Investigating the Impact of Model Instability on Explanations and Uncertainty

Anonymous ACL submission

Abstract

Explainable AI methods facilitate the understanding of model behaviour, yet, 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. 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 perturbations and measure the effect on the output of pre-trained language models and different uncertainty metrics. Realistic perturbations have minimal effect on performance and explanations, yet masking has a drastic effect. We find that high uncertainty doesn't necessarily imply low explanation plausibility; the correlation between the two metrics can be moderately positive when noise is exposed during the training process. This suggests that noise-augmented models may be better at identifying salient tokens when uncertain. Furthermore, when predictive and epistemic uncertainty measures are over-confident, the robustness of a saliency map to perturbation can indicate model stability issues. Integrated Gradients shows the overall greatest robustness to perturbation, while still showing model-specific patterns in performance; however, this phenomenon is limited to smaller Transformer-based language models.

1 Introduction

Though language models 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 unfamiliar 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).

Due to this instability, 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, and if the instability of an explanation can provide insight into the model's performance. 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;

Noise type	Example text
(unperturbed)	"an artful intelligent film that stays within the confines of a well-established genre"
token-MASK	"an [MASK] [MASK] film that stays within the confines of a [MASK] genre"
token-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 where 25% of human-salient tokens are perturbed

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 formatting, like social media), we must understand how this "noise" in the data affects model's performance, confidence, and explainability. As text perturbations can introduce some ambiguity into the data that is not present at training time, they 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 relationship 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, via a controlled experiment by artificially injecting varying degrees and types of realistic noise (see Table 1) and measuring the impact on model explanations and uncertainty. In this manner, we also investigate the relationship between explanation plausibility (the agreement between model saliency and ground-truth annotations) and model uncertainty. To assess if explanation instability reflects model instability, we limit our investigation to gradient-based techniques, rather than model approximation explanations. Here, we provide the following **contributions**:

- We evaluate the relationship between uncertainty and explanation plausibility 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 plausibility 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 high uncertainty does not imply low explanation plausibility; models trained with noisy data can still generate coherent explanations despite high uncertainty amid noise. Furthermore, we argue that explanation instability can give some insight into model performance and can show patterns in saliency attribution: Common, realistic perturbations (like synonym replacement) have smaller effects on model performance and saliency maps, yet l33t speak and token replacement have a larger impact. This pattern is seen typically strongest in Integrated Gradients, which also shows the greatest robustness for smaller language models.

2 Related Work

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, these (often image) studies do not investigate the causes of the instabilities or how they relate to other measures, like uncertainty.

Noise on language model performance Several other studies have looked specifically at the effect

of noise on the performance and confidence of BERT-related models. Surprisingly, there are contrasting effects of noise on machine and human ability to perform natural language understanding tasks. Perturbations that do 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 noise-augmented data (Zhang et al., 2022b). However, as these studies focus on BERT-related models, there is limited focus on other model types, like GPT, and they also do not evaluate explanations.

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. Therefore, we include epistemic uncertainty as a measure in our study.

Uncertainty and XAI Other works in the intersection of uncertainty and XAI 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.

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').

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 sources of uncertainty. In our paper, we investigate the effect of different scales of perturbations on a range of popular language models, including GPT2 and OPT. In addition, to avoid conflating sources of uncertainty, we use multiple measures of uncertainty to assess the relationship between model instability and saliency attribution.

3 Methods

3.1 Datasets

We limit relevant tasks and datasets for this investigation to publicly available English datasets. We select simple, popular text classification tasks 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 have 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 of the sufficiency of highlighted text as explanations (Wiegreffe and Marasović, 2021).

3.2 Models

We test the performance of five 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), $\text{GPT-2}_{\text{medium}}$ (Radford et al., 2019), and OPT-350M (Zhang et al., 2022a), chosen due to their variety in pretraining and their popularity. We describe their finetuning in Appendix A.

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 ($\alpha \in \{0, .05, .10, .25, .50, .70, .80, .90, .95\}$); more details are in Appendix B.1.

We use three **hierarchies** of preferential perturbation: `random`, `human`, and `gradient`. Random-hierarchy is determined randomly, though the pattern of perturbed words is preserved across increasing levels of perturbation. Human-hierarchy 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 perturbation approach in Appendix D.1. Gradient-hierarchy 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. This was determined after taking a subsample of the datapoints and choosing the aggregation method giving 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., 2022b; Moradi and Samwald, 2021). At the word level, we replace words with tokens, such as `MASK`, as done in perturbation-based studies (Madsen et al., 2021). We also compare `MASK` replacement with `UNK` tokens replacement. We convert the entire word to l33t speak (`l33t`) (Eger et al., 2019; Zhang et al., 2022b), 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.2.

3.4 Explanation techniques

We focus on local gradient-based explanations, which use backpropagation to compute a saliency heatmap over input features for a specific datapoint to audit a model's decision. These explanations have been shown to perform best across many metrics, models, and tasks (Atanasova et al., 2020), and, compared to perturbation-based techniques, like LIME, which approximate model performance, are model-centric and should give a more faithful representation of the instability within the model, rather than the technique. The simplest implementation uses the gradient of the input as the saliency score (Simonyan et al., 2013); however, this can be very noisy (Smilkov et al., 2017). Therefore, we rely on modified versions: **SmoothGrad** (`SG`) returns the average saliency map obtained by 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). **InputXGradients** (`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 (Kokhlikyan et al., 2019).

3.5 Evaluation design

Throughout our investigations, we use the following metrics: To measure model performance, we use **accuracy**. To measure uncertainty, we use two measures: Following similar perturbation studies, we include **predictive uncertainty**, the conflated,

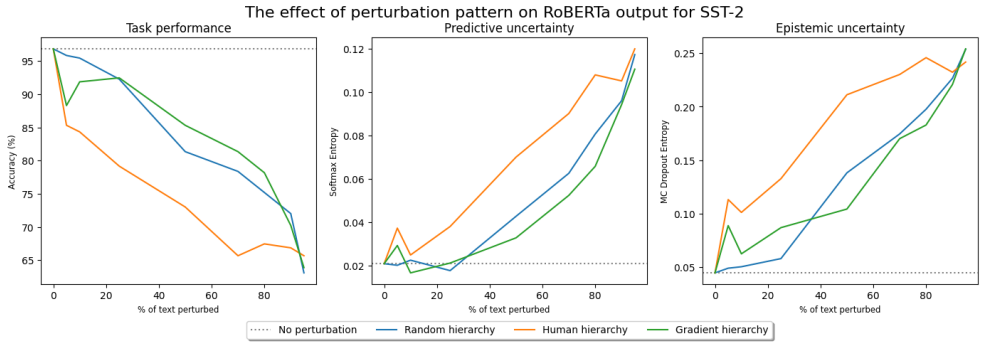

Figure 1: The effect of increasing text perturbation (averaged across perturbation type) on RoBERTa performance and uncertainty across three different hierarchies: (1) Random; (2) Human, following human annotation and POS tags; and (3) Gradient, following ranking of Hotflip gradients. Dotted lines show the value at $\alpha = 0.0$.

popular measure of uncertainty, measured via the entropy of the softmax logits (to reduce overconfidence (Pearce et al., 2021)). We define **epistemic uncertainty**, the uncertainty in the model's parameters, using MC Dropout entropy, following Kendall and Gal (2016). We define **explanation robustness** as the average Pearson correlation between saliency map$_{\alpha = .05}$ and saliency map$_{\alpha = .00}$. We also define **explanation plausibility** as the Mean Average Precision (MAP) of model gradients to the human annotations. There are many metrics to evaluate explanation quality, and each with pitfalls (Ju et al., 2022); we chose this one for its applicability for human-XAI collaboration and evaluation. We first evaluate its suitability by assessing the change in model performance and confidence between the perturbation of human- and gradient-ranked salient tokens. For all saliency map comparisons, we combine all gradients back to word level.

We first look at general trends in performance, uncertainty, and explanation plausibility with increasing perturbation across models and datasets. As a Kolmogorov–Smirnov test of the plausibility and uncertainty measures violates the assumption of normality ($p < 10^{-5}$), we use Spearman's Rank Correlation (SciPy v1.11.4) to find the correlation between the explanation plausibility and uncertainty measures at a datapoint level for correctly-predicted datapoints. We then compare the robustness of the saliency maps across model, dataset, and perturbation type.

4 Results

We present the motivation and results of each investigation. We show the results for SST-2 but summarize and interpret the results for all investigated datasets; the data for all datasets are in the Appendix.

4.1 Noise on uncertainty and explanations

4.1.1 The effect of perturbation hierarchy

To ensure the faithfulness of using human annotations as ground-truth for salient tokens (Ju et al., 2022), we showcase the impact of different hierarchies of perturbation (as described in §3.3) on model performance, uncertainty and explanations in Figure 1 and Appendix C.1. All perturbations impair model performance, uncertainty, and explanation plausibility, but human-hierarchical perturbation has the greatest impact up to very high levels of perturbation across all tasks, suggesting that these human-salient tokens are vital signals for the models. While random and gradient-based perturbation generally have similar impacts on task performance, uncertainty and explanation plausibility, gradient-based perturbation strategies have a stronger impact on these metrics at low levels of perturbation ($\alpha = .05$), which lessens with slightly more perturbation ($\alpha = .1$), and suggests that gradient-based perturbation techniques have their greatest efficacy at low levels of perturbation.

