How does Uncertainty Impact Explanation Coherence?

Anonymous ACL submission

Abstract

Explainable AI methods facilitate the understanding of model behaviour. However, small, imperceptible perturbations to inputs can vastly distort explanations. As these explanations are typically evaluated holistically, before model deployment, it is difficult to assess when a particular explanation is trustworthy. In contrast, uncertainty is easily measured at inference time and in an unsupervised fashion. Some studies have tried to create confidence estimators for explanations, but none have investigated an existing link between uncertainty and explanation quality. We artificially simulate epistemic uncertainty in text input by introducing noise at inference time. In this large-scale empirical study, we insert different levels of noise in a myriad of ways and measure the effect on PLM output and uncertainty metrics. We find that uncertainty and explanation coherence have a task-dependant correlation which can be moderately positive and potentially stems from noise exposed during the training process; this suggests that these models may be better at identifying salient tokens when uncertain, which can be used for human-AI collaboration. While this quality can be at odds with robustness to noise, Integrated Gradients typically shows good robustness and a relatively strong correlation to uncertainty given perturbed data. This suggests that uncertainty is not only an indicator of output reliability, but could also be a potential indicator of explanation coherence.

1 Introduction

Though large language models like ChatGPT have become increasingly popular for personal and industrial use, these black-box models have been prone to perpetuate discrimination and output hallucinations (Augenstein et al., 2023; Bang et al., 2023; Weidinger et al., 2021). To use these models safely, it is important to instil a level of trust in their output. Some methods of instilling trust in a model output include *uncertainty* estimation and *eXplainable AI* (XAI). Uncertainty is a reflection of a model's confidence in its output, given, for example, ambiguous or noisy data. While uncertainty can be estimated at inference time in an unsupervised manner, XAI is typically holistically evaluated for a model and task (Chen et al., 2022; Hedström et al., 2023). However, XAI techniques give unstable explanations given small changes in input data (Adebayo et al., 2018; Alvarez-Melis and Jaakkola, 2018; Lakkaraju and Bastani, 2020). While these studies have been critiqued for inserting unnatural noise into the input data, even relatively realistic perturbations to images can disrupt most gradient-based saliency map techniques (Amorim et al., 2023).

Therefore, it is difficult to know when we can trust a specific explanation. Ideally, we would like to use XAI to understand why a model succeeds and fails to identify points of failure in a model pipeline– these failures could arise from mistakes in the model training or ambiguity within the data. It is vital to understand when explanations are trustworthy, as the inclusion of XAI can cause an over-reliance on models (Bauer et al., 2023; van der Waa et al., 2021), give users the false impression of global task understanding (Chromik et al., 2021), and lead to overall poorer performance than if no human-AI collaboration (Schmidt et al., 2020). Therefore, we would like to assess if the uncertainty of a model's output can give any indication of an explanation's quality. We expect noise at inference time, especially for text data: Words can be accidentally ablated, mispelled or otherwise mutated. Different authors have distinct linguistic styles. New words emerge or change in meaning. Thanks to this noise, many SOTA language models suffer out-of-distribution issues and, thus, fail in real-world applications (Alipanahi et al., 2022; Ribeiro et al., 2020). As large language models rely on drawing from large amounts of data (often stemming from sources with variable writing styles and

Noise type	Example text
(unperturbed)	“an artful intelligent film that stays within the confines of a well-established genre”
MASK	“an [MASK] [MASK] film that stays within the confines of a [MASK] genre”
UNK	“an [UNK] [UNK] film that stays within the confines of a [UNK] genre”
charinsert	“an artfuVl intDelligent film that stays within the confines of a well-Mestablished genre”
charswap	“an artfjl intellhgent film that stays within the confines of a Pell-established genre”
butterfingers	“an artdul intelligegt film that stays within the confines of a well-esfablished genre”
l33t	“an @r7fu1 1n7311193n7 film that stays within the confines of a w311-357@611543d genre”
synonym	"an disingenous sound film that stays within the confines of a good-established genre"

Table 1: All 7 types of perturbation visualized on a datapoint at 25% human-hierarchy perturbation

formatting, like social media), we must understand how this “noise” in the data affects a model’s performance, confidence, and explainability. As text perturbations can introduce some ambiguity into the data that is not present at training time, it should affect a model’s reported uncertainty alongside its explanation. Given the variety of language models available, it is also vital to compare how this differs across different models and XAI methods.

In this paper, we conduct a large-scale empirical investigation into the effect of noise on Pre-trained Language Models (PLMs), via a controlled experiment where we artificially inject varying degrees and types of noise (see Table 1) and measure the impact on model explanations and uncertainty. In this manner, we also investigate the relationship between explanation coherence and model certainty. Here, we provide the following **contributions**:

- We evaluate the relationship between uncertainty and explanation coherence given perturbed and unperturbed data.;
- We assess on a large-scale how the degree of artificial noise at inference time affects model performance, confidence and explanation coherence across a variety of transformer-based language models, degrees of perturbation, and methods of perturbation;
- We compare four popular XAI methods in their robustness to noise across noise types and models at different levels of perturbation.

We find that uncertainty metrics often show a low, positive correlation to explanation coherence; however, the correlation between epistemic uncertainty and explanation coherence can become negative with noise insertion, if there is no noise present during training. Given perturbed data, this relationship often becomes weakest with SmoothGrad and strongest with Guided Backpropogation and Integrated Gradients; Integrated Gradients and SmoothGrad show the greatest robustness to noise, suggesting that saliency maps can be robust while maintaining a relationship with uncertainty.

2 Related Work

2.1 Measures of trustworthiness

There are many ways to assess a model’s trustworthiness for a task or inference. The confidence in an output can be quantified via its uncertainty, and the reasonability of an output can be assessed via XAI. Furthermore, the overall quality of an XAI method can be evaluated, either via the similarity to human annotations or via other metrics like robustness to noise or conciseness (Hedström et al., 2023; Chen et al., 2022; Atanasova et al., 2020). There is some controversy within these measures: Models that output explanations with high similarity to human-annotations may result in unfaithful explanations, as models may not actually rely on this information to compute their output (Jin et al., 2023). Moreover, these explanations can also be unstable and prone to large changes in output given small changes in input data (Adebayo et al., 2018; Alvarez-Melis and Jaakkola, 2018; Lakkaraju and Bastani, 2020; Hedström et al., 2023; Chen et al., 2022). However, as these studies assess for explanation changes given imperceptible changes in (often image) data, we lack understanding as to how these explanations change on large-scale perturbations.

2.2 Noise on PLM Performance

Several other studies have looked specifically at the effect of noise on the performance and confidence of BERT-related models. Surprisingly, many of these found contrasting effects of noise on machine and human ability to perform natural language understanding tasks. Perturbations that would not affect a human’s ability to understand text significantly perturb BERT performance (Jin et al., 2019; Wang et al., 2022), yet perturbations that worsen human performance do not affect model performance (Feng et al., 2018; Gupta et al., 2021; Sinha et al., 2021). The impact of different kinds of noise differs across model types (Moradi and Samwald, 2021), and the more "learnable" a kind of noise is for a model, the less performance decays given augmented data (Zhang et al., 2022). However, as these studies focus on BERT-related models, there is limited focus on other model families, like GPT, and they typically do not evaluate explanations.

2.3 Uncertainty Measures

The 'learnability' of a trait or type of noise can be likened to *epistemic uncertainty*, which is a measure of uncertainty in a model's parameters. This is believed to be malleable given more training time and data (Gal and Ghahramani, 2015). In contrast, *aleatoric uncertainty* stems from noise inherent in the data generation process (Kendall and Gal, 2016). Many studies conflate the two forms of uncertainty by only looking at the softmax of the output logits as a measure of confidence (hereon named *predictive uncertainty*). However, these measures can be prone to over-confidence. For example, when provided highly perturbed data, model confidence increases, even with the addition of calibration methods (Feng et al., 2018; Gupta et al., 2021). As these studies use the conflated measure of predictive uncertainty, it is difficult to ascertain the cause of this confidence increase.

