THE EFFECT OF MODEL SIZE ON LLM POST-HOC EXPLAINABILITY VIA LIME

Henning Heyen*
University College London
henning.heyen.22@ucl.ac.uk

Amy Widdicombe
University College London
amy.widdicombe.17@ucl.ac.uk

Noah Y. Siegel
University College London
noah.siegel.22@ucl.ac.uk

María Pérez-Ortiz
University College London
maria.perez@ucl.ac.uk

Philip Treleaven
University College London
p.treleaven@ucl.ac.uk

ABSTRACT

Large language models (LLMs) are becoming bigger to boost performance. However, little is known about how explainability is affected by this trend. This work explores LIME explanations for DeBERTaV3 models of four different sizes on natural language inference (NLI) and zero-shot classification (ZSC) tasks. We evaluate the explanations based on their **faithfulness** to the models' internal decision processes and their **plausibility**, i.e. their agreement with human explanations. Our results suggest some extent of misalignment between the LIME explanations and the models' internal processes as model size increases.

1 INTRODUCTION

Research has shown that performance in language models depends strongly on scale and less on model shape (Kaplan et al., 2020), where scale refers to the number of parameters, the training dataset size, and the amount of compute for training. For instance, OpenAI's series of Generative Pre-Trained Transformers (GPT) has grown from 1.5 billion parameters for GPT-2 to 175 billion parameters for GPT-3 which helped improve across various NLP tasks Brown et al. (2020). This trend seems likely to continue.

As LLMs grow in size and performance and are increasingly deployed in high-stakes applications, the need to understand and explain their behaviour becomes more crucial. Post-hoc explainability methods such as LIME Ribeiro et al. (2016) are one way of attempting to do this. Although these methods have been widely applied to LLMs (Madsen et al., 2022), to the best of our knowledge no research has been conducted on the impact of model size on the quality of these kinds of explanations. Here we begin to fill this gap by investigating the impact of model size on the quality of LIME explanations. We apply two approaches to assess the quality of explanations, namely faithfulness (Chan et al., 2022) and plausibility (DeYoung et al., 2020). While faithfulness aims to measure the extent to which an explanation reflects the true internal decision processes, plausibility assesses the quality of the explanations based on their agreement with human-generated explanations.

We find that, even though model performance increases with model size, the agreement between human-generated and LIME-generated explanations does not. This indicates some extent of misalignment between the explanations and the true internal decision processes. Our findings also imply possible flaws in removal-based faithfulness metrics based on the NLP task which points to more general limitations for post-hoc explainabilty. This study serves as a first attempt to understand how post-hoc explainability is affected by model size. We hope that this research encourages others

*Primary author

to further explore this area and to that end we provide an extensible code repository[1] for others to build on.

2 Methodology

Models and Datasets We use fine-tuned DeBERTaV3 models from Huggingface of four different sizes, ranging from 22 to 304 million parameters[2]. Note that state-of-the-art models exhibit parameter counts in the order of billions. Due to computational constraints, we could not experiment with larger models. The models were fine-tuned on two standard natural language inference (NLI) datasets, matched MNLI (Williams et al., 2018) and SNLI (Bowman et al., 2015). Instead of SNLI, we use e-SNLI (Camburu et al., 2018) which extends SNLI by human annotated highlights indicating the most important tokens with respect to the label. Additionally, we apply the models in a zero-shot classification (ZSC) setting using the CoS-e (Rajani et al., 2019) dataset. CoS-e consists of commonsense questions with five candidate labels where the candidate labels differ for each question[3]. Similarly to e-SNLI, CoS-e contains human annotated highlights[4]

Explainability Method and Metrics There exist different notions of explainability in NLP. One is in the form of free-text natural language explanations and another is in the form of highlight-based, typically post-hoc explanations. While the former approach commonly requires human evaluation the latter can be measured more objectively because post-hoc techniques in NLP are normally mappings from tokens to real-valued importance scores. Common techniques can be categorised as gradient-based, attention weight-based and perturbation-based. While gradient-based techniques are highly vulnerable to adversarial perturbation (Wang et al., 2020) several studies argue that explanations based on attention weights are unreliable. The study "Attention is not Explanation" (Jain & Wallace, 2019), for instance, identified different attention distributions yielding equivalent predictions. We believe perturbation-based methods avoid some of these pitfalls - we use one such method, LIME Ribeiro et al. (2016), in our experiments. There is no single framework for evaluating post-hoc explanations due to a lack of consensus on what constitutes a high-quality explanation. In this work, we examine two different approaches, namely faithfulness and plausibility:

Faithfulness as discussed in Chan et al. (2022) aims to measure to what extent the explanation reflects the model's internal decision process. Generally, faithfulness metrics rely on removing tokens from the input sequence based on the explanation and measuring the change in prediction. While several faithfulness metrics exist, they are "not always consistent with each other and even lead to contradictory conclusions" (Chan et al., 2022) - they compared six faithfulness metrics and concluded that **comprehensiveness** is the most diagnostic and the least complex. Based on this we use comprehensiveness for our experiments. Proposed by DeYoung et al. (2020), comprehensiveness suggests that an explanation is faithful if the prediction strongly deviates when the most important tokens (as identified by the explanation method) are removed from the input sequence[5]. More formally,

$$\text{COMP}(\mathbf{x}, c, k) = p(c \mid \mathbf{x}; \theta) - p(c \mid \mathbf{x} \backslash \mathbf{x}_k; \theta), \tag{1}$$

where $p(c \mid \mathbf{x}; \theta)$ denotes the model's prediction for class c on the entire input sequence and $p(c \mid \mathbf{x} \backslash \mathbf{x}_k; \theta)$ denotes the model's prediction when the top k most important tokens $\mathbf{x}_k$ are removed from the input string. Practically, we obtain the most important tokens by taking the top t percent tokens from the list of tokens with associated real-valued importance scores generated by a post-hoc explainability method such as LIME. We denote this explanation as $\mathbf{x}_t$. To enhance the metric's reliability, the original paper proposes aggregated comprehensiveness which averages the comprehensiveness over different lengths of explanations. In this work, we use bins $t \in T = \{10\%, 30\%, 50\%\}$ to vary the length of explanations. The aggregated comprehensiveness can be defined as,

$$\text{COMP}_{\text{agg}}(\mathbf{x}, c) = \sum_{t \in T} \text{COMP}(\mathbf{x}, c, t) = \sum_{t \in T} \left(p(c \mid \mathbf{x}; \theta) - p(c \mid \mathbf{x} \setminus \mathbf{x}_{:t}; \theta) \right). \tag{2}$$

[1] https://github.com/henningheyen/ICLR-Submission

[2] For architectural specifics see Table 4 in the appendix

[3] The Huggingface API internally transforms zero-shot classification problems to an NLI problem

[4] For examples from MNLI, e-SNLI and CoS-e refer to Table 5, 6 and 7 in the appendix.

[5] For a visualization of the comprehensiveness metric see Figure 1 in the appendix

Plausibility, as compared to faithfulness, defines the quality of an explanation by the intersection between the highlights generated by a post-hoc explainability technique and some human-annotated highlights. In other words, plausibility measures the "agreement between extracted and human rationale" (DeYoung et al., 2020). Plausibility fundamentally differs from faithfulness in that plausible explanations do not reveal whether the model actually relied on the explanation. Assessing plausibility typically requires human evaluation (Strout et al., 2019). However, more recently some existing datasets have been extended by human-annotated highlights (DeYoung et al., 2020) which allows for more quantitative evaluation of plausibility. In this paper, we use two datasets with human-annotated highlights, nameley CoS-e (Rajani et al., 2019) and e-SNLI (Camburu et al., 2018). Similarly to DeYoung et al. (2020) we measure plausibility by the **intersection over union (IOU)**. As proposed by DeYoung et al. (2020), we take the number of most important tokens according to the average explanation length provided by humans for each dataset [6]. Suppose $\mathbf{x}_1$ is the set of tokens from the human explanation and $\mathbf{x}_2$ is the set of generated tokens. Then IOU can be formalised by:

$$\text{IOU}(\mathbf{x}_1, \mathbf{x}_2) = \frac{|\mathbf{x}_1 \cap \mathbf{x}_2|}{|\mathbf{x}_1 \cup \mathbf{x}_2|}. \quad (3)$$

3 Experiments and Results

The results for Experiment 1 and 2 are summarised in Table 1. Experiment 3 results and line plot visualisations can be found in the appendix.