4.1.2 The effect of perturbation type

To assess the impact of our various perturbation types, we show the effect of the investigated noise types (see Table 1) in Figure 2. Though all perturbation types adversely impact task performance and explanation plausibility, this effect is typically smaller for more 'realistic' perturbations, especially `synonym` and `butterfingers`. Across datasets and models, we typically see that special token replacements have the greatest detrimental effect (particularly `MASK`). This trend can be seen across models and datasets, as shown in Appendix

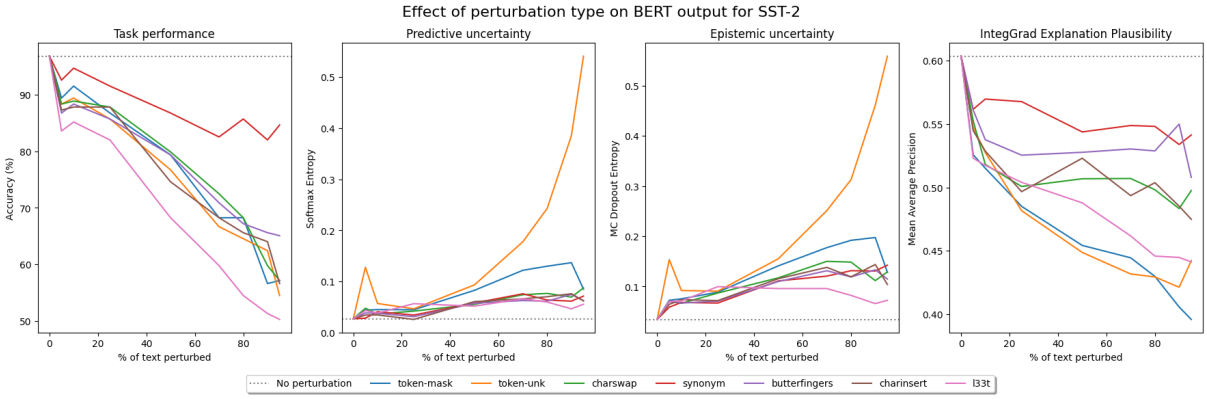

Figure 2: The effect of increasing text perturbation on BERT accuracy, uncertainty and explanation plausibility across the different types of perturbation, averaged across perturbation hierarchy. Dotted lines show $\alpha = 0.0$.

C.2. Typically, we see an inverse relationship between task performance and uncertainty for all perturbation types; In Figure 2, a steep decrease in performance from `l33t` perturbation comes with little to no increase in either uncertainty measure. Furthermore, while we do often see a trend for uncertainty to increase with increased perturbation, it is not at the same rate of performance decline, mimicking over-confidence issues reported in other studies (Feng et al., 2018; Gupta et al., 2021; Pearce et al., 2021). As we show in Appendix C.2, this is especially pronounced with RoBERTa.

4.2 The relationship between uncertainty and explanation plausibility

While we see a general increase in uncertainty and a decrease in explanation quality with perturbation, we want to investigate at a data point level if greater uncertainty of a model output implies lower plausibility across explanation techniques. Therefore, we assess the correlation between uncertainty and explanation plausibility across all datasets, saliency maps, and models in Table 3. We see similar patterns in correlation between all attribution methods and before and after perturbation with some exceptions: SmoothGrad (`SG`) typically shows much weaker correlation after perturbation, whereas Guided Backpropagation (`GBP`) and Integrated Gradients (`IG`) show the strongest. While we would expect increased uncertainty to imply decreased explanation plausibility, this is not always the case; While SST-2 shows a low negative correlation between epistemic uncertainty and explanation plausibility after perturbation, this is not the case for SemEval and HateXplain, which has a low-to-moderate positive correlation that exists before and after perturbation. Therefore, greater uncertainty of an output does not necessarily imply a degradation in explanation quality.

4.3 Robustness across perturbation type

To investigate how noise introduces instability in explanation maps, we assess how saliency maps are impacted by our perturbations. We show the robustness values across each model in Figure 3 and see distinct patterns that are shared across most saliency maps. For example, Integrated Gradients (`IG`), InputXGrad (`IXG`), and Guided Backpropagation (`GBP`) all show reduced robustness to `l33t` perturbations and increased robustness to `synonym` at low-levels of perturbation. This aligns with the differences we see in task performance in Figure 2. We also see different patterns emerge across different models; RoBERTa has general lower robustness to `UNK` tokens, and ELECTRA to `MASK`. `IG` shows the greatest overall robustness for the models BERT, RoBERTa, and ELECTRA, but SmoothGrad (`SG`) has greater robustness for GPT2 and OPT. `GBP`, which typically has low robustness for all other models, shows the greatest robustness for OPT. These patterns are preserved across datasets (See Appendix C.3), where perturbations to which we find decreased robustness typically also further deteriorate model performance. Furthermore, on datasets with lower performance (HateXplain), we also see decreased overall robustness.

In summary, while perturbation decreases model performance and explanation plausibility, it has a task and perturbation-dependent effect on uncertainty. Furthermore, high uncertainty of an output does not necessarily imply low explanation plausibility, as we find a moderate positive correlation be-

		Before Perturbation								Including Perturbed Text							
		Predictive uncertainty				Epistemic uncertainty				Predictive uncertainty				Epistemic uncertainty			
dataset	**model**	**GBP**	**IXG**	**IG**	**SG**	**GBP**	**IXG**	**IG**	**SG**	**GBP**	**IXG**	**IG**	**SG**	**GBP**	**IXG**	**IG**	**SG**
	BERT	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
	ELECTRA	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
SST-2	RoBERTa	**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
	GPT2	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
	OPT	**-0.205**	**-0.205**	-0.030	-0.030	**-0.159**	**-0.159**	-0.099	-0.099	**-0.152**	**-0.152**	-0.130	-0.130	**-0.219**	**-0.219**	-0.109	-0.109
	BERT	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
	ELECTRA	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
SemEval	RoBERTa	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
	GPT2	**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
	OPT	0.224	0.224	**0.226**	**0.226**	**0.230**	**0.230**	0.224	0.224	0.164	0.164	**0.165**	**0.165**	**0.167**	**0.167**	0.164	0.164
	BERT	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	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
HateXplain	RoBERTa	**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	**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
	OPT	**0.459**	**0.459**	0.428	0.428	0.432	0.432	**0.456**	**0.456**	**0.408**	**0.408**	0.391	0.391	0.382	0.382	**0.410**	**0.410**

Table 3: The Spearman Rank Correlation between explanation plausibility and both measures of uncertainty across model, dataset, and saliency technique. We bold the strongest correlation for each comparison.

tween the measures on some datasets. Where uncertainty measures fail to align with model accuracy on some perturbation patterns, saliency map robustness can provide additional indication of model performance patterns; Integrated Gradients typically shows the greatest robustness to all types of noise; however, we can see model-specific patterns in susceptibility to adversarial perturbation.

5 Discussion

While noise consistently deteriorates model performance and explanation plausibility, 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); though the observed decrease is not at the same rate of performance decline. However, both cited studies perturb at the word and sentence structure level, unlike our study. Furthermore, we see a similar pattern between the predictive and epistemic uncertainty measures suggesting that overconfidence after perturbation stems from the training process. Overall, human-based perturbations have the strongest effect on task performance and uncertainty measures, and gradient-based perturbation is only more effective than random perturbation at low levels of noise ($\alpha = .05$). This suggests that the human annotations are faithful indicators of saliency, as their perturbation degrades model performance more than gradient-based approaches, justifying our use of plausibility as a quality metric.

Across all models, realistic perturbations, such as `butterfingers` or `synonym` have the smallest impact on task performance and explanation plausibility, yet masking has the greatest impact. Furthermore, `MASK` has the greatest effect on both measures of uncertainty. Furthermore, when we look at model-level comparisons in Appendix C.2 and C.3, we find that the perturbations that have the maximal impact on model performance (e.g. `MASK` for ELECTRA in SemEval), also have a unique impact on saliency maps at low levels of perturbation, suggesting that the instability we see in explanations can provide some signal to model performance in the absence of labels. While SmoothGrad shows good all-around robustness to noise due to its regularization, it does not show specific patterns in model instability. In contrast, Integrated Gradients has relatively high robustness for smaller language models at low levels of perturbation and shows increased robustness to perturbations that minimally impact model performance (`synonym` and `butterfinger`). While lack of robustness is typically viewed as a deficiency of an XAI technique (Hedström et al., 2023), we believe it can also be a signal of model instability: Epistemic uncertainty approximation measures work to perturb the model decision boundary to obtain a datapoint's likelihood of class correspondence as an indicator of uncertainty. In contrast, perturbation, by introducing stochasticity in an input (much like SmoothGrad), also suggests the proximity of a datapoint to the decision boundary. In cases where popular

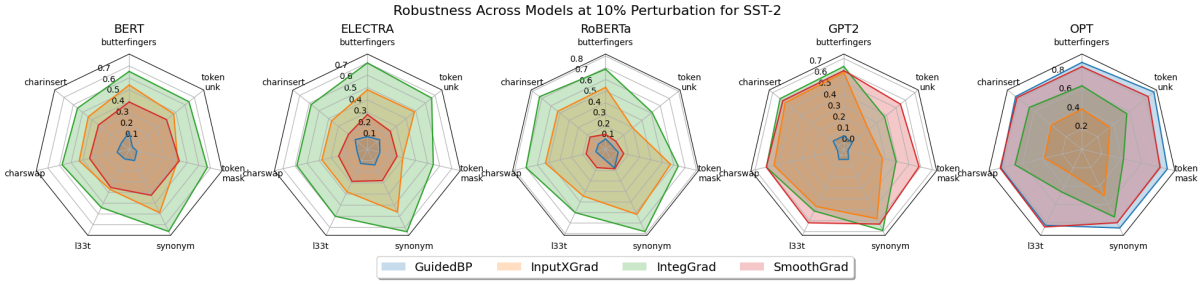

Figure 3: Model-level differences of the correlation to the unperturbed saliency map at low levels of perturbation. We separately show the effect on BERT, RoBERTa, ELECTRA, GPT2 and OPT.

uncertainty measures are prone to over-confidence, a lack of robustness may give some indication of uncertainty for a particular data point.