2.4 Uncertainty and XAI

Other works in the intersection of uncertainty and XAI try to quantify the uncertainty of a given explanation, by developing new models (Bykov et al., 2020) or looking at ensemble explanations (Chai, 2018; Slack et al., 2020; Marx et al., 2023), or they attempt to explain the causes of a model's uncertainty (Brown and Talbert, 2022; Watson et al., 2023). In Marx et al. (2023), they find that the size of the dataset is inversely proportional to the uncertainty of the explanations, which suggests that, with increased training data, XAI techniques tend to converge and that epistemic uncertainty may affect XAI explanations. However, these methods do not look at existing links between XAI and uncertainty and look mainly at image and synthetic datasets.

In summary, most studies investigating noise on model output look only at small levels of perturbation and focus on a small subset of language models (if any). Furthermore, they conflate different aspects of uncertainty or create new measures. In our paper, we investigate the effect of different scales of perturbations on a range of popular language models, including GPT2. In addition, to avoid conflating sources of uncertainty, we specifically examine the interaction between XAI and a common measure of epistemic uncertainty to assess the relationship between the two model outputs.

Dataset	Task	Size
SemEval 2013 Task 2	Sentiment Classification	Training: 4133 Annotated Test: 1659
SST-2 + Hummingbird	Sentiment Classification	Training: 67349 Annotated Test: 62
HateXplain	Hatespeech Detection	Training: 15383 Annotated Test: 1142

Table 2: Our training and test datasets. We restrict our test datapoints to those including human-annotated explanations ('Annotated Test').

3 Methods

3.1 Datasets

We identify relevant tasks and datasets for this investigation by limiting ourselves to publicly available datasets in the English language. We select simple, popular text classification tasks (sentiment classification, hatespeech detection) with text that has been annotated for importance at word-level granularity by multiple (2+) annotators. We summarize the datasets in Table 2. Within sentiment classification, we choose two datasets: Hummingbird (Hayati et al., 2021) and the Semeval-2013 Task 2 dataset (Nakov et al., 2013). Hummingbird is a re-annotated subset of several datasets, including the SST-2 dataset (Socher et al., 2013). We restrict the Hummingbird Sentiment test dataset to only datapoints originating from the SST-2 validation set, and train on the SST-2 train dataset. We remove neutral datapoints from SemEval-2013 dataset and HateXplain (Mathew et al., 2020), to avoid issues with the sufficiency of highlighted text as explanations (Wiegreffe and Marasović, 2021).

3.2 Models

We test the performance of four different open-source large pre-trained language models: $\text{BERT}_{\text{base}}$ (Devlin et al., 2018), $\text{RoBERTa}_{\text{base}}$ (Liu et al., 2019), ELECTRA (Clark et al., 2020) and $\text{GPT-2}_{\text{medium}}$ (Radford et al., 2019), chosen due to their variety in pretraining and their popularity. We describe their finetuning in Appendix A.

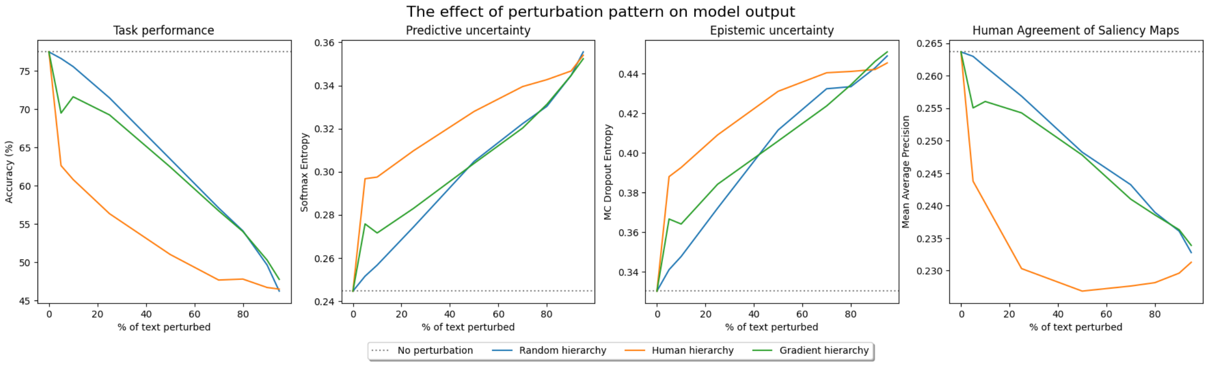

Figure 1: The effect of increasing text perturbation on mean model performance, confidence, and explanation coherence across three different hierarchies: (1) Random perturbation; (2) Human-based perturbation, following human annotation and POS tags; and (3) Gradient-based perturbation, following ranking of Hotflip gradients.

3.3 Perturbations

At test time, we introduce varying levels, hierarchies, and types of perturbations to simulate epistemic uncertainty. A singular type of perturbation is applied to space-delimited words following different hierarchies for increasing **levels**, or proportions, of the text (0%, 5%, 10%, 25%, 50%, 70%, 80%, 90%, 95%).

We use three **hierarchies** for preferential perturbation: `random`-importance, `human`-importance, and `gradient`-importance. Random-importance is determined randomly, though the pattern of perturbed words is preserved across increasing levels of perturbation. Human-importance is determined by the word-level annotations of the dataset. Non-annotated words are then ranked via their part-of-speech tag. We assess the efficacy of this strategic POS perturbation approach in Appendix C.1. Gradient-importance is calculated specific to each model as it is ranked by words with the greatest average change according to the Hotflip candidates table (Ebrahimi et al., 2018). When combining tokens to create full words, we take the mean of token gradients to create the final gradient. This was determined after taking a subsample of the datapoints and choosing the aggregation method that gave the lowest mean ranking to NLTK stopwords.

We introduce seven different noise **types** to the datapoints (see Table 1), selected from previous work in text perturbation: At a fine-grained level, we introduce a random character into a random section of the word (`charinsert`), randomly replace a character in a word (`charswap`) or replace a random character with a character nearby on a qwerty keyboard (`butterfingers`). These insertions have been implemented in other studies on adversarial perturbation in text (Zhang et al., 2022; Moradi and Samwald, 2021). At the word level, we replace words with tokens, such as `MASK`, as has been done in perturbation-based studies (Madsen et al., 2021). We also compare `MASK` replacement with `UNK` tokens, to assess if Masked Language Modelling in pre-training tasks helps models better handle `MASK`-related perturbations. We also convert the entire word to l33t speak (`l33t`) (Eger et al., 2019; Zhang et al., 2022), and swap the word with a semantically related word (`synonym`) using publicly available corpora (Pavlick et al., 2015; Fellbaum, 1998; Loper and Bird, 2002), manually-made dictionaries (e.g. for public Twitter IDs) or randomly generated replacements (e.g. for URLs). Not all words have valid synonyms; therefore, we are only able to perturb about 16.2% of words in the Hummingbird dataset and 18.4% of the SemEval dataset. These mainly consist of rare or slang words, and non-parseable hashtags or misspellings in the case of the SemEval dataset. Our precise rules for synonym replacement can be found in Appendix B.

3.4 Explanation techniques

We focus on local gradient-based explanations as they have been shown to perform best across a range of metrics, models, and tasks (Atanasova et al., 2020). These explanation measures use backpropagation to compute a saliency map over input features for a specific datapoint to audit a model's decision. The simplest implementation uses the gradient of the input as the saliency score (Simonyan et al., 2013); however, the output can be very noisy (Smilkov et al., 2017). Therefore, we rely on modified versions of the technique: **SmoothGrad** (`SG`) returns the average saliency map obtained by

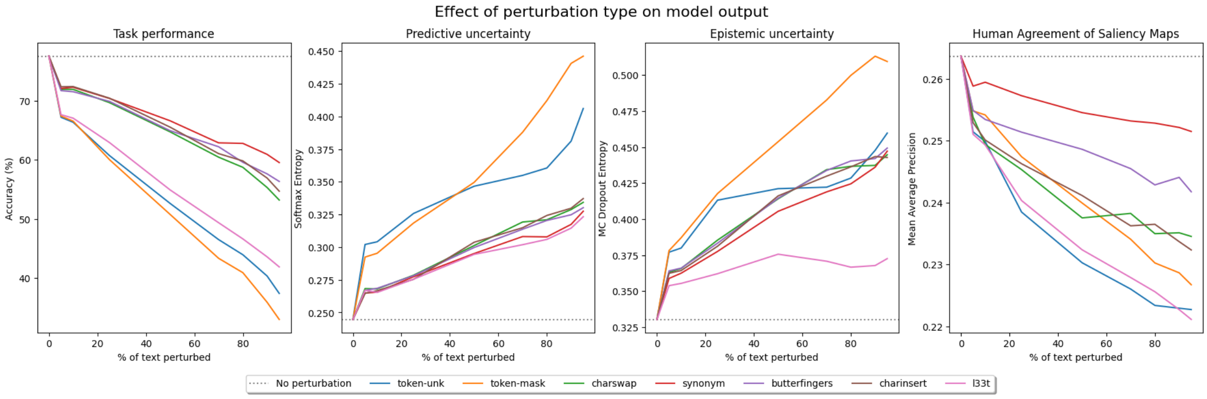

Figure 2: The effect of increasing text perturbation on mean model performance, confidence and explanation coherence across the different types of perturbation.