Experiment 1 First, the four DeBERTaV3 models were evaluated in terms of performance on the validation sets of all three datasets (MNLI, e-SNLI, CoS-e). We report on model accuracy and 95% confidence intervals. We find that, as expected, performance improves monotonically with increasing model size for all three datasets. We can conclude that the models' capabilities are different enough to reason about the effect of model size on the LIME explanations.

Experiment 2 We then compute LIME explanations[7] for each model on a subset of 100 test samples from each dataset. We had to use a subset due to the computational intensity of LIME. The explanations were always calculated with respect to the predicted class, not necessarily the correct class. For each explanation, the aggregated comprehensiveness and IOU scores were computed according to equation 2 and 3 respectively. Note that IOU scores were only feasible for e-SNLI and CoS-e instances since MNLI does not provide human-annotated highlights. For each model, we report on the mean comprehensiveness and mean IOU scores across all 100 explanations. We observe an overall increase in comprehensiveness with the highest scores on the largest model for all three datasets suggesting that faithfulness of LIME increases with model size. IOU, on the other hand, stays almost constant across all model sizes for both datasets indicating that the plausibility of the LIME explanations is uncorrelated with model size.

Experiment 3 Lastly, we investigated both metrics with respect to the labels (*entailment, neutral, contradiction*) for MNLI and e-SNLI[8]. The goal is to see how the metrics change with model size when we condition on the label. We observe that comprehensiveness improves for contradictory sentence pairs with larger model sizes in MNLI, while no consistent pattern emerges in e-SNLI. Generally, we find that neutral sentence pairs achieved lower comprehensiveness scores than contradiction pairs. IOU stays almost constant across different model sizes regardless of the label.

Discussion Overall we find that for all three datasets, the largest model achieved the highest comprehensiveness score suggesting that with LIME larger models yield more faithful explanations. However, the IOU score stays constant across different model sizes suggesting that the plausibility of the explanations is uncorrelated with model size and performance. Interestingly this would imply that the agreement with human-annotated highlights does not improve with model performance which indicates an inherent misalignment between the generated explanations and the true internal decision process. This finding seems contradictory to our previous result that with LIME larger models yield more faithful explanations. Splitting the metrics by labels could reveal potential flaws with the comprehensiveness metric in the NLI setting as we found significantly lower scores

[6]Mean explanation-input-ratio e-SNLI: 0.19 ($\pm$0.193), CoS-e: 0.26 ($\pm$0.137)

[7]For a LIME example see Figure 5 in the appendix.

[8]Results are shown in Table 2, Figure 3 and Figure 4 in the appendix, the labels are balanced, see Table 8

Dataset	Model Size	Comprehensiveness	IOU	Accuracy	95% C.I.
MNLI	xsmall	0.785 ($\pm$ 0.022)	–	0.878	(0.871, 0.885)
	small	0.817 ($\pm$ 0.022)	–	0.878	(0.872, 0.884)
	base	0.796 ($\pm$ 0.027)	–	0.900	(0.894, 0.906)
	large	**0.823** ($\pm$ 0.027)	–	**0.902**	(0.896, 0.908)
e-SNLI	xsmall	0.726 ($\pm$ 0.022)	**0.282** ($\pm$ 0.017)	0.920	(0.915, 0.925)
	small	0.724 ($\pm$ 0.026)	0.259 ($\pm$ 0.016)	0.922	(0.917, 0.927)
	base	0.764 ($\pm$ 0.025)	0.254 ($\pm$ 0.016)	0.931	(0.926, 0.936)
	large	**0.778** ($\pm$ 0.025)	0.256 ($\pm$ 0.017)	**0.932**	(0.927, 0.937)
CoS-e	xsmall	0.304 ($\pm$ 0.018)	0.233 ($\pm$ 0.013)	0.331	(0.305, 0.355)
	small	0.316 ($\pm$ 0.019)	0.231 ($\pm$ 0.014)	0.336	(0.306, 0.362)
	base	0.356 ($\pm$ 0.020)	**0.235** ($\pm$ 0.012)	0.359	(0.330, 0.383)
	large	**0.391** ($\pm$ 0.022)	0.230 ($\pm$ 0.012)	**0.378**	(0.349, 0.406)

