Faithfulness Measurable Masked Language Models

Andreas Madsen [1 2] **Siva Reddy** [1 3 4] **Sarath Chandar** [1 2 5]

Abstract

A common approach to explaining NLP models is to use importance measures that express which tokens are important for a prediction. Unfortunately, such explanations are often wrong despite being persuasive. Therefore, it is essential to measure their faithfulness. One such metric is if tokens are truly important, then masking them should result in worse model performance. However, token masking introduces out-of-distribution issues, and existing solutions that address this are computationally expensive and employ proxy models. Furthermore, other metrics are very limited in scope. This work proposes an inherently faithfulness measurable model that addresses these challenges. This is achieved using a novel fine-tuning method that incorporates masking, such that masking tokens become in-distribution by design. This differs from existing approaches, which are completely model-agnostic but are inapplicable in practice. We demonstrate the generality of our approach by applying it to 16 different datasets and validate it using statistical in-distribution tests. The faithfulness is then measured with 9 different importance measures. Because masking is in-distribution, importance measures that themselves use masking become consistently more faithful. Additionally, because the model makes faithfulness cheap to measure, we can optimize explanations towards maximal faithfulness; thus, our model becomes indirectly inherently explainable.

[1]Mila, Montreal, Canada [2]Computer Engineering and Software Engineering Department, Polytechnique Montreal, Montreal, Canada [3]Computer Science and Linguistics, McGill University, Montreal, Canada [4]Facebook CIFAR AI Chair [5]Canada CIFAR AI Chair. Correspondence to: Andreas Madsen <andreas.madsen@mila.quebec>, Siva Reddy <siva.reddy@mila.quebec>, Sarath Chandar <sarath.chandar@mila.quebec>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

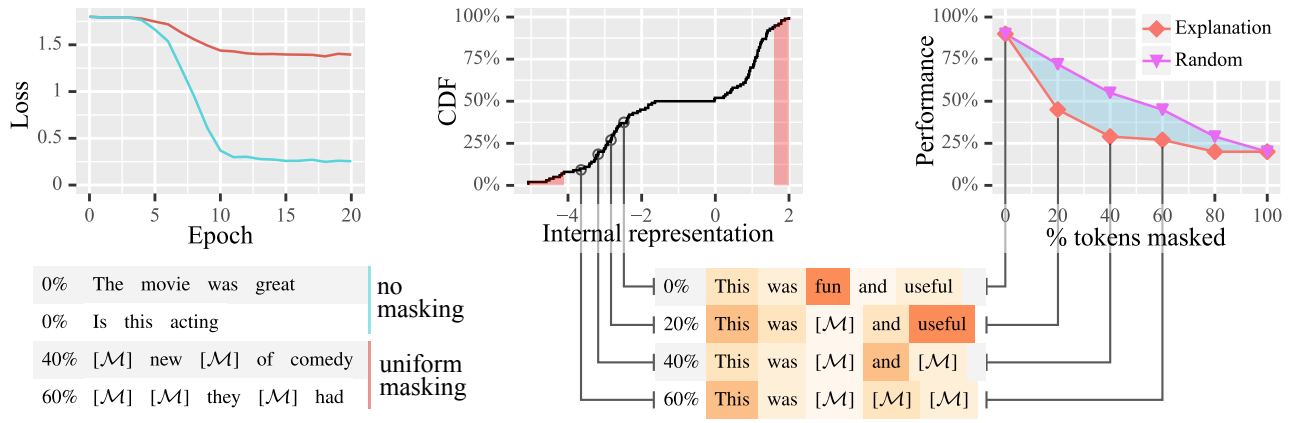
As machine learning models are increasingly being deployed, the demand for interpretability to ensure safe operation increases (Doshi-Velez & Kim, 2017). In NLP, importance measures such as attention or integrated gradient are a popular way of explaining which input tokens are important for making a prediction (Bhatt et al., 2019). These explanations are not only used directly to explain models but are also used in other explanations such as contrastive (Yin & Neubig, 2022), counterfactuals (Ross et al., 2021), and adversarial explanations (Ebrahimi et al., 2018).