In the case of hatespeech detection, `UNK` and `l33t` surprisingly reduce uncertainty (see Appendix C.2); this could explain the positive correlation between uncertainty and explanation plausibility for HateXplain, as highly perturbed examples will show lower uncertainty as plausibility decreases. The dataset is compiled from Twitter, and character substitutions may be used to hide potentially offensive terms. While there is no class difference regarding the proportion 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 an indicator of a class, owing to the high "learnability" of this perturbation (Zhang et al., 2022b). So, when noise is learned to be an indicator of class, uncertainty may show a positive correlation with output quality and explanation plausibility. However, we also see a weaker, positive relationship with the Twitter-based SemEval dataset, and we do not see saliency attributed to l33t noise in Appendix D.2; therefore, models trained with noisy data may show this positive relationship. This suggests that when these models have greater uncertainty, they may be more precise at identifying salient tokens. Other studies also suggest performance improvements after training models with noisy data (Yu et al., 2024). We show in Appendix D.4 that, at very high perturbation, the strength of this relationship weakens (due to lack of meaningful tokens), but can remain weakly positive for simple tasks.

Altogether, this suggests that the effect of perturbation on language models must be considered holistically across noise type and training data; realistic perturbations, which are expected to be more present in data drawn from social media, have a smaller impact on performance, uncertainty, and explanations. While uncertainty is not always a faithful indicator of local instability, weakened robustness to perturbations can provide additional information for model performance. Furthermore, as high uncertainty does not necessarily imply low explanation plausibility with noisy datasets, noise-augmented training may not only help model performance in out-of-domain tasks, but may also help ensure coherent explanations in low-confidence domains. For future work, we recommend the use of Integrated Gradients for smaller language models as it gives a more holistic depiction of model performance in adversarial conditions; however, as models scale in size, other gradient-based explanation techniques are more robust.

6 Conclusion

We provide an empirical investigation across language models, noise perturbations, and saliency maps to investigate a relationship between uncertainty and explanation plausibility. Following an array of perturbation techniques, we show that noise injection simultaneously affects model performance, uncertainty, and explanation plausibility. We do not find a strong negative relationship between uncertainty and explanation plausibility; model fine-tuned with noisy data typically show a moderately positive correlation between plausibility and uncertainty, which suggests that these models may even be better at identifying salient tokens when uncertain. We also show that the instability of a saliency map to noise can also provide insights into a model's performance, and suggest Integrated Gradients for future work in Human-XAI collaboration, due to its robustness to noise for smaller language models.

Limitations

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

Though we do compare many popular language models, more model types would have made an interesting comparison. 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 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. It would also have been interesting to explore larger language models ($>$ 1B parameters), like LLAMA (Touvron et al., 2023); however, our focus on gradient-based explanations makes such an investigation very computationally expensive. Furthermore, our requirement for human annotations limited the possible number of datasets for investigation; however, our pilot studies on other popular NLP tasks found very similar results to those reported in this study.

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).

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.

Yiming Ju, Yuanzhe Zhang, Zhao Yang, Zhongtao Jiang, Kang Liu, and Jun Zhao. 2022. Logic traps in evaluating attribution scores. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5911–5922, Dublin, Ireland. Association for Computational Linguistics.

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.

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.

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.

Xiaowei Yu, Yao Xue, Lu Zhang, Li Wang, Tianming Liu, and Dajiang Zhu. 2024. Exploring the Impact of Information Entropy Change in Learning Systems. Accepted to ICLR 2024.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Singh Koura, Anjali Sridhar, Tianlu Wang, Luke Zettlemoyer, and Meta Ai. 2022a. OPT: Open Pre-trained Transformer Language Models.

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

A Model training specifications

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 $\text{BERT}_{\text{base}}$,, which is 110 million parameters. We use $\text{RoBERTa}_{\text{base}}$, which is 125 million parameters. ELECTRA is 110 million parameters. We rely on $\text{GPT2}_{\text{medium}}$, which is 345 million parameters, and OPT-350M, which is 350 million parameters. BERT, RoBERTa, and ELECTRA are trained and assessed on Titan RTX GPUs; GPT2 and OPT are 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[1]. While we cannot find hyperparameters reaching the performance described in the original RoBERTa-base (94.8%) article (Liu et al., 2019), we choose the hyperparameters specified by this model card [2], 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) and achieves an accuracy of 92% on the evaluation set. For OPT, we used the hyperparameters specified by the huggingface model card[3], which achieved an accuracy of 91% on the evaluation set.

SemEval					
model	**BERT**	**RoBERTa**	**ELECTRA**	**GPT2**	**OPT**
Learning rate	1e-5	1e-5	3e-6	8e-5	7e-6
Batch size	16	16	8	32	32
Epochs	3	3	5	7	1
Random seed	37	37	24	42	42
Adam ϵ	1e-8	1e-8	1e-8	1e-8	1e-8
Adam β 1	0.9	0.9	0.9	0.9	0.9
Adam β 2	0.999	0.999	0.999	0.999	0.999
LLRD	None	None	None	None	None
Decay type	Linear	Linear	Linear	Cosine	Cosine
Warmup Fraction	0	0	0	0.01	0.01
Attention Dropout	0.1	0.1	0.1	0.1	0.1
Dropout	0.1	0.1	0.1	0.1	0.1
Weight Decay	0	0	0	0.1	0.1
Test Accuracy	92%	94%	91%	91%	91%

Table 4: Final hyperparameters for all investigated models on the SemEval dataset

A.2 SemEval, 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 for SemEval, and HateXplain are shown in Tables 4 and 5.

B Perturbation specifications

B.1 Proportion replaced

When perturbing a text by α α indicates the proportion of the text that is modified by a perturbation type. Texts are never fully perturbed, so, if $\alpha = 0.95$, and the length of the text (N_{tokens}) is fewer than 20 tokens, at least one token is left unmodified. For very short texts, as seen in SST-2, the data point is left unmodified until $\alpha \times N_{tokens} \geq 1$. Effects of perturbation level are only assessed at an aggregated level to visualize the rate of metric change with increasing perturbation.

B.2 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 replaced 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

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

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

[3] https://huggingface.co/tianyisun/opt-350m-finetuned-sst2

HateXplain					
model	**BERT**	**RoBERTa**	**ELECTRA**	**GPT2**	**OPT**
Learning rate	2e-5	6e-6	2e-5	5e-5	9e-6
Batch size	32	32	8	32	8
Epochs	5	5	2	6	1
Random seed	2	2	6	42	42
Adam ϵ	1e-8	1e-8	1e-8	1e-8	1e-8
Adam β 1	0.9	0.9	0.9	0.9	0.9
Adam β 2	0.999	0.999	0.999	0.999	0.999
LLRD	None	None	None	None	None
Decay type	Linear	Linear	Linear	Cosine	Cosine
Warmup Fraction	0	0	0	0.01	0.01
Attention Dropout	0.1	0.1	0.1	0.1	0.1
Dropout	0.1	0.1	0.1	0.1	0.1
Weight Decay	0	0	0	0.1	0.1
Test Accuracy	68%	69%	70%	66%	69%

Table 5: Final hyperparameters for all investigated models on the HateXplain dataset

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 Model and dataset-level differences

While the results for BERT and SST-2 are visualized in the article, we provide the results for all investigated models and datasets below.

C.1 The effect of perturbation hierarchy on uncertainty and explanations

We show the results of our investigations into the effect of perturbation hierarchies (see §4.1) across our investigated datasets in Figure 4. **Results:** Across all 4 datasets, we find that human-hierarchied perturbation has the strongest impact on task performance, uncertainty, and explanation plausibility. Furthermore, we can see that, while gradient and random-hierarchical perturbation has typically quite similar impact, the difference is greatest at low ($\alpha = .05$) levels of perturbation, and begins to diminish at higher levels ($\alpha = .1$). Interestingly, we see high levels of perturbation have a parabolic relationship with uncertainty in the HateXplain dataset.

C.2 The effect of perturbation type on model output

We show the results of our investigations into the differential effect of perturbation types (see §4.1) across our datasets and models. We present the effect of perturbations across models on accuracy and predictive and epistemic uncertainty for the SST-2 dataset in Figure 5, and the effect of perturbations on explanation plausibility in Figure 6. Similarly, we provide model-specific graphs for the SemEval dataset in Figures 7 and 8, and the HateXplain dataset in Figures 9 and 10. **Results:** We typically see a steep decrease in accuracy with increasing perturbation across all perturbations, models, datasets. Typically, we also see an increase in predictive and epistemic uncertainty; though, we see some exceptions with the HateXplain datasets, and this increase does not always correspond to the performance decrease. We generally see a decreasing trend in explanation plausibility with increasing perturbation, but this relationship is not as strong as the other observed metrics. Typically, this decrease is steepest with Integrated Gradients and all BERT explanations. Interestingly, we do not see SmoothGrad explanation plausibility change with perturbation with OPT and GPT2, which is related to the robustness of the combination as we see in Appendix C.3 and the poor initial plausibility score.