perturbing the original input with Gaussian noise (Smilkov et al., 2017). **Guided Backpropogation** (GBP) uses a different computation of gradients (by ignoring all negative values) to visually improve its saliency maps (Springenberg et al., 2014). **InputX-Gradients** (IXG) considers both the importance of the feature and the strength of the expressed dimension (Shrikumar et al., 2016). **IntegratedGradients** (IG) accumulates the gradients between an input of interest and a neutral baseline (Sundararajan et al., 2017). We use the Captum implementations of these saliency maps (Kokhlikyan et al., 2019).

3.5 Evaluation design

For comparisons to the human annotations and across models, we combine all gradients back to word level (i.e. space-delimited). We use **accuracy** as a reflection of model output quality. To measure model confidence, we use several measures of uncertainty: We calculate **predictive uncertainty** (PRU), which is traditionally reported in the literature, via the entropy of the softmax logits (to reduce overconfidence (Pearce et al., 2021)). We approximate **epistemic uncertainty** (EPU) via the entropy of model predictions after 100 inferences with dropout left on (Kendall and Gal, 2016). As a measure of **explanation coherence**, we take the Mean Average Precision (MAP) of model gradients with respect to the human-level annotations.

As a Kolmogorov–Smirnov test of the MAPs and both measures of entropy violate the assumption of normality ($p < 10^{-5}$), we use Spearman's Rank Correlation[1] to assess shared trends across models and datasets. We calculate the correlation coefficient between the MAP of the gradients to the human annotations and both measures of entropy at a data-point level. We only include datapoints that are correctly predicted, to ensure the relevance of the annotated explanations. We divide our investigation between perturbed and un-perturbed data, and across model, attribution method, and dataset, to assess the generalisability of findings.

Finally, to evaluate the change in explanation coherence with noise, we calculate the Pearson correlation of the new saliency maps with the original saliency maps and the perturbation pattern.

4 Results

4.1 Noise on uncertainty and explanations

4.1.1 The effect of perturbation prioritization

We present the aggregated effect of different hierarchies of perturbation as described in §3.3 in Figure 1. All perturbations impair model performance, uncertainty, and explanation coherence, but human-prioritised perturbation has the greatest impact up to very high levels of perturbation. While random and gradient-based perturbation generally have similar impact on task performance, uncertainty and explanation coherence, gradient-based perturbation strategies have a stronger impact on these metrics at low levels of perturbation. Interestingly, the decrease in explanation coherence is markedly smaller given increasing perturbation than that for task performance and uncertainty.

4.1.2 The effect of perturbation type

We show the aggregated effect of the investigated noise types listed in Table 1 in Figure 2. Though all perturbation types adversely impact task performance and human agreement, this effect is smaller for `synonym` and `butterfinger`. In contrast, token

[1]We use the implementententation in SciPy v1.11.4

replacements have the greatest detrimental effect. Surprisingly, while most perturbations augment uncertainty as they increase in scale, we do not see this with `l33t` perturbation and epistemic uncertainty. This is investigated further in Appendix C.2 and find it owes to dataset-level differences. We further show-case model-level differences in Figure 3 and in Appendix C.3, where we find that BERT and RoBERTa show the greatest increase in uncertainty given `MASK` tokens and decrease in uncertainty with increasing `l33t` speak. This is surprising, given that, while previous studies using l33t perturbation (Zhang et al., 2022; Eger et al., 2019) do not report confidence measures, Zhang et al. (2022) note that this perturbation was one of the most "learnable" perturbations for the models, which we expect to correlate with epistemic uncertainty.

4.2 The relationship between uncertainty and explanation coherence

We assess the correlation between uncertainty and explanation coherence across all datasets, saliency maps, and models in Table 3. Before perturbation, we surprisingly find a tendency for low to moderate positive correlation between uncertainty and explanation quality for the SemEval and HateXplain datasets. While SST-2 shows a weak correlation between the metrics before perturbation, this becomes moderately negative after perturbation. Typically, attribution methods that show a stronger correlation to uncertainty levels before perturbation continue to show a relatively stronger correlation given perturbed data. We see similar patterns in correlation between all attribution methods; however, SmoothGrad (`SG`) typically shows much weaker correlation after perturbation, whereas Guided Backpropagation (`GBP`) and Integrated Gradients (`IG`) show the strongest.

4.3 The change in explanation with increasing noise

In Figure 4, we visualize the robustness of saliency maps across low and high levels of perturbation. At low levels of perturbation (10%), `IG` shows the greatest correlation to the original saliency map regardless of the type of noise introduced to the datapoint. At higher levels, `SG` has the greatest general robustness to noise. Interestingly, at high levels of perturbation, while `SG` is equally robust to all types of perturbation, `IG` and `IXG` show greater robustness to `synonym` and `charswaps`.

We also investigate model-level differences at low levels of perturbation in Figure 5 and find that Integrated Gradients shows the greatest robustness for the models BERT, RoBERTa, and ELECTRA. However, SmoothGrad has the greatest robustness for GPT2. Figure 4 also shows the correlation to noise across saliency map and perturbation types. None of the saliency maps show any strong correlation to noise. Therefore, despite lower saliency being attributed to previously salient tokens given increasing noise, models do not seem to attribute saliency to the input noise instead.

In summary, while perturbation decreases model performance and explanation coherence, it has a task-dependent effect on uncertainty. We also see dataset-level differences in correlations between explanation coherence and uncertainty, which is often moderately positive; the strength of this association given perturbed data differs also between saliency maps, where SmoothGrad is typically weakest. Furthermore, Integrated Gradients is most robust against all types of noise at low levels of perturbation for most models. SmoothGrad shows greater robustness for GPT2 and for all models at high levels of perturbation.

5 Discussion

While noise consistently deteriorates model performance and explanation coherence, the impact of increasing noise on model confidence varies across model and task. Unlike previous studies, we do not typically see an increase in confidence after perturbation (Feng et al., 2018; Gupta et al., 2021), but rather a decrease. However, both studies perturb at the word and sentence structure-level, unlike our study. Interestingly, we see the greatest difference between perturbation patterns at low levels of perturbation. Overall, human-based perturbations have the strongest effect on task performance and uncertainty measures. Gradient-based perturbation is only more effective than random perturbation at low levels of perturbations. This suggests that these human annotations are faithful indicators of salient tokens, as their perturbation degrades model performance more than gradient-based approaches.

Across all models, realistic perturbations, such as `charswap` or `synonym` have the smallest impact on task performance and explanation coherence, yet masking has the greatest impact. Furthermore, `MASK` has the greatest effect on both measures of confidence. This is surprising, given that both BERT and RoBERTa have masked-language mod-

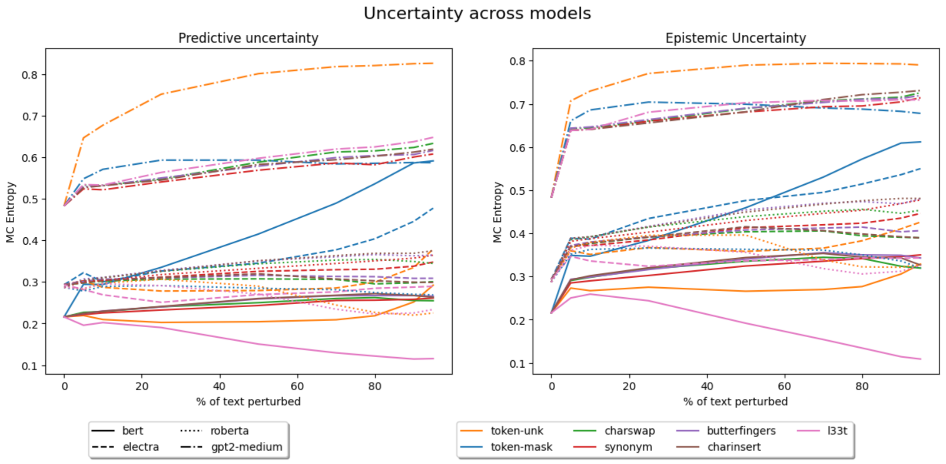

Figure 3: The differential effect of increasing levels of text perturbation on predictive (left) and epistemic uncertainty (right) across 7 different kinds of noise between our four investigated models.