Table 1: Mean comprehensiveness and IOU scores with mean standard errors on 100 test samples for each dataset across all model sizes and accuracy scores on full validation sets with 95% confidence intervals. IOU could not be computed on MNLI as this dataset does not provide human annotated highlights as ground truth explanations.

for neutral sentence pairs. The problem with comprehensiveness in an NLI setting could be that removing highlighted tokens from a neutral pair might very well result in another neutral prediction which limits the applicability of comprehensiveness in this case. More generally, this shows how post-hoc explanations in NLP lack expressiveness. Highlights might not suffice to fully explain LLMs. This observation limits our finding that LIME explanations are more faithful for larger models. We conclude that the applicability of comprehensiveness is task-dependent and more coherent explainability metrics and techniques are needed.

4 Conclusion

Our work serves as a first attempt to understand how NLP explainability is affected by increasingly larger language models. We showed that token removal-based faithfulness metrics such as comprehensiveness are task-dependent and that highlight-based explainability techniques generally lack expressiveness as suggested by our results from Experiment 3. Our analysis of plausibility indicates that LIME explanations might not capture the true internal decision processes and that there exists an inherent misalignment. Besides post-hoc explanations similar alignment problems have previously been observed. For example, a study called "Language Models Don't Always Say What They Think" (Turpin et al., 2023) found that the prediction in chain of thought prompting can be manipulated although the explanations sound plausible which poses a risk of overtrusting large language models.

Future Work Future research could repeat our experiments using other perturbation-based post-hoc techniques such as Anchors (Ribeiro et al., 2018) or SHAP (Lundberg & Lee, 2017) to validate our observations. Additionally, other tasks such as sentiment analysis, text summarisation or language modelling could be explored. Furthermore, non of our models have gone through Reinforment Learning from Human Feedback (RLHF) (Ouyang et al., 2022). We believe that the results on plausibility might significantly change with RLHF. Note that our large models had 304 million parameters. The current state-of-the-art models, however, are much bigger, consisting of billions of parameters. While computational feasibility with perturbation-based techniques is one concern, the effect on the explainability of very large models is still unexplored and should be the subject of future investigation. More broadly, this work aims to emphasize the urgent need for an NLP explainability framework that is human-interpretable, objective, scalable and expressive.

Summary Our work investigated LIME explanations on fine-tuned DeBERTaV3 models of different sizes in an NLI and ZSC setting. We applied two approaches to capture the quality of explanations, namely faithfulness and plausibility. We measured faithfulness with comprehensiveness and plausibility with IOU. Our results showed improved comprehensiveness for LIME explanations with increasing model size. However, we identified limitations of the comprehensiveness metric in the

NLI setting and for post-hoc explanations in NLP more generally. Given that performance increased with model size raises questions on why agreement with human annotations does not increase. We suggest that there is some extent of misalignment between the model's internal decision process and its LIME explanation. This work aims to serve as an initial step towards understanding the effect of model size on LLM post-hoc explainability. We believe there is an urgent need for further investigations on LLM explainability and a more coherent explainability framework for LLMs. If we fail to faithfully explain the decisions of increasingly large language models we risk that those models pursue unexpected objectives rather than agreeing with human values and intentions.

Author Contributions

If you'd like to, you may include a section for author contributions as is done in many journals. This is optional and at the discretion of the authors.

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31, 2018.

Chun Sik Chan, Huanqi Kong, and Liang Guanqing. A comparative study of faithfulness metrics for model interpretability methods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5029–5038, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.345. URL https://aclanthology.org/2022.acl-long.345.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408.

Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Andreas Madsen, Siva Reddy, and Sarath Chandar. Post-hoc interpretability for neural nlp: A survey. *ACM Computing Surveys*, 55(8):1–42, 2022.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4932–4942, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1487. URL https://aclanthology.org/P19-1487.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Julia Strout, Ye Zhang, and Raymond Mooney. Do human rationales improve machine explanations? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 56–62, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4807. URL `https://aclanthology.org/W19-4807`.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*, 2023.