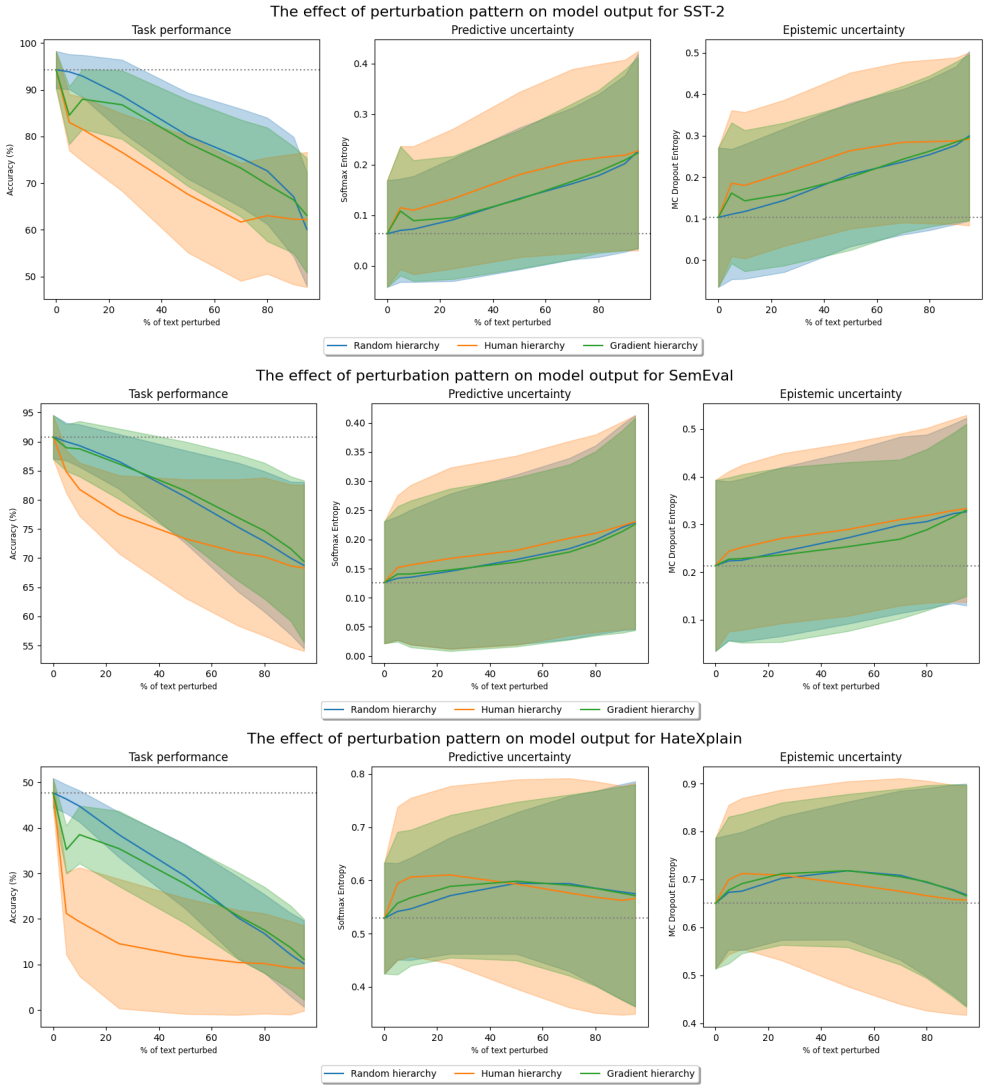

Figure 4: We show the differential effect of our perturbation hierarchies across the different datasets investigated. Values are averaged over all 7 perturbation types.

Across all datasets, we find similar behaviour between special token replacements (`token-unk` and `token-mask`) as well as between character-level changes (`charswap`, `charinsert`, `butterfingers`). `synonym` and `butterfingers` typically have the smallest effect on all measures. Interestingly, `l33t` has a very task-dependent effect: For SemEval, it has a moderate effect on all model outputs. In HateXplain, it has a very strong negative effect on model performance and explanation plausibility, yet decreases model uncertainty. Generally, we see increasing uncertainty with increasing levels of perturbation for all models and noise types as well as decreasing accuracy. Typically, perturbations decrease accuracy at a similar rate as they increase uncertainty, except in the case of `l33t`, `UNK`, and `MASK`. With SST-2, `l33t`, `UNK`, and `MASK` perturbations most impact all models' accuracy, yet we do not see this reflected in the uncertainty curves. Similarly, for HateXplain, these perturbations reduce uncertainty for BERT, ELECTRA, RoBERTa, and OPT. Overall, GPT2 outputs much greater predictive and epistemic uncertainty relative to the other base models, and RoBERTa shows only slight increase in uncertainty with increased perturbation, even when accuracy is just as perturbed as other models (SST-2), suggesting a tendency for over-confidence.

C.3 Robustness across perturbation types

We look at model-level differences in saliency map robustness across datasets at low levels of perturbation in Figure 11. We look at robustness at high levels of perturbation in Appendix D.3. **Results**: Typically, we see the greatest overall robustness across all perturbation types for Integrated Gradient. However, GPT2 and OPT typically has the greatest robustness with SmoothGrad. Guided backpropogation is typically very unrobust for all models, save for OPT, where it shows surprising robustness. For all models, we see a similar shape on the radar plot appear by Guided backpropagation, Integrated Gradients and InputXGrad, which is consistent for each model across datasets. Saliency maps on BERT typically show decreased robustness to `l33t` perturbation, and increased robustness to `butterfingers` and `synonym`. For RoBERTa and GPT2, we see a decreased robustness to `UNK` and `l33t` perturbations, whereas robustness is lowest to `MASK` perturbations for ELECTRA. OPT shows consistent poor robustness to `l33t`, and relatively low robustness to `UNK` and `MASK`, though the extent of this changes between tasks.

D Extra investigations

D.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 12, and denote the average location of a change in strategy with a dotted line. **Results**: We can see that POS-hierarchied 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.

D.2 Saliency map correlation to noise

To assess if decreased robustness of a saliency map technique to a particular perturbation type stems from attribution of saliency to the perturbed input, we assess the Pearson correlation of the output saliency map to the perturbed tokens, and visualize the output across dataset and model in a radar plot in Figure 13. **Results**: While we see equivalent lack of correlation to all types of noise for InputXGrad and GuidedBP saliency maps, SmoothGrad shows differing behaviour according to model type. For most models, SmoothGrad shows a slight negative correlation to `l33t` and `UNK` tokens; however, SmoothGrad does not show this particular aversion to `UNK` with RoBERTa and it does not show a particular aversion to `l33t` with GPT2. Furthermore, SmoothGrad applied on ELECTRA shows a consistent aversion to `UNK` and `l33t` tokens. All correlation values are very low, with a magnitude under 0.3. This behaviour for SmoothGrad may stem from its regularization process, and may also give some indication of model instability or stability, as these specific perturbations also have a strong detrimental effect on model performance. Ultimately, no model and saliency map combinations appear to preferentially attribute saliency to any type of perturbed tokens/words.