		Before Perturbation								**Including Perturbed Text**							
		Predictive uncertainty				**Epistemic uncertainty**				**Predictive uncertainty**				**Epistemic uncertainty**			
model	**dataset**	**GBP**	**IXG**	**IG**	**SG**	**GBP**	**IXG**	**IG**	**SG**	**GBP**	**IXG**	**IG**	**SG**	**GBP**	**IXG**	**IG**	**SG**
BERT	SST-2	0.076	0.068	-0.128	**-0.155**	-0.052	**-0.060**	0.041	0.039	**-0.104**	-0.099	-0.069	-0.069	-0.240	-0.228	**-0.248**	-0.219
	SemEval	0.237	0.248	0.238	**0.249**	0.235	**0.247**	0.234	**0.247**	0.149	**0.165**	0.150	**0.165**	0.151	**0.166**	0.148	0.164
	HateXplain	0.268	**0.270**	0.265	0.262	0.211	0.229	0.263	**0.267**	0.293	0.178	**0.297**	0.181	0.243	0.139	**0.259**	0.148
ELECTRA	SST-2	0.040	0.002	-0.050	**-0.089**	**-0.127**	-0.065	-0.050	-0.058	**-0.096**	**-0.096**	-0.043	-0.050	**-0.383**	-0.380	-0.164	-0.175
	SemEval	0.200	**0.232**	0.199	**0.232**	0.201	**0.233**	0.199	0.231	0.162	**0.169**	0.162	**0.169**	0.163	**0.171**	0.162	0.170
	HateXplain	0.565	0.458	**0.573**	0.464	0.539	0.430	**0.568**	0.462	0.444	0.240	**0.452**	0.247	0.425	0.221	**0.448**	0.244
RoBERTa	SST-2	**0.088**	0.048	0.030	-0.000	**-0.367**	-0.330	-0.174	-0.200	**-0.124**	-0.101	-0.084	-0.069	**-0.357**	-0.324	-0.267	-0.246
	SemEval	0.213	**0.234**	0.212	**0.234**	0.215	**0.235**	0.213	**0.235**	0.149	**0.155**	0.148	0.154	0.149	**0.155**	0.147	0.153
	HateXplain	**0.529**	0.434	0.517	0.424	0.502	0.407	**0.503**	0.408	**0.396**	0.218	0.390	0.213	0.371	0.195	**0.379**	0.201
GPT2	SST-2	0.078	-0.033	**0.124**	-0.014	-0.150	**-0.237**	-0.036	-0.088	**-0.092**	-0.068	-0.013	-0.004	-0.232	**-0.241**	-0.094	-0.068
	SemEval	**0.220**	0.181	0.218	0.182	**0.221**	0.184	0.219	0.181	**0.127**	0.120	**0.127**	0.121	**0.128**	0.122	0.127	0.120
	HateXplain	**0.393**	0.278	0.386	0.278	0.380	0.270	**0.399**	0.284	**0.300**	0.106	0.298	0.105	0.291	0.097	**0.304**	0.110

Table 3: The Spearman Rank Correlation between explanation coherence (MAP) and both measures of uncertainty across model, dataset and saliency map. We bold the saliency map with the strongest correlation for each comparison.

eling pretraining (Devlin et al., 2018; Liu et al., 2019), and calls into question the use of MASK tokens for faithfulness measures (Madsen et al., 2023).

In the case of hatespeech detection, UNK and l33t surprisingly reduce data and model uncertainty (see Figure 7); this could explain the positive correlation between uncertainty and explanation coherence for HateXplain, as highly perturbed examples will show lower uncertainty as explanation coherence decreases. The dataset is compiled from Twitter, and we suspect that numeric characters may be used to hide potentially offensive terms. While there is no class difference regarding the number of words containing letters and numbers (0: 0.695 %, 1: 0.975 %, 2: 0.912 %), at manual inspection, we find examples of l33t-like speak in Classes 0 and 2 (e.g. h0e) that we do not find in the neutral class (e.g. WW2). The existence of these examples in the training data may have made the noise more easily learned by the models as an indicator of a class, owing to the high "learnability" of this perturbation (Zhang et al., 2022). So, when noise is learned to be an indicator of class, uncertainty may show a positive correlation with output quality and explanation coherence. However, we also see a weaker, positive relationship with the Twitter-based SemEval dataset, and we do not see an increased correlation to l33t noise in Figure 4; therefore, models trained with noise-augmented data (or large amounts of social media data, like large language models) may show this positive relationship. This suggests that when these models have greater uncertainty, they may still be more precise at identifying salient tokens amid noise. Other studies also suggest performance improvements after training models with noisy data (Anonymous, 2023). We show in Appendix C.4 that, at very high

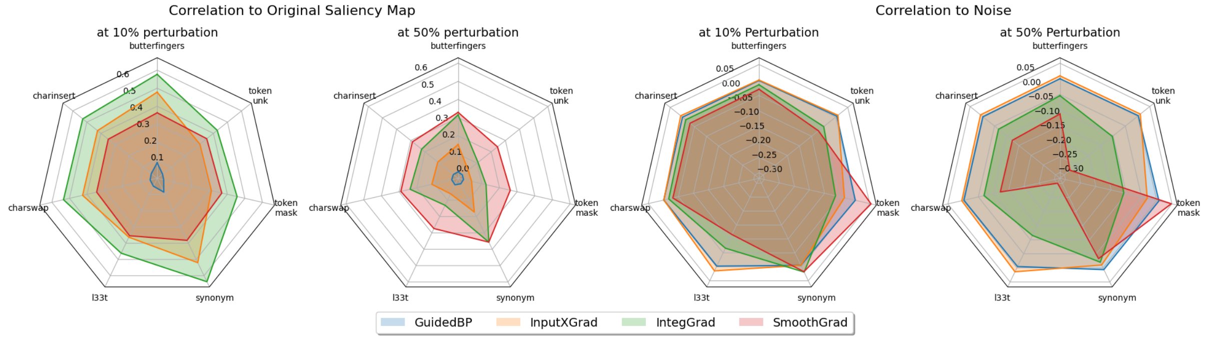

Figure 4: The correlation of various saliency maps to the original saliency map and noise patterns at high and low levels of perturbation. The axes denote the different types of noise. The color denotes the saliency map.

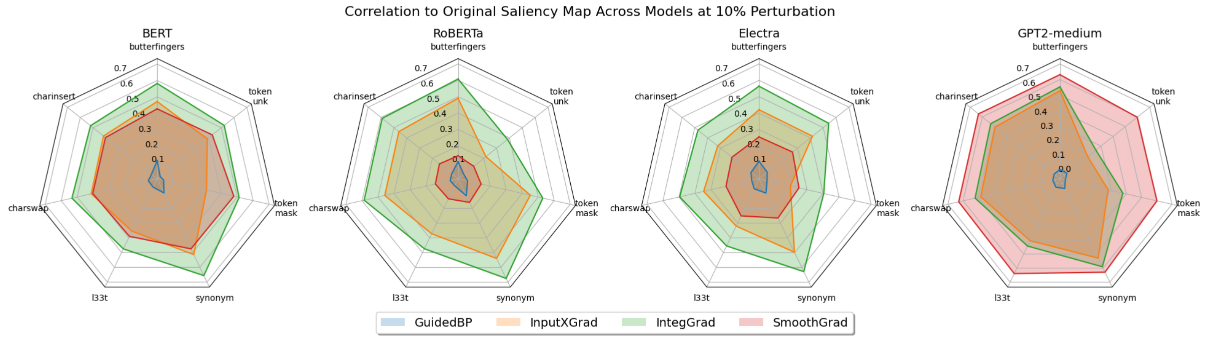

Figure 5: Model-level differences of the correlation to the unperturbed saliency map at low levels of perturbation. We separately show the effect on BERT, RoBERTa, ELECTRA, and GPT2.

perturbation levels, the strength of this relationship weakens (due to lack of meaningful tokens), but may still remain weakly positive for simple tasks.