Junlin Wang, Jens Tuyls, Eric Wallace, and Sameer Singh. Gradient-based analysis of NLP models is manipulable. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 247–258, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.24. URL `https://aclanthology.org/2020.findings-emnlp.24`.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL `https://aclanthology.org/N18-1101`.

A APPENDIX

A.1 ADDITIONAL TABLES

Dataset	**Model Size**	**Label**	**Comprehensiveness**	**IOU**
MNLI	xsmall	contradiction	0.810 ($\pm$ 0.038)	-
		entailment	0.759 ($\pm$ 0.042)	-
		neutral	0.788 ($\pm$ 0.034)	-
	small	contradiction	0.853 ($\pm$ 0.041)	-
		entailment	0.805 ($\pm$ 0.039)	-
		neutral	0.794 ($\pm$ 0.035)	-
	base	contradiction	0.895 ($\pm$ 0.034)	-
		entailment	0.768 ($\pm$ 0.050)	-
		neutral	0.728 ($\pm$ 0.049)	-
	large	contradiction	0.939 ($\pm$ 0.018)	-
		entailment	0.750 ($\pm$ 0.057)	-
		neutral	0.789 ($\pm$ 0.046)	-
e-SNLI	xsmall	contradiction	0.805 ($\pm$ 0.034)	0.289 ($\pm$ 0.031)
		entailment	0.713 ($\pm$ 0.039)	0.315 ($\pm$ 0.025)
		neutral	0.663 ($\pm$ 0.035)	0.244 ($\pm$ 0.028)
	small	contradiction	0.744 ($\pm$ 0.042)	0.286 ($\pm$ 0.029)
		entailment	0.783 ($\pm$ 0.051)	0.249 ($\pm$ 0.025)
		neutral	0.652 ($\pm$ 0.040)	0.242 ($\pm$ 0.030)
	base	contradiction	0.808 ($\pm$ 0.038)	0.264 ($\pm$ 0.027)
		entailment	0.786 ($\pm$ 0.043)	0.291 ($\pm$ 0.025)
		neutral	0.701 ($\pm$ 0.045)	0.211 ($\pm$ 0.028)
	large	contradiction	0.759 ($\pm$ 0.046)	0.259 ($\pm$ 0.024)
		entailment	0.809 ($\pm$ 0.038)	0.292 ($\pm$ 0.026)
		neutral	0.768 ($\pm$ 0.042)	0.220 ($\pm$ 0.036)

Table 2: Mean comprehensiveness and IOU scores with mean standard errors on 100 test samples split by label for both NLI datasets across all model sizes. IOU could not be computed on MNLI as this dataset does not provide human-annotated highlights as ground truth explanations.

	MNLI	**e-SNLI**	**CoS-e**
xsmall	2min 3s	1min 8s	34min 35s
small	2min 40s	1min 40s	44min 28s
base	5min 20s	3min 35s	1h 27min 7s
large	15min 38s	12min 45s	4h 35min 50s

Table 3: Compute time for all LIME explanations of 100 test instances from each dataset across all model sizes on Nvidia's T4 GPU, 51GB RAM.

	Parameters (in millions)	**Layers**	**Hidden Size**	**Attention Heads**
large	304	24	1024	12
base	86	12	768	12
small	44	6	768	12
xsmall	22	12	384	6

Table 4: Architecture comparison for DeBERTaV3 models.

Premise	**Hypothesis**	**Label**
Look, there's a legend here.	See, there is a well-known hero here.	Entailment
Yeah, I know, and I did that all through college and it worked too.	I did that all through college but it never worked.	Contradiction
Boats in daily use lie within feet of the fashionable bars and restaurants.	Bars and restaurants are interesting places.	Neutral

Table 5: Natural language inference examples from the MNLI dataset.

Premise	**Hypothesis**	**Label**
An adult dressed in black holds a stick .	An adult is walking away, empty-handed .	Contradiction
A child in a yellow plastic safety swing is laughing as a dark-haired woman in pink and coral pants stands behind her.	A young mother is playing with her daughter in a swing.	Neutral
A man in an orange vest leans over a pickup truck .	A man is touching a truck.	Entailment

Table 6: Natural language inference examples from the e-SNLI dataset. Highlighted tokens indicate human-annotated explanations.