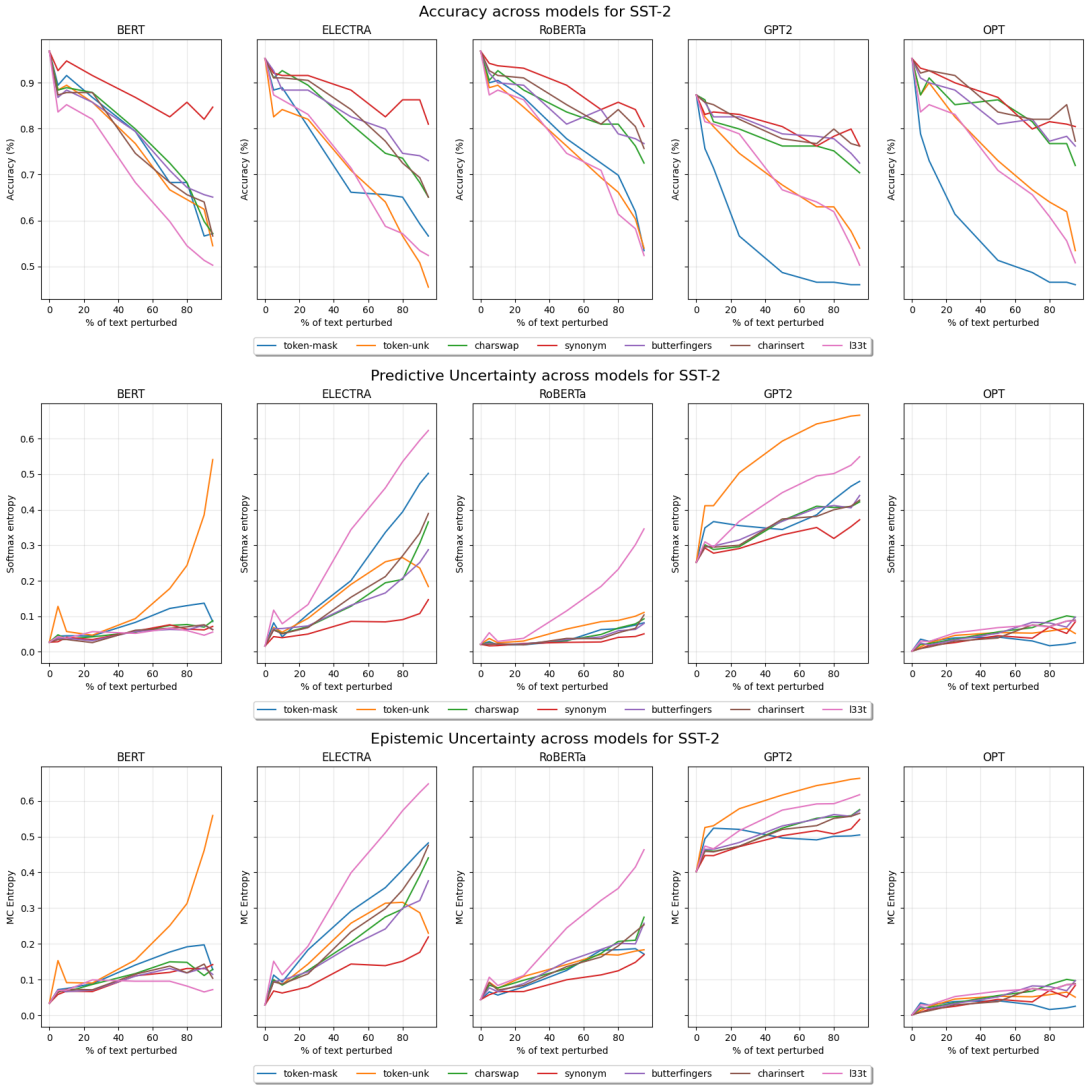

Figure 5: We show the differential effect of increasing levels of text perturbation on model accuracy and both measures of uncertainty on the **SST-2** dataset. Values are averaged over all hierarchies.

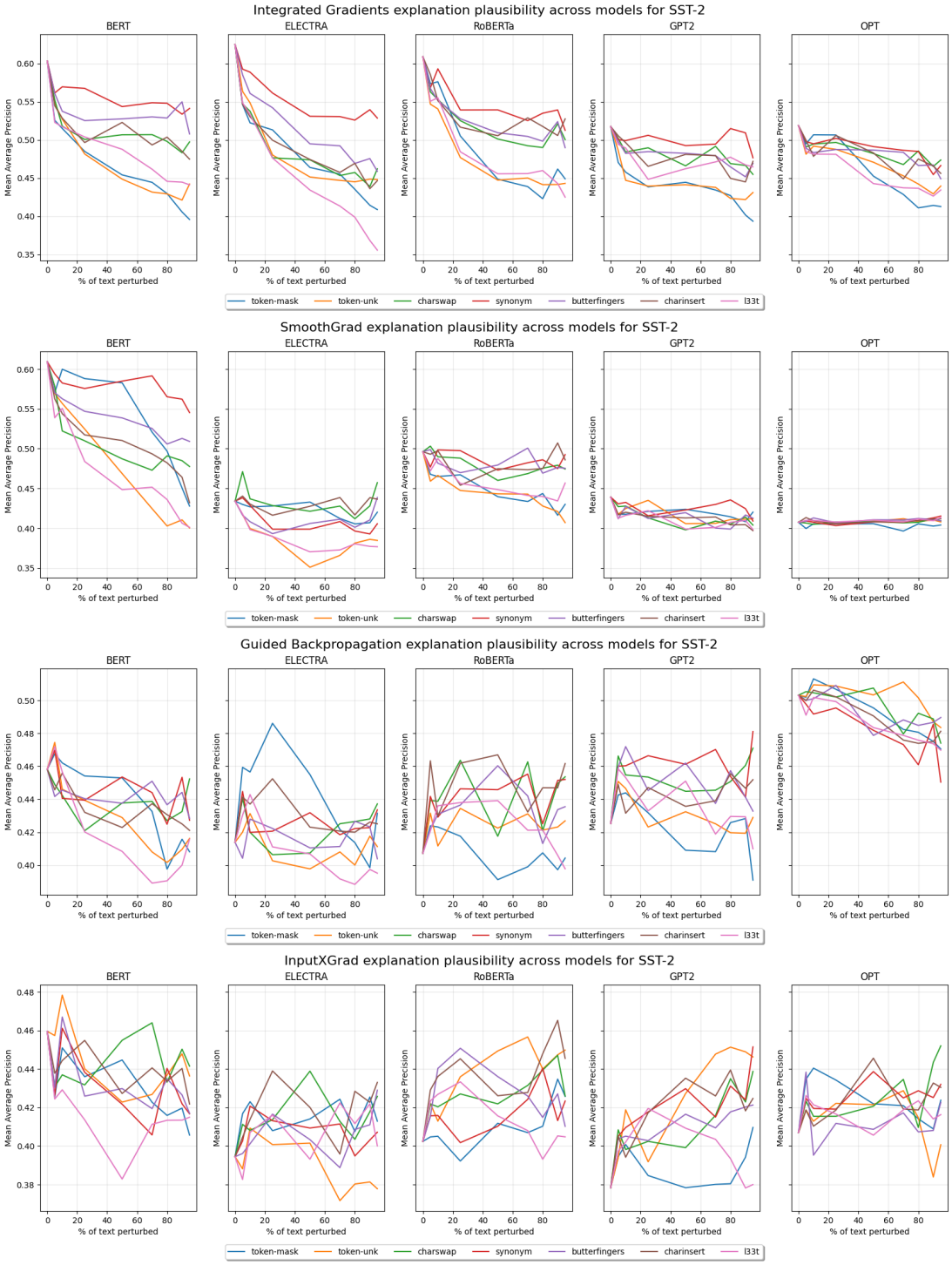

Figure 6: We show the differential effect of increasing levels of text perturbation on the explanation plausiblity of all saliency maps **SST-2** dataset. Values are averaged over all hierarchies.

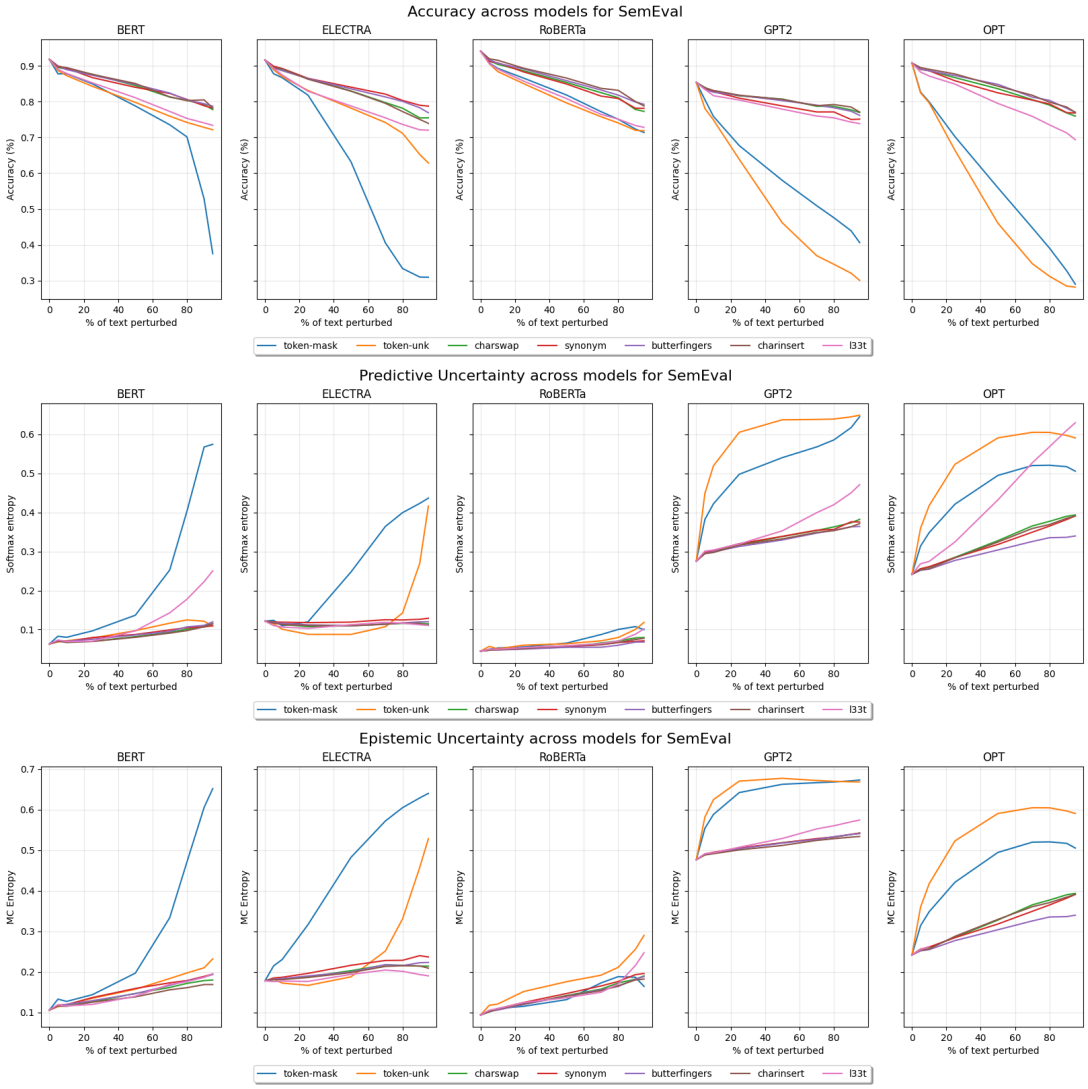

Figure 7: We show the differential effect of increasing levels of text perturbation on model accuracy and both measures of uncertainty on the **SemEval** dataset. Values are averaged over all hierarchies.