SmoothGrad shows the greatest all-around robustness to noise but a weak correlation to uncertainty after perturbation. Similarly, Guided backpropagation shows low robustness, but a relatively strong correlation to uncertainty given noisy data. In contrast, Integrated Gradients shows relatively strong correlations to uncertainty but also high robustness for most models at low levels of perturbation. At high levels of perturbation, it and InputXGrad show increased robustness to 'realistic' perturbations (`synonym` and `butterfinger`), which minimally impact model performance (see Figure 2). Therefore, saliency maps can still be robust while correlating to model uncertainty, and patterns in a saliency map's robustness may also relate to model performance.

We recommend that future XAI evaluation and human-XAI collaboration studies consider uncertainty metrics as an additional measure of XAI quality. The relationship between uncertainty and explanation coherence for a model and dataset should be assessed pre-deployment, and an XAI method with adequate robustness and correlation to uncertainty for the model should be chosen. Not only could this help indicate explanation quality at inference time, it may also suggest if noise-augmented training data is needed or if active learning can use strategic word-level human annotations to improve explanation coherence (Nguyen et al., 2019).

6 Conclusion

We provide an empirical investigation across language models, noise perturbations, and saliency maps to investigate a relationship between uncertainty and explanation coherence. Following an array of perturbation techniques, we show that noise injection simultaneously affects model performance, uncertainty, and explanation coherence. However, models fine-tuned on noisier data typically show a moderately positive correlation between explanation coherence and uncertainty, which suggests that these models may be better at identifying salient tokens when uncertain. We also suggest Integrated Gradients for future work in Human-XAI collaboration, due to its robustness to noise and relatively strong correlation to uncertainty given perturbed data.

Limitations

We do not investigate aleatoric uncertainty in this study, as our main experimental setup was to simulate epistemic uncertainty by introducing noise not present in the training data. However, we do assess across different datasets sources, with differing levels of latent noise and aleatoric uncertainty, and find highly correlated results for a shared task. However, future work should consider further disambiguating aleatoric uncertainty in their comparisons. In addition, given our investigation into epistemic uncertainty, it could also be interesting to assess how the observed robustness changes in models fine-tuned with noise-augmented training data. Future studies could also consider simulating uncertainty in other methods, perhaps at other points of the pipeline.

Though we do compare many popular language model types, we could have also chosen to investigate even more. Models with visual encoding, for example PIXEL (Rust et al., 2023), may handle different types of noise differently; visual perturbations, like l33t speak, may show a lesser effect on PIXEL model performance and confidence, whereas more semantic changes, like synonym replacement, may have a larger effect. However, given the format of our study, the saliency maps would be difficult to compare across all model types. Furthermore, we only investigate 3 datasets and 4 language models, which, while more extensive than similar studies, still does not include all popular NLP tasks or extremely large language models (XLMs), like LLAMA (Touvron et al., 2023).

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity Checks for Saliency Maps.

Babak Alipanahi, Farhad Hormozdiari, Alexander D'amour, Katherine Heller, Dan Moldovan, Ben Adlam, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory Mclean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D Sculley. 2022. Underspecification Presents Challenges for Credibility in Modern Machine Learning. Technical report.

David Alvarez-Melis and Tommi S. Jaakkola. 2018. On the Robustness of Interpretability Methods.

José P. Amorim, Pedro H. Abreu, João Santos, Marc Cortes, and Victor Vila. 2023. Evaluating the faithfulness of saliency maps in explaining deep learning models using realistic perturbations. *Information Processing and Management*, 60(2).

Anonymous. 2023. Exploring the impact of information entropy change in learning systems. In *Submitted to The Twelfth International Conference on Learning Representations*. Under review.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A Diagnostic Study of Explainability Techniques for Text Classification.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. 2023. Factuality Challenges in the Era of Large Language Models.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. *CoRR*, abs/2302.04023.

Kevin Bauer, | Moritz Von Zahn, Oliver Hinz, and Moritz Von Zahn. 2023. Please Take Over: XAI, Delegation of
}Authority, and Domain Knowledge. Technical report.

Katherine E Brown and Douglas A Talbert. 2022. Using Explainable AI to Measure Feature Contribution to Uncertainty. Technical report.

Kirill Bykov, Marina M. C. Höhne, Klaus-Robert Müller, Shinichi Nakajima, and Marius Kloft. 2020. How Much Can I Trust You? – Quantifying Uncertainties in Explaining Neural Networks.

Lucy R Chai. 2018. Uncertainty Estimation in Bayesian Neural Networks And Links to Interpretability.

Zixi Chen, Varshini Subhash, Marton Havasi, Weiwei Pan, Finale Doshi-Velez, and John A Paulson. 2022. WHAT MAKES A GOOD EXPLANATION?: A HARMONIZED VIEW OF PROPERTIES OF EXPLANATIONS.

Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I Think i Get Your Point, AI! The Illusion of Explanatory

Depth in Explainable AI. In *International Conference on Intelligent User Interfaces, Proceedings IUI*, pages 307–317. Association for Computing Machinery.

Kevin Clark, Minh-Thang Luong, Google Brain, Quoc V Le Google Brain, and Christopher D Manning. 2020. ELECTRA: PRE-TRAINING TEXT ENCODERS AS DISCRIMINATORS RATHER THAN GENERATORS.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-Box Adversarial Examples for Text Classification. Technical report.

Steffen Eger, Gözde Gül Şahin, Andreas Rücklé, Ji-Ung Lee, Claudia Schulz, Mohsen Mesgar, Krishnkant Swarnkar, and Edwin Simpson. 2019. Text Processing Like Humans Do: Visually Attacking and Shielding NLP Systems. Technical report.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Shi Feng, Eric Wallace, Alvin Grissom Ii, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of Neural Models Make Interpretations Difficult. pages 3719–3728.

Yarin Gal and Zoubin Ghahramani. 2015. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning.

Ashim Gupta, Giorgi Kvernadze, and Vivek Srikumar. 2021. BERT & Family Eat Word Salad: Experiments with Text Understanding.

Shirley Anugrah Hayati, Dongyeop Kang, and Lyle Ungar. 2021. Does BERT Learn as Humans Perceive? Understanding Linguistic Styles through Lexica.

Anna Hedström, tu-berlinde Leander Weber, Dilyara Bareeva, Daniel Krakowczyk, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M-C Höhne. 2023. Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond. *Journal of Machine Learning Research*, 24:1–11.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2019. Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment.

Weina Jin, Xiaoxiao Li, and Ghassan Hamarneh. 2023. Rethinking AI Explainability and Plausibility.

Alex Kendall and Yarin Gal. 2016. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? Technical report.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Jonathan Reynolds, Alexander Melnikov, Natalia Lunova, and Orion Reblitz-Richardson. 2019. Pytorch captum. `https://github.com/pytorch/captum`.

Himabindu Lakkaraju and Osbert Bastani. 2020. "How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations.

Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez, and Ion Stoica. 2018. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit.

Andreas Madsen, Nicholas Meade, Vaibhav Adlakha, and Siva Reddy. 2021. Evaluating the Faithfulness of Importance Measures in NLP by Recursively Masking Allegedly Important Tokens and Retraining.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2023. Faithfulness Measurable Masked Language Models.

Charlie Marx, Youngsuk Park, Hilaf Hasson, Yuyang Wang, Stefano Ermon, and Jun Huan. 2023. But Are You Sure? An Uncertainty-Aware Perspective on Explainable AI. Technical report.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection.

Milad Moradi and Matthias Samwald. 2021. Evaluating the robustness of neural language models to input perturbations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1558–1570, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Preslav Nakov, Zornitsa Kozareva, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Theresa Wilson. 2013. SemEval-2013 Task 2: Sentiment Analysis in Twitter. Technical report.

Vu-Linh Nguyen, Sébastien Destercke, and Eyke Hüllermeier. 2019. Epistemic Uncertainty Sampling.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. Technical report.

