mr.] [Barak] says [he] is withholding the money until the Palestinians abide by cease - fire agreements . Earlier Thursday [Mr.] [Barak] ruled out an early resumption of peace talks , even with the United States acting as intermediary . (Eve) (Conette) reports from Jerusalem . Defending what | MALE

2- Once again there 'll be two presidential candidates missing from the debate . Pat Buchanan hardly registers on the political radar this year . And Ralph Nader , who may make the difference between a [Gore] or [Bush] win in several places . (ABC) ('s) (Linda) (Douglas) was with | FEMALE

Table 10: Addition interpretation examples with LSTM.

PTB

1- (U.S.) (Trade) (Representative) (Carla) (Hills) said the first dispute-settlement (panel) set up under the U.S.-Canadian " free (trade) " (agreement) has ruled that (Canada) 's [restrictions] on [exports] of (Pacific) (salmon) and (herring) | PLURAL

2- Individual [investors] , (investment) [firms] and [arbitragers] who speculate in the [stocks] of (takeover) [candidates] can suffer (liquidity) and (payment) [problems] when [stocks] dive ; those [investors] often | PLURAL

3- (U.S.) [companies] wanting to expand in (Europe) | PLURAL

4- CURBING [WAGE] (BOOSTS) will get high [priority] again in 1990 collective [bargaining] , a [Bureau] of [National] [Affairs] [survey] of 250 (companies) with (pacts) expiring next [year] | PLURAL

5- TEMPORARY (WORKERS) have good (educations) , the [National] [Association] of [Temporary] [Services] | PLURAL

CoNLL

1- [Israeli] [Prime] [Minister] [Ehud] [Barak] says [he] is freezing tens of millions of dollars in tax payments to the Palestinian Authority . [Mr.] [Barak] says [he] is withholding the money until the Palestinians abide by cease - fire agreements . Earlier Thursday [Mr.] [Barak] ruled out an early resumption of peace talks , even with the United States acting as intermediary . (Eve) (Conette) reports from Jerusalem . Defending what | FEMALE

2- Once again there 'll be two presidential candidates missing from the debate . Pat Buchanan hardly registers on the political radar this year . And Ralph Nader , who may make the difference between a [Gore] or [Bush] win in several places . (ABC) ('s) (Linda) (Douglas) was with | FEMALE

Table 11: Addition interpretation examples with QRNN.

PTB

1- (U.S.) (Trade) (Representative) (Carla) (Hills) said the first dispute-settlement (panel) set up under the U.S.-Canadian “ free (trade) ” (agreement) has ruled that (Canada) ’s [restrictions] on [exports] of (Pacific) (salmon) and (herring) | PLURAL

2- Individual [investors] , (investment) [firms] and [arbitragers] who speculate in the [stocks] of (takeover) [candidates] can suffer (liquidity) and (payment) [problems] when [stocks] dive ; those [investors] often | PLURAL

3- (U.S.) [companies] wanting to expand in (Europe) | PLURAL

4- CURBING [WAGE] (BOOSTS) will get high [priority] again in 1990 collective [bargaining] , a [Bureau] of [National] [Affairs] [survey] of 250 (companies) with (pacts) expiring next [year] | PLURAL

5- TEMPORARY (WORKERS) have good (educations) , the [National] [Association] of [Temporary] [Services] | SINGULAR

CoNLL

1- [Israeli] [Prime] [Minister] [Ehud] [Barak] says [he] is freezing tens of millions of dollars in tax payments to the Palestinian Authority . [Mr.] [Barak] says [he] is withholding the money until the Palestinians abide by cease - fire agreements . Earlier Thursday [Mr.] [Barak] ruled out an early resumption of peace talks , even with the United States acting as intermediary . (Eve) (Conette) reports from Jerusalem . Defending what | FEMALE

2- Once again there ’ll be two presidential candidates missing from the debate . Pat Buchanan hardly registers on the political radar this year . And Ralph Nader , who may make the difference between a [Gore] or [Bush] win in several places . (ABC) (’s) (Linda) (Douglas) was with | MALE

Table 12: Addition interpretation examples with Transformer.