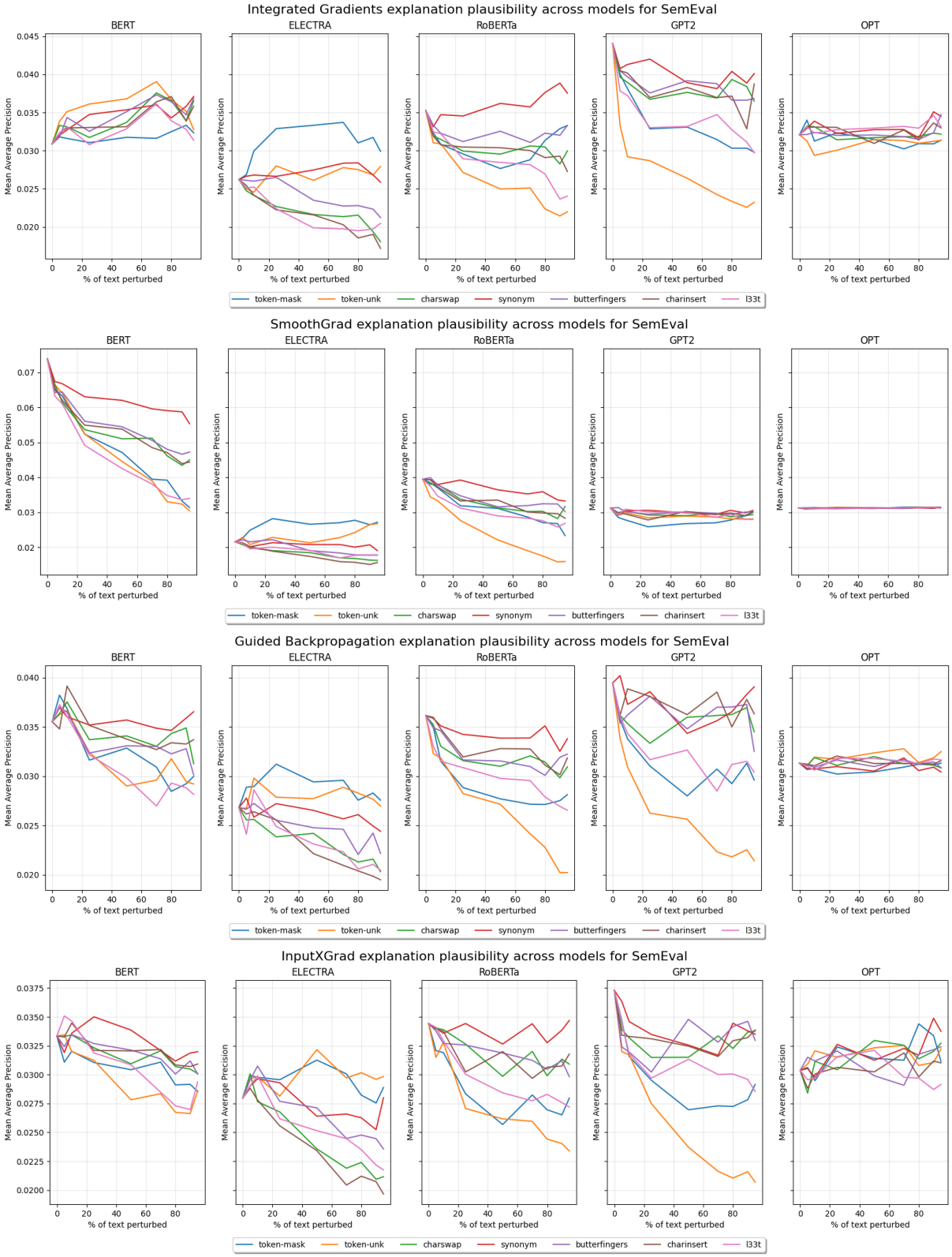

Figure 8: We show the differential effect of increasing levels of text perturbation on the explanation plausiblity of all saliency maps **Semeval** dataset. Values are averaged over all hierarchies.

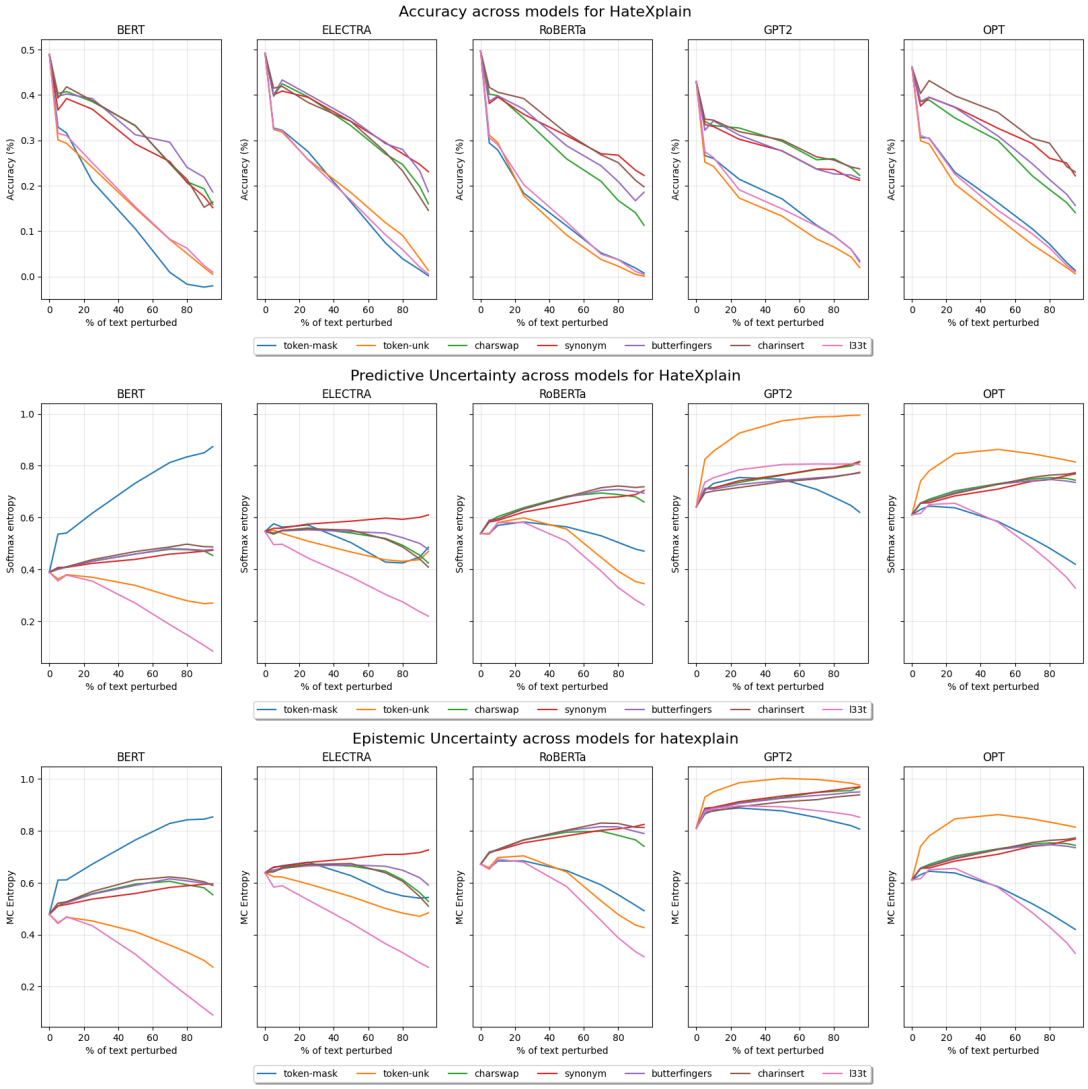

Figure 9: We show the differential effect of increasing levels of text perturbation on model accuracy and both measures of uncertainty on the **HateXplain** dataset. Values are averaged over all hierarchies.