Tim Pearce, Alexandra Brintrup, and Jun Zhu. 2021. Understanding Softmax Confidence and Uncertainty.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. Technical report.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. Technical report.

Phillip Rust, Jonas F Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam De Lhoneux, and Desmond Elliott. 2023. LANGUAGE MODELLING WITH PIXELS.

Philipp Schmidt, Felix Biessmann, and Timm Teubner. 2020. Transparency and trust in artificial intelligence systems.

Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. 2016. Not Just a Black Box: Learning Important Features Through Propagating Activation Differences.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps.

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. 2021. UnNatural Language Inference. pages 7329–7346.

Dylan Slack, Sophie Hilgard, Sameer Singh, and Himabindu Lakkaraju. 2020. Reliable Post hoc Explanations: Modeling Uncertainty in Explainability.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. SmoothGrad: removing noise by adding noise.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. 2014. Striving for Simplicity: The All Convolutional Net.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models.

Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291.

Boxin Wang, Chejian Xu, Xiangyu Liu, Yu Cheng, and Bo Li. 2022. SemAttack: Natural Textual Attacks via Different Semantic Spaces.

David S Watson, Joshua O'Hara, Niek Tax, Richard Mudd, and Ido Guy. 2023. Explaining Predictive Uncertainty with Information Theoretic Shapley Values. *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, Iason Gabriel, and <lweidinger@deepmind Com>. 2021. Ethical and social risks of harm from Language Models.

Sarah Wiegreffe and Ana Marasović. 2021. Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing.

Yunxiang Zhang, Liangming Pan, Samson Tan, and Min-Yen Kan. 2022. Interpreting the Robustness of Neural NLP Models to Textual Perturbations. Technical report.

A Hyperparameters

The pre-trained models are connected to a classification head and fine-tuned on the datasets listed in Table 2 using either previously reported optimal hyperparameters or with hyperparameters we identified by exploring the search space with raytuning (Liaw et al., 2018). We use pre-trained tokenizers specific to each model. For BERT, we rely on $BERT_{base}$, which is 110 million parameters. We use $RoBERTa_{base}$, which is 125 million parameters. ELECTRA is 110 million parameters. We rely on $GPT-2_{medium}$, which is 345 million parameters. BERT, RoBERTa, and ELECTRA are trained and assessed on Titan RTX GPUs; GPT2 is trained and assessed on A100 GPUs.

A.1 SST-2

Our BERT model uses the hyperparameters reported by the best-performing BERT-base model on the SST-2 task, which achieves 92.3% accuracy on the evaluation set[2]. While we cannot find hyperparameters reaching the performance described in the

[2]https://huggingface.co/gchhablani/bert-base-cased-finetuned-sst2

original RoBERTa-base (94.8%) article (Liu et al., 2019), we choose the hyperparameters specified by this model card [3], which achieves an accuracy of 94.5% on the evaluation set. Our ELECTRA model uses the best-performing hyperparameters listed in the original article (Clark et al., 2020), which achieves an accuracy of 96.0% on the evaluation set. Our GPT2 model uses the hyperparameters listed in the original article (Radford et al., 2019).

A.2 SemEval and HateXplain

Model hyperparameters are identified using a hyperparameter search space with a learning rate between $1e-6$ and $1e-4$, epochs between 1 and 10, and a batch size of (4, 8, 16, 32).

Our final hyperparameters are shown in the tables below:

BERT, SemEval

Learning Rate	1e-5
Batch Size	16
Epochs	3
Random Seed	37
Adam ϵ	1e-8
adam $\beta 1$	0.9
adam $\beta 2$	0.999
LLRD	None
Decay Type	Linear
Warmup Fraction	0
Attention Dropout	0.1
Dropout	0.1
Weight Decay	0.0

RoBERTa, SemEval

Learning Rate	1e-5
Batch Size	16
Epochs	3
Random Seed	37
Adam ϵ	1e-8
adam $\beta 1$	0.9
adam $\beta 2$	0.999
LLRD	None
Decay Type	Linear
Warmup Fraction	0
Attention Dropout	0.1
Dropout	0.1
Weight Decay	0.0

[3]https://huggingface.co/Bhumika/RoBERTa-base-finetuned-sst2

ELECTRA, SemEval

Learning Rate	3e-6
Batch Size	8
Epochs	5
Random Seed	24
Adam ϵ	1e-8
adam $\beta 1$	0.9
adam $\beta 2$	0.999
LLRD	None
Decay Type	Linear
Warmup Fraction	0
Attention Dropout	0.1
Dropout	0.1
Weight Decay	0.0

GPT2, SemEval

Learning Rate	8e-5
Batch Size	32
Epochs	7
Random Seed	42
Adam ϵ	1
adam $\beta 1$	0.9
adam $\beta 2$	0.999
LLRD	None
Decay Type	Cosine
Warmup Fraction	0.01
Attention Dropout	0.1
Dropout	0.1
Weight Decay	0.1

B Synonym Replacement

Across all synonym replacements, we preserve the case of the original word (e.g. HAPPY! becomes GLAD!). In addition, we use NLTK POS tagger to tag each word to a part of speech for more precise synonym mapping. If NLTK is unable to find a part of speech, or it must be dropped when merging multiple tokens (e.g. if one token is not a punctuation mark or a possession-indicator), then we ignore part of speech.

We followed the following hierarchical rules for synonym replacement:

1. Tokens beginning with `http://t.co/` or `https://t.co/` are replaced with a similar randomly-generated URL string following a similar regex pattern

2. Tokens beginning with a #, we remove the #, find a synonym, and then re-add the #.

3. Tokens beginning with a are replaced with another random Twitter ID found in the test set.

4. Determinants are re-

BERT, HateXplain

Learning Rate	2e-5
Batch Size	32
Epochs	5
Random Seed	2
Adam ϵ	1e-8
adam $\beta 1$	0.9
adam $\beta 2$	0.999
LLRD	None
Decay Type	Linear
Warmup Fraction	0
Attention Dropout	0.1
Dropout	0.1
Weight Decay	0.0

ELECTRA, HateXplain

Learning Rate	2e-5
Batch Size	8
Epochs	2
Random Seed	6
Adam ϵ	1e-8
adam $\beta 1$	0.9
adam $\beta 2$	0.999
LLRD	None
Decay Type	Linear
Warmup Fraction	0
Attention Dropout	0.1
Dropout	0.1
Weight Decay	0.0

RoBERTa, HateXplain

Learning Rate	6e-6
Batch Size	32
Epochs	5
Random Seed	2
Adam ϵ	1e-8
adam $\beta 1$	0.9
adam $\beta 2$	0.999
LLRD	None
Decay Type	Linear
Warmup Fraction	0
Attention Dropout	0.1
Dropout	0.1
Weight Decay	0.0

GPT2, HateXplain

Learning Rate	5e-5
Batch Size	32
Epochs	6
Random Seed	42
Adam ϵ	1e-8
adam $\beta 1$	0.9
adam $\beta 2$	0.999
LLRD	None
Decay Type	Cosine
Warmup Fraction	0.01
Attention Dropout	0.1
Dropout	0.1
Weight Decay	0.1

placed another random determinant (['a', 'an', 'the', 'this', 'that']). Similarly question determinants are replaced with other question determinants. (['that', 'what', 'whatever', 'which', 'whichever'])

5. Proper nouns are replaced with a randomly generated first name or last name. If the original name ends with a "'s", this is removed and then re-added to the synonym.

6. If the word is a quote ["'", "''", "`", "``", '"'], bracket ["(", ")", "{", "}", "[", "]", '/'], punctuation mark ['.', '!', '?', ','], or sentence break ['-', '--', ',', ':', ';'], it is replaced by another quote, bracket, punctuation mark or sentence break.

7. If the word is an arabic number (e.g. 7), it is replaced by its english equivalent (e.g. seven).

8. If a word has a synonym in WordNet or a word with an Equivalence relation in PPDB 2.0, we randomly select a synonym from the set. If a synonym is longer than one word, the words are hyphenated (This is done to simplify matching of saliency maps between perturbations).

9. If the word starts or ends with a quote, bracket, punctuation mark or line break, we remove the character, find a synonym and then re-add the character in question.

10. If there are hyphens, periods or '//' spaced throughout the word, we use the punctuation mark to parse the word and find a replacement word for one of the word subsections.