Question	**Candidate Labels**	**Label**
He was a sloppy eater , so where did he leave a mess?	sailboat, desk, closet, table, apartment	table
Where can someone get a new saw ?	hardware store, toolbox, logging camp, tool kit, auger	hardware store
Many homes in this country are built around a courtyard. Where is it?	hospital, park, spain, office complex, office	spain

Table 7: Zero shot classification examples from the CoS-e dataset. Highlighted tokens indicate human-annotated explanations.

Dataset	**Contradiction**	**Entailment**	**Neutral**
MNLI	32	36	32
e-SNLI	33	32	35

Table 8: Number of observations by label for MNLI and e-SNLI for 100 explained test samples.

A.2 Additional Figures

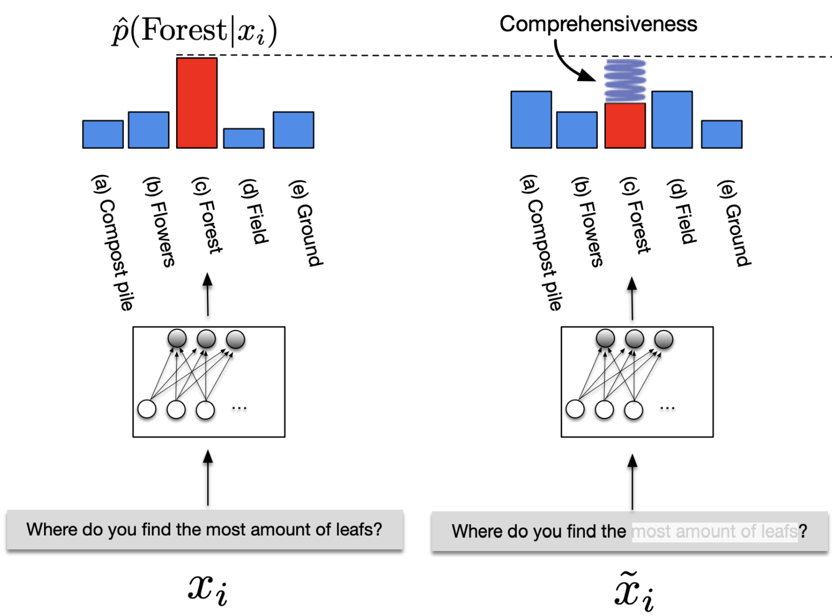

Figure 1: Visualistion of comprehensiveness metric on CoS-e instance from (DeYoung et al., 2020). Comprehensiveness suggests that an explanation is faithful if the prediction strongly deviates when the most important tokens (as identified by the explanation method) are removed from the input sequence.

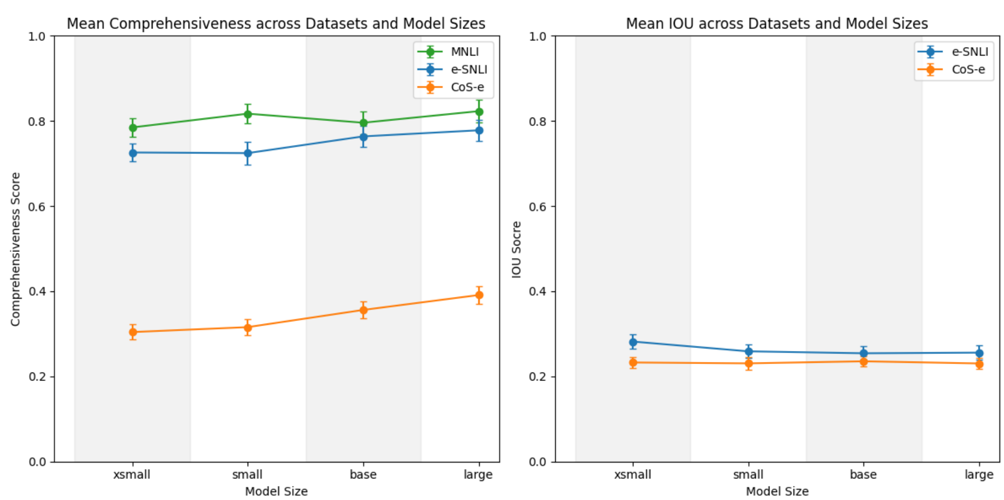

Figure 2: Mean comprehensiveness (left) and IOU (right) scores with mean standard errors on 100 test samples for each dataset across all model sizes. IOU could not be computed on MNLI as this dataset does not provide human-annotated highlights as ground truth explanations.