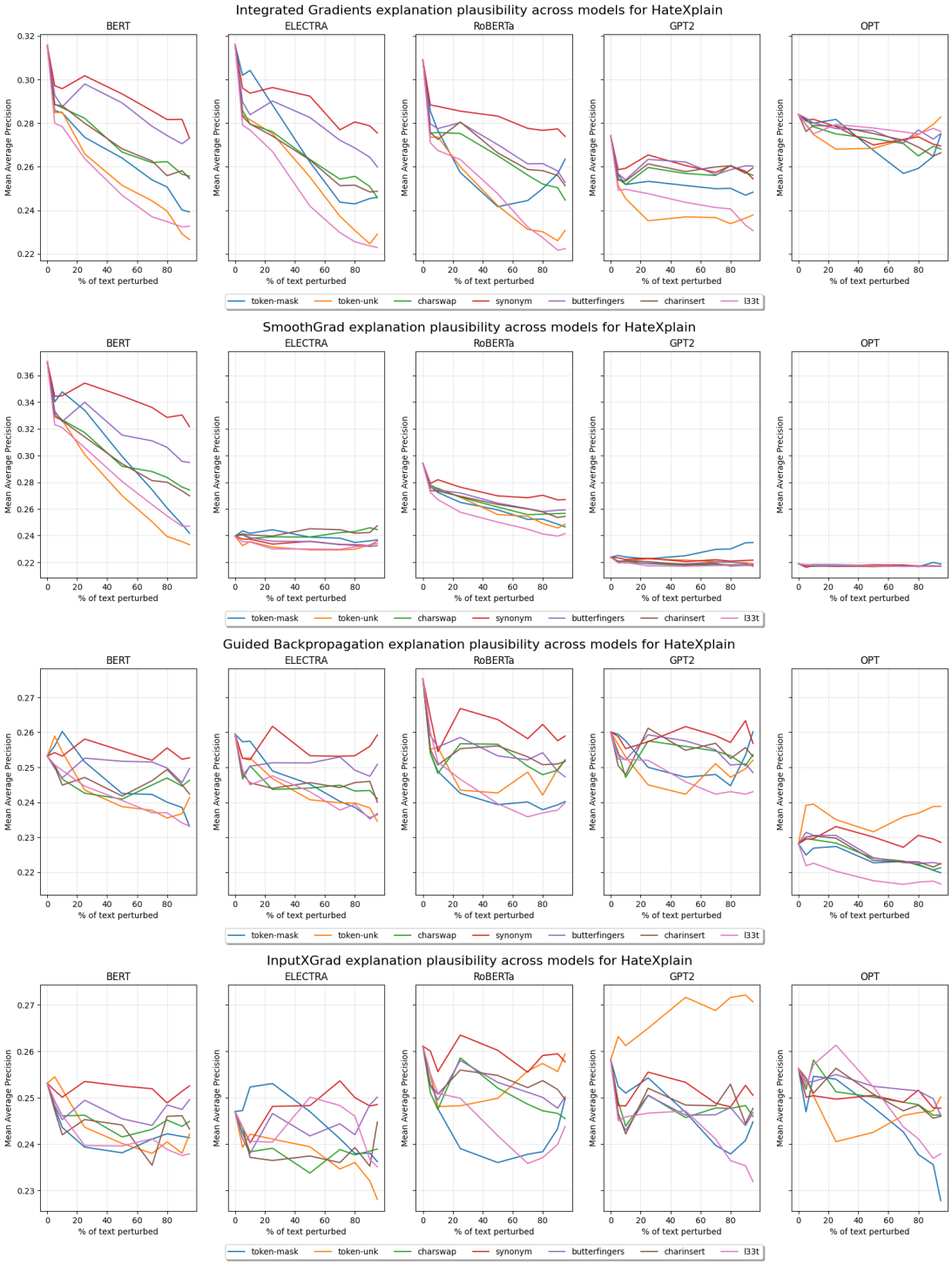

Figure 10: We show the differential effect of increasing levels of text perturbation on the explanation plausiblity of all saliency maps **HateXplain** dataset. Values are averaged over all hierarchies.

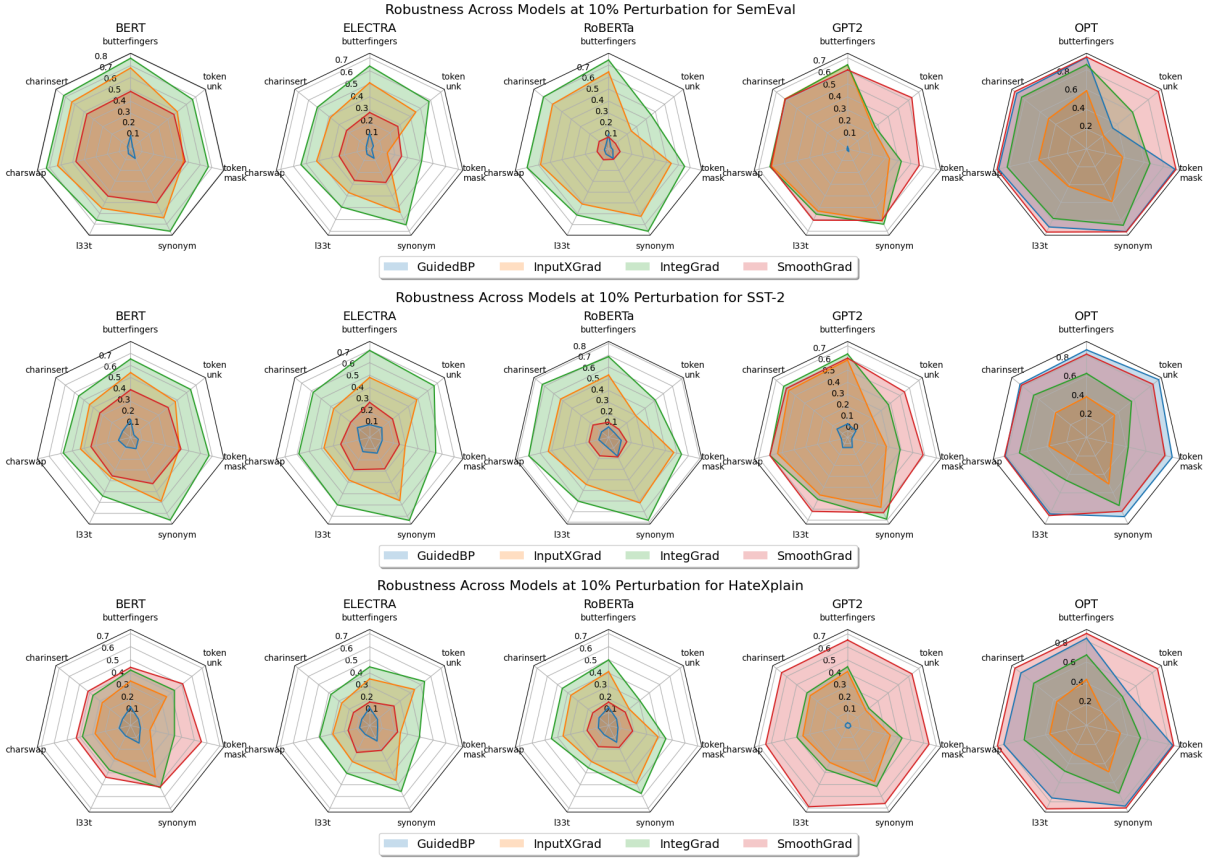

Figure 11: We show the robustness across models of our saliency maps at low levels of perturbation across different tasks

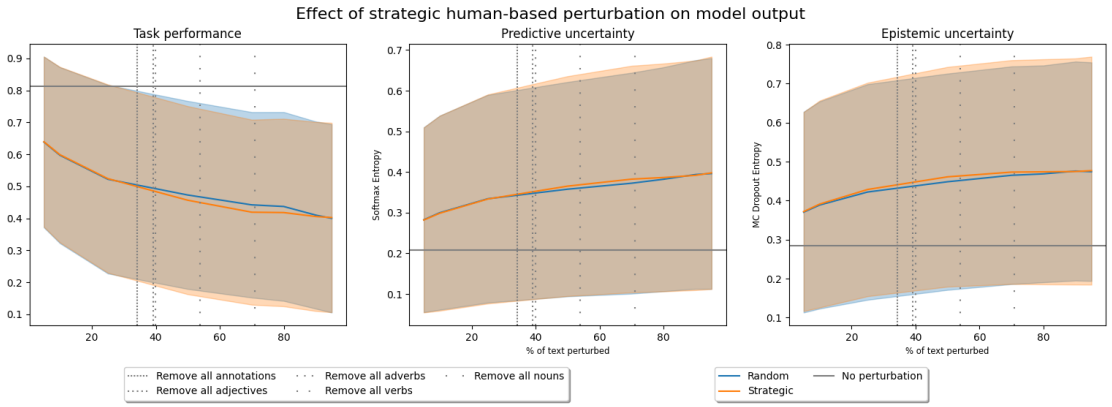

Figure 12: We compare the effect of two different methods of human-based perturbation on model accuracy, confidence and explanation plausibility. Human-Random randomly perturbs tokens after all annotated tokens are perturbed. Human-Strategic preferentially perturbs tokens based on their POS. Vertical lines denote the average location of strategy shift for the Human-Strategic perturbation hierarchy.