11. If a word has a forward or reverse entailment in PPDB 2.0, we randomly choose one as a replacement. (e.g. berry for fruit or fruit for berry).

12. If no synonym has been found with using POS tags, I will expand my search in WordNet and PPDB 2.0 without the POS tag.

13. If the word ends with the popular suffixes '-ish', '-ness', or '-less', we remove the suffix, find a synonym, and then re-add the suffix in question.

C Extra investigations

C.1 Human-Random vs Human-Strategic

To assess the efficacy of our human-strategic approach (and if POS tag-level perturbations affect model performance), we compare human-random and human-strategic perturbation in Figure 6, and denote the average location of a change in strategy with a dotted line. **Results**: We can see that POS-prioritized perturbation does adversely affect model performance and uncertainty. However, we find that after all adjectives, adverbs, verbs, and nouns have been perturbed, further perturbation does not show any increasing impact on model performance or uncertainty until the text is nearly completely perturbed. Interestingly, we find that POS-based perturbation does somewhat improve saliency map quality, it is on a very small scale (maximum difference is .003).

C.2 Task-level differences

While we find that our results for accuracy and explanation coherence are fairly well correlated across models (see Table 4) and datasets (see Table 5), both included uncertainty measures (see §3.5) given increasing noise shows only a correlation between the datasets SemEval and SST-2 and the models BERT and ELECTRA. In addition, the human agreement of InputXGrad and GuidedBP does not show a strong correlation across all models.

We further show the task-level differences in uncertainty in Figure 7. **Results**: Special token replacements (with mask or unknown tokens) have the greatest effect on model accuracy; however, this is not translated to the uncertainty and explanation coherence measures. While special token replacements and L33t speech cause the greatest increase in uncertainty for sentiment classification tasks, the introduction of unknown tokens and l33t speak actually reduce model uncertainty in the hatespeech detection task.

C.3 Model-level differences

We showcase model-level differences in reported uncertainty in Figure 3 and in Tables 6 and 7. **Results:** Generally, we see increasing uncertainty with increasing levels of perturbation for all models and noise types. GPT2 outputs much greater predictive and epistemic uncertainty relative to the other base models. GPT2 and RoBERTa show lightly decreasing uncertainty with UNK token and MASK token replacement. ELECTRA's uncertainty is less impacted by random character insertion, relative to BERT and RoBERTa, and BERT and RoBERTa show the greatest decrease in uncertainty with increasing l33t speak in a dataset. Overall, we find that RoBERTa gives fairly high confidence at high perturbation, despite low performance (50.4% at 95% perturbation), yet, in contrast, ELECTRA, BERT, and GPT-2 are more honest regarding uncertainty.

We look at model-level differences in noise correlation at low-levels of perturbation in Figure 8. **Results**: While we see equal lack of correlation to all types of noise for InputXGrad and GuidedBP saliency maps, SmoothGrad shows different behaviour according to model type. For most models, SmoothGrad shows a slight negative correlation to l33t speak and unknown tokens; however, SmoothGrad does not show this particular aversion to unknown tokens with RoBERTa and it does not show a particular aversion to l33t speak with GPT2.

C.4 Uncertainty and explanation coherence at high levels of perturbation

We investigate the correlation between explanation coherence and our two uncertainty measures at very high levels of perturbation (90% and 95%) in Table 8, to assess if the previously observed relationship breaks down after salient tokens are removed. In this comparison, we also include incorrectly guessed datapoints. **Results:** In SST-2, which has no noise in its training data, we continue to observe a moderately negative relationship between uncertainty and explanation coherence. SemEval, which is an easier task than HateXplain, seems to conserve a very weak positive relationship between uncertainty and explanation coherence across models and attribution methods. However, for HateXplain, this correlation disappears (ca. 0.0), which suggests that the model can no longer identify salient tokens.

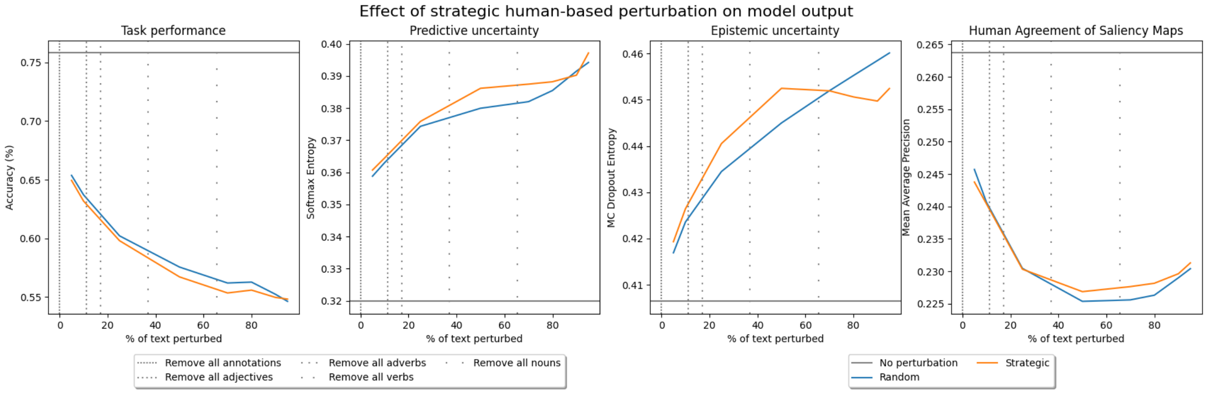

Figure 6: We compare the effect of two different methods of human-based perturbation on model accuracy, confidence and explanation coherence. Human-Random randomly perturbs tokens after all annotated tokens are perturbed. Human-Strategic preferentially perturbs tokens based on their POS. Vertical lines denote the average location of strategy shift for the Human-Strategic perturbation hierarchy.

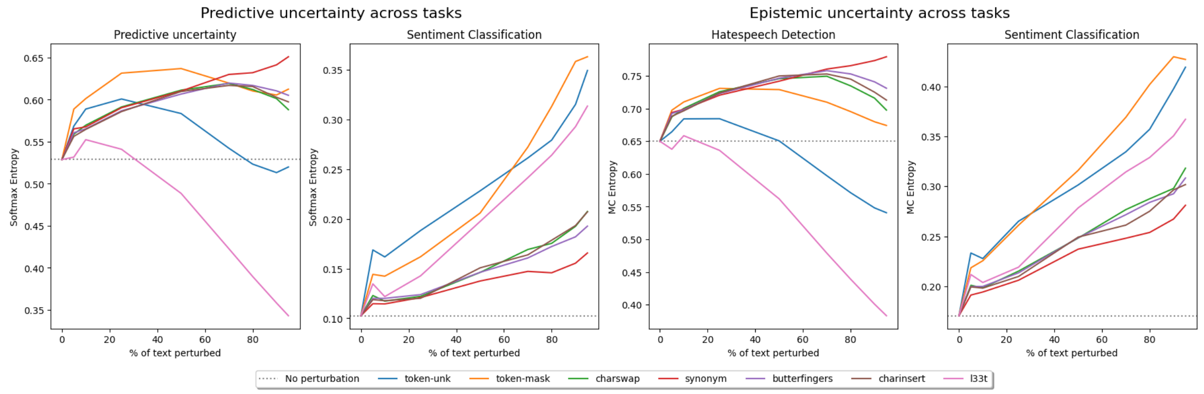

Figure 7: We show the differential effect of increasing levels of text perturbation on **predictive uncertainty** (left two graphs) and **epistemic uncertainty** (right two) across 8 different kinds of noise between the tasks of Hatespeech Detection (left) and Sentiment Classification (right), next to an unperturbed dataset