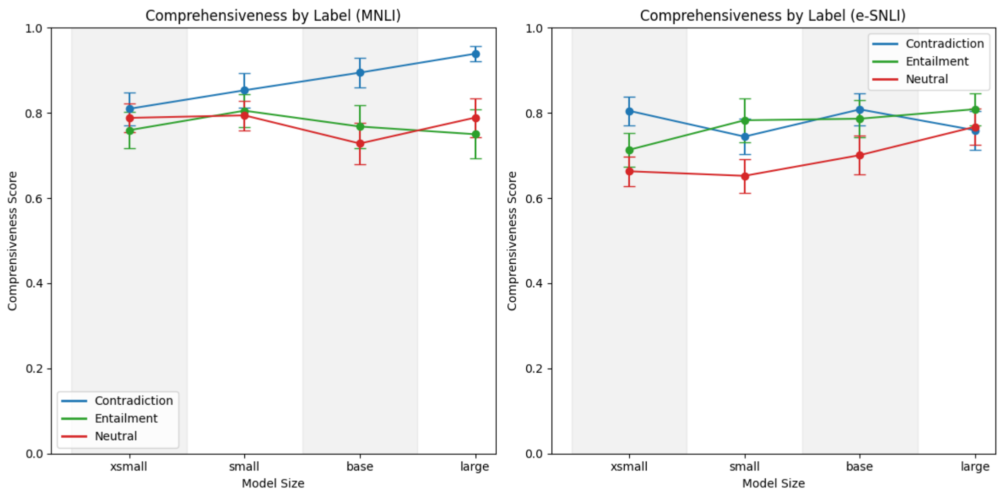

Figure 3: Mean comprehensiveness scores with mean standard errors on 100 test samples for MNLI (left) and e-SNLI (right) across all model sizes. Note how neutral sentence pairs achieve generally lower comprehensiveness scores than contradictive sentence pairs.

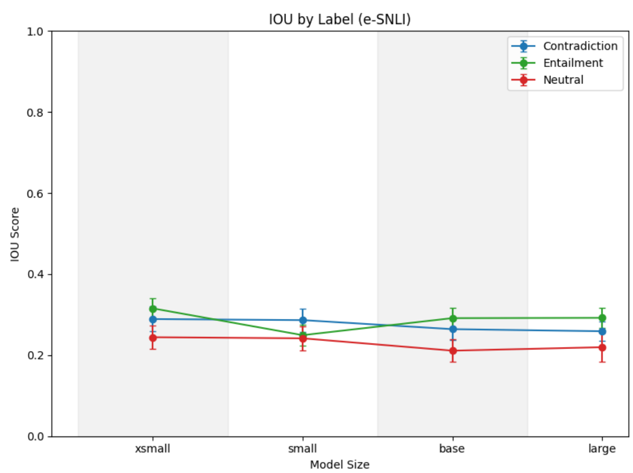

Figure 4: Mean comprehensiveness scores with mean standard errors on 100 test samples for CoS-e across all model sizes. Note how IOU scores are almost constant as the model size increases regardless of the label.

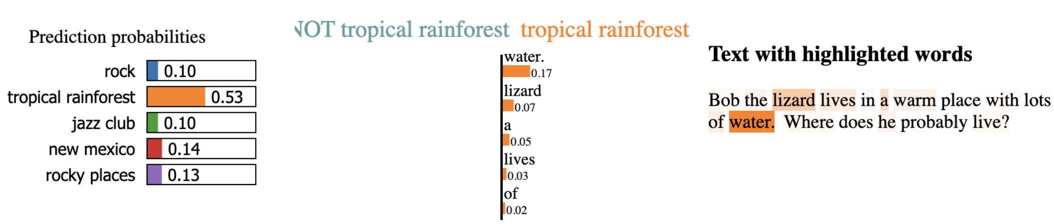

Figure 5: LIME example on a CoS-e instance using the xsmall DeBERTaV3 model. LIME maps every token to a real-valued importance score.