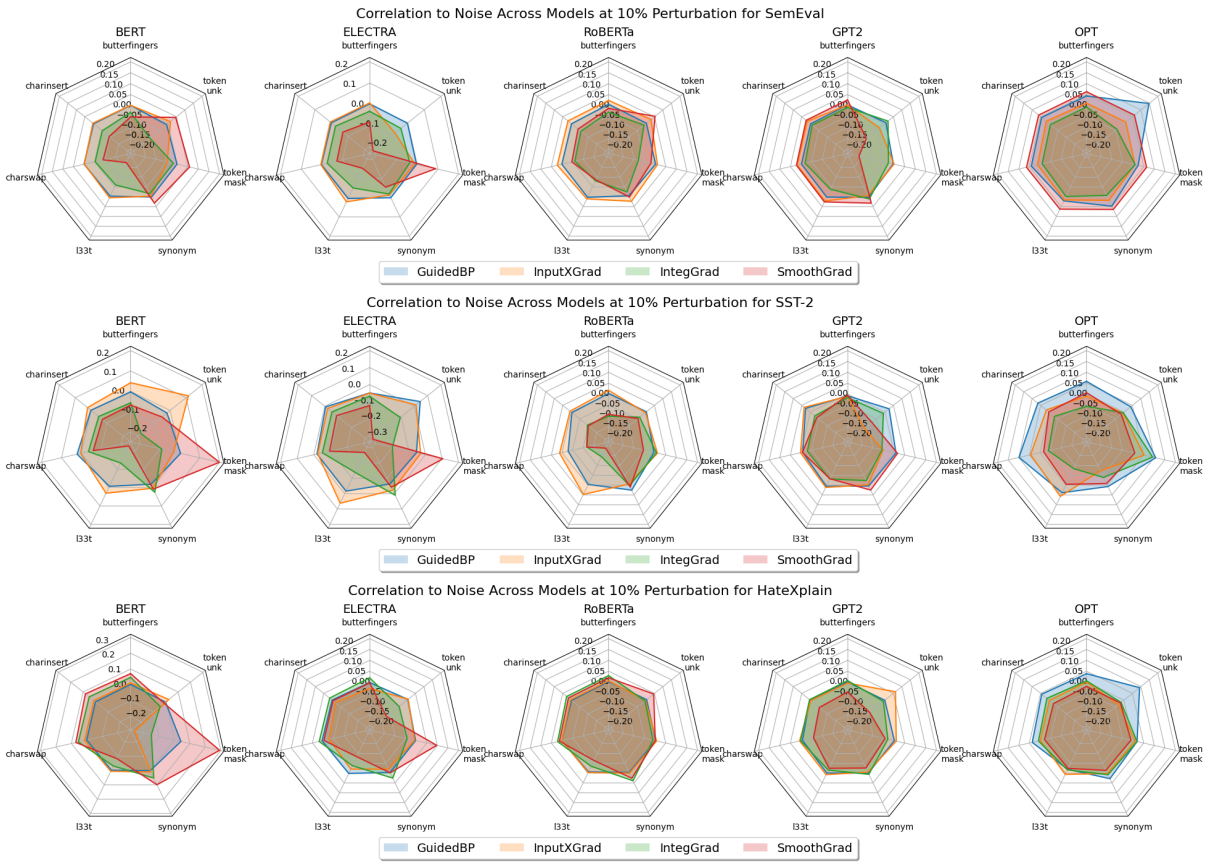

Figure 13: We show the correlation to noise across models of our saliency maps at low levels of perturbation across different tasks

D.3 Saliency map robustness at high levels of noise

To assess saliency map robustness at high levels of noise, we present the same investigation performed in §4.3 but with $\alpha = .5$ in Figure 14. **Results**: We see similar patterns as those described in Appendix C.3, but overall lower robustness. One exception is SmoothGrad on the larger language models (GPT2 and OPT) and Guided Backpropagation for OPT, which still seem to show high general robustness. However, though the general patterns of the saliency map$_{\alpha = 0.0}$ is preserved, we can see in Figures 6, 8, 10, the quality of the original saliency maps are typically quite low, in terms of agreement to human annotations. Similarly to §4.3, we can see that robustness is typically lower to `l33t` and `UNK` perturbations for BERT and GPT2, `l33t` and `MASK` for ELECTRA, and `l33t`, `MASK` and `UNK` for OPT.

D.4 Uncertainty and explanation plausibility at high levels of perturbation

We investigate the correlation between explanation plausibility and our two uncertainty measures at very high levels of perturbation ($\alpha \in \{.90, .95\}$) in Table 6, 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 plausibility. SemEval, which is an easier task than HateXplain, seems to conserve a very weak positive relationship between uncertainty and explanation plausibility 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.

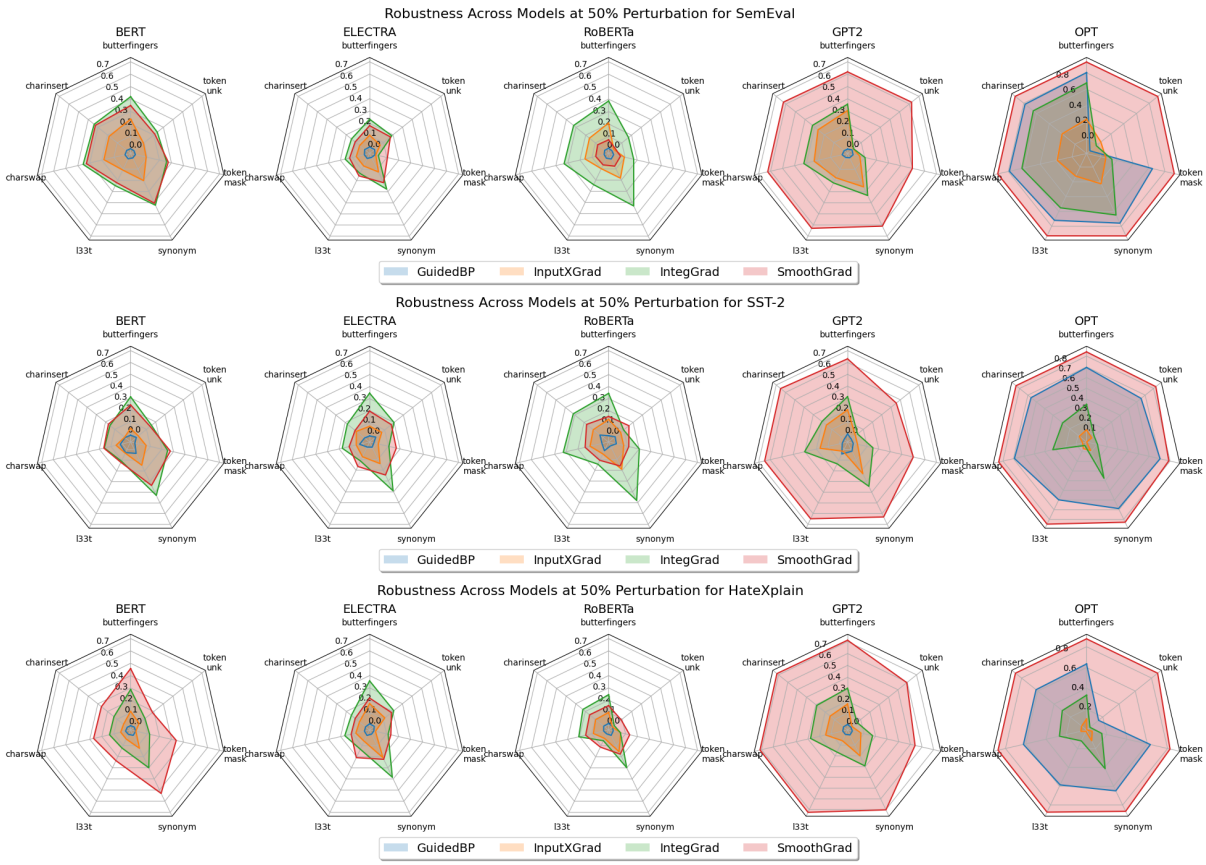

Figure 14: We show the robustness of our saliency maps at high levels of perturbation ($\alpha = 0.50$) across different tasks

		Predictive Uncertainty				Epistemic Uncertainty			
Dataset	**Model**	**GBP**	**IXG**	**IG**	**SG**	**GBP**	**IXG**	**IG**	**SG**
	BERT	-0.016	0.020	-0.015	**0.092**	**-0.162**	-0.100	-0.089	-0.011
	ELECTRA	**-0.122**	-0.114	-0.048	-0.032	**-0.308**	-0.289	-0.160	-0.151
SST-2	**RoBERTa**	**-0.169**	-0.123	-0.153	-0.130	**-0.315**	-0.254	-0.244	-0.178
	GPT2	**-0.075**	-0.017	-0.070	-0.016	**-0.159**	-0.100	-0.096	-0.048
	OPT	-0.013	-0.013	**-0.054**	**-0.054**	**-0.070**	**-0.070**	-0.020	-0.020
	BERT	0.088	**0.103**	0.088	**0.103**	0.089	**0.104**	0.087	0.103
	ELECTRA	**0.103**	0.096	**0.103**	0.096	**0.105**	0.097	0.104	0.097
SemEval	**RoBERTa**	**0.106**	**0.106**	**0.106**	**0.106**	**0.108**	0.106	0.104	0.104
	GPT2	0.064	**0.083**	0.065	**0.083**	0.065	**0.085**	0.065	0.084
	OPT	**0.074**	**0.074**	0.073	0.073	0.067	0.067	**0.076**	**0.076**
	BERT	-0.049	**-0.078**	-0.049	**-0.078**	-0.040	-0.060	-0.041	**-0.064**
	ELECTRA	-0.054	-0.084	-0.061	**-0.091**	-0.033	-0.059	**-0.060**	-0.090
HateXplain	**RoBERTa**	-0.021	-0.054	-0.023	**-0.055**	-0.009	-0.036	-0.020	**-0.052**
	GPT2	0.134	0.090	**0.140**	0.094	0.126	0.097	**0.134**	0.092
	OPT	**0.019**	**0.019**	0.003	0.003	**-0.025**	**-0.025**	0.005	0.005

Table 6: The Spearman Rank Correlation between explanation plausibility (MAP) and both measures of uncertainty across model, dataset and saliency map at high levels of perturbation ($\alpha \in \{0.90, 0.95\}$) All datapoints (correctly and incorrected guessed) are included. We bold the saliency map with the strongest correlation for each comparison.