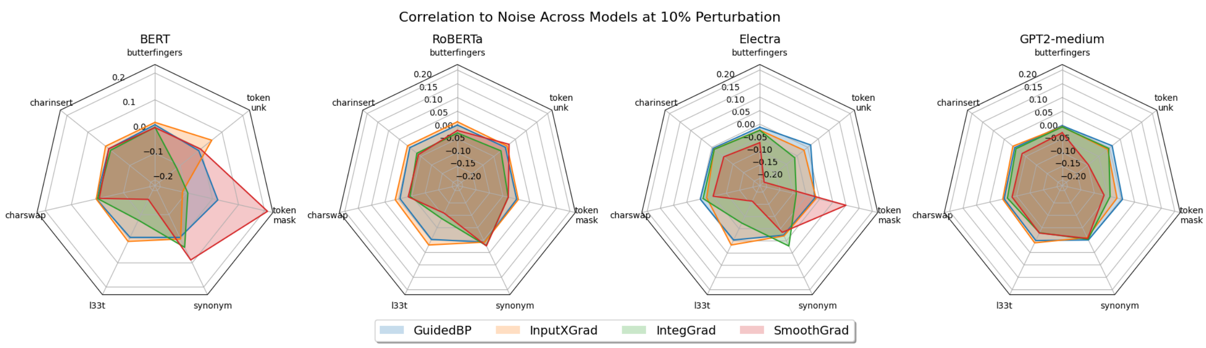

Figure 8: We show model-level differences of the correlation to noise at low levels of perturbation. We separately show the effect on BERT, RoBERTa, ELECTRA, and GPT2.

dataset	dataset	accuracy	PRU	EPU	GBP MAP	IXG MAP	IG MAP	SG MAP
HateXplain	SemEval	0.799 *	-0.215	-0.550	0.923 *	0.792 *	0.870 *	0.951 *
HateXplain	SST-2	0.825 *	-0.269	-0.500	0.870 *	0.505	0.952 *	0.970 *
SemEval	SST-2	0.976 *	0.986 *	0.964 *	0.908 *	0.581	0.800 *	0.939 *

Table 4: The Spearman's Rank Correlation of accuracy, confidence and explanation coherence metrics between datasets across increasing noise of different types of perturbation. A star is drawn next to values with a p < .0001. Dataset differences are further investigated in Figures 7

model	model	accuracy	PRU	EPU	GBP MAP	IXG MAP	IG MAP	SG MAP
BERT	ELECTRA	0.958 *	0.750 *	0.914 *	0.797 *	0.413 *	0.856 *	0.689 *
BERT	RoBERTa	0.910 *	0.416 *	0.464 *	0.479 *	0.147	0.901 *	0.790 *
BERT	GPT2	0.941 *	-0.007	0.081	0.589 *	0.065	0.865 *	0.753 *
ELECTRA	RoBERTa	0.968 *	0.225	0.398	0.250	-0.180	0.832 *	0.407 *
ELECTRA	GPT2	0.927 *	-0.163	0.132	0.358	-0.180	0.668 *	0.717 *
RoBERTa	GPT2	0.897 *	-0.061	0.210	0.845 *	0.559 *	0.848 *	0.554 *

Table 5: The Spearman's Rank Correlation of metrics between models across increasing noise of different noise kinds. A star is drawn next to values with a p < .0001. Model differences are further visualized in Figure 3.

lvl	5	10	25	50	70	80	90	95
Replace with UNK token								
BERT	14	12	9	10	11	13	33	53
RoBERTa	16	23	26	18	7	2	1	6
ELECTRA	26	10	5	7	12	32	51	60
GPT2	57	58	59	60	62	63	64	61
Replace with MASK token								
BERT	58	57	59	60	61	63	64	62
RoBERTa	14	20	21	15	13	12	10	11
ELECTRA	46	24	47	57	61	62	63	64
GPT2	16	22	35	32	26	23	21	24
Swap random character								
BERT	20	22	29	32	36	44	34	31
RoBERTa	24	32	36	42	46	49	44	41
ELECTRA	16	29	35	27	23	13	15	22
GPT2	7	6	15	31	49	50	54	51
Replace with Synonym								
BERT	16	21	25	30	39	37	41	35
RoBERTa	25	28	35	40	43	47	51	50
ELECTRA	28	33	44	48	49	50	52	54
GPT2	2	1	12	20	29	28	38	36
Butterfinger mispelling								
BERT	18	23	27	43	49	50	48	45
RoBERTa	30	34	38	45	54	57	55	53
ELECTRA	25	36	37	42	41	34	39	38
GPT2	4	8	17	25	37	41	47	42
Random character insert								
BERT	19	24	28	42	46	51	47	38
RoBERTa	27	31	37	48	56	59	61	58
ELECTRA	21	31	40	43	19	17	14	20
GPT2	3	5	14	27	34	40	48	45
Convert to l33t speak								
BERT	7	8	6	5	4	2	1	3
RoBERTa	17	22	19	9	4	3	5	8
ELECTRA	8	3	1	2	4	6	9	11
GPT2	10	9	19	43	52	53	55	56

Table 6: Rank of aleatoric uncertainty across perturbation type and model with increasing levels of perturbation. High numbers indicate higher levels of uncertainty.

lvl	5	10	25	50	70	80	90	95
Replace with UNK token								
BERT	13	10	12	9	11	14	23	28
RoBERTa	18	22	32	33	10	6	7	14
ELECTRA	12	9	11	10	14	30	47	49
GPT2	50	58	59	61	64	63	62	60
Replace with MASK token								
BERT	53	51	57	60	61	63	64	62
RoBERTa	12	17	20	15	13	11	5	4
ELECTRA	36	27	51	59	61	62	64	63
GPT2	16	26	40	31	25	24	20	18
Swap random character								
BERT	17	24	32	36	42	39	30	31
RoBERTa	24	29	36	41	45	47	42	43
ELECTRA	15	22	35	38	39	34	26	24
GPT2	7	8	15	28	39	45	55	54
Replace with Synonym								
BERT	15	18	26	34	37	45	48	46
RoBERTa	23	25	35	40	46	49	54	50
ELECTRA	16	18	29	48	50	52	54	55
GPT2	1	5	14	23	29	30	36	37
Butterfinger mispelling								
BERT	21	22	29	41	52	49	47	38
RoBERTa	28	34	38	48	53	57	59	51
ELECTRA	19	23	32	43	46	45	41	42
GPT2	6	10	17	27	35	46	52	47
Random character insert								
BERT	20	25	33	44	50	43	40	35
RoBERTa	26	30	37	44	52	58	60	55
ELECTRA	17	21	33	44	40	31	25	28
GPT2	2	3	13	21	41	53	57	48
Convert to l33t speak								
BERT	7	8	6	5	4	2	1	3
RoBERTa	16	21	19	8	2	1	3	9
ELECTRA	7	4	1	2	3	5	6	8
GPT2	4	9	22	34	42	38	43	33

Table 7: Rank of epistemic uncertainty across perturbation type and model with increasing levels of perturbation. Larger numbers indicate higher numbers of uncertainty.

		Predictive Uncertainty				Epistemic Uncertainty			
Model	**Dataset**	**GBP**	**IXG**	**IG**	**SG**	**GBP**	**IXG**	**IG**	**SG**
BERT	SST-2	-0.016	0.020	-0.015	**0.092**	**-0.162**	-0.100	-0.089	-0.011
	SemEval	0.088	**0.103**	0.088	**0.103**	0.089	**0.104**	0.087	0.103
	HateXplain	-0.049	**-0.078**	-0.049	**-0.078**	-0.040	-0.060	-0.041	**-0.064**
ELECTRA	SST-2	**-0.122**	-0.114	-0.048	-0.032	**-0.308**	-0.289	-0.160	-0.151
	SemEval	**0.103**	0.096	**0.103**	0.096	**0.105**	0.097	0.104	0.097
	HateXplain	-0.054	-0.084	-0.061	**-0.091**	-0.033	-0.059	**-0.060**	-0.090
RoBERTa	SST-2	**-0.169**	-0.123	-0.153	-0.130	**-0.315**	-0.254	-0.244	-0.178
	SemEval	**0.106**	**0.106**	**0.106**	**0.106**	**0.108**	0.106	0.104	0.104
	HateXplain	-0.021	-0.054	-0.023	**-0.055**	-0.009	-0.036	-0.020	**-0.052**
GPT2	SST-2	**-0.075**	-0.017	-0.070	-0.016	-0.159	**-0.100**	-0.096	**-0.048**
	SemEval	0.064	**0.083**	0.065	**0.083**	0.065	**0.085**	0.065	0.084
	HateXplain	0.134	0.090	**0.140**	0.094	0.126	0.097	**0.134**	0.092

Table 8: The Spearman Rank Correlation between explanation coherence (MAP) and both measures of uncertainty across model, dataset and saliency map at high levels of perturbation (90% and 95%). All datapoints (correctly and incorrected guessed) are included. We bold the saliency map with the strongest correlation for each comparison.