Unfortunately, importance measures (IMs) are often found to provide false explanations despite being persuasive (Jain & Wallace, 2019; Hooker et al., 2019). For example, a given IMs might not be better at revealing important tokens than pointing at random tokens (Madsen et al., 2022a). This presents a risk, as false but persuasive explanations can lead to unsupported confidence in a model. Therefore, it's important to measure faithfulness. Jacovi & Goldberg (2020) defines faithfulness as: "how accurately it (explanation) reflects the true reasoning process of the model". In this work, we propose a methodology that enables existing models to support measuring faithfulness by design.

Measuring faithfulness is challenging, as there is generally no known ground-truth for the correct explanation. Instead, faithfulness metrics have to use proxies. One such proxy is the *erasure-metric* by Samek et al. (2017): if tokens are truly important, then masking them should result in worse model performance compared to masking random tokens.

However, masking tokens can create out-of-distribution issues. This can be solved by retraining the model after allegedly important tokens have been masked (Hooker et al., 2019; Madsen et al., 2022a). Unfortunately, this is computationally expensive, leaks the gold label, and the measured model is now different from the model of interest.

In general, the cost of proxies has been some combination of incorrect assumptions, expensive computations, or using a proxy-model (Jain & Wallace, 2019; Bastings et al., 2022; Madsen et al., 2022a). Based on previous work, we propose the following desirable, which to the best of our knowledge, no previous method satisfies in all aspects, but we satisfy:



(a) **Masked fine-tuning.** In-distribution support for masking any permutation of tokens is achieved by uniformly masking half of the mini-batch during fine-tuning. The other half is left unmasked to maintain regular unmasked performance.

(b) **In-distribution validation.** CDFs of the model's embeddings given a masked validation dataset, provide in-distribution p-values and validate that test observations masked according to an explanation are in-distribution.

(c) **Faithfulness metric.** Observations are masked according to an explanation. A model performance lower than masking random tokens means the explanation is faithful. A larger area between the curves means more faithful.

Figure 1. To measure faithfulness, a *faithfulness measurable masked language model* is created (a), then the model is checked for out-of-distribution issues given an explanation (b), and finally, the faithfulness is measured by masking allegedly important tokens (c). – *[M] is the masking token.*

a) The method does not assume a known true explanation.

b) The method measures faithfulness of an explanation w.r.t. a specific model instance and single observation. For example, it is not a proxy-model that is measured.

c) The method uses only the original dataset, e.g. does not introduce spurious correlations.

d) The method only uses inputs that are in-distribution w.r.t. the model.

e) The method is computationally cheap by not training/fine-tuning repeatedly and only computes explanations of the test dataset.

f) The method can be applied to any classification task.

The key idea is to use the erasure-metric but without solving the out-of-distribution (OOD) issue using retraining, which avoids all the limitations. This is achieved by including masking in the fine-tuning procedure of masked language models, such that masking is always in-distribution (Figure 1a). This is possible because language models are heavily over-parameterized and can thus support such additional complexity. Although our approach applies to Masked Language Models (MLMs), we suspect future work could apply this idea to any language model with sufficient capacity.

Our approach is significantly different from previous literature, which is completely model agnostic. Instead, we fine-tune a model such that measuring faithfulness is easy by design. We call such designed models: **inherently faithfulness measurable models** (FFMs).

To validate that masking is in-distribution, we generalize previous OOD detection work from computer vision (Matan et al., 2022). This serves as a statistically grounded meta-validation of the faithfulness measure itself (Figure 1b). Finally, once the model is validated, the erasure-metric can be applied (Figure 1c).

Note, the concept of an *inherently faithfulness measurable model* (FMM) is significantly different from inherently explainable models, which are interpretable by design (Jacovi & Goldberg, 2020). An FMM does not guarantee that an explanation exists, and an inherently explainable model doesn't provide a means of measuring faithfulness.

However, with an FMM, measuring faithfulness is computationally cheap. Therefore, optimizing an explanation w.r.t. faithfulness is possible, as proposed by Zhou & Shah (2023). However, they did not solve the OOD issue caused by masking, as it was "orthogonal" to their idea, but our *inherently faithfulness measurable model* fills that gap, making it indirectly inherently explainable.

Finally, for completeness, we compare a large variety of existing explanation methods and modify some existing explanations to be able to separate positive from negative contributions. In general, we find that the explanations that take advantage of masking (occlusion-based) are more faithful than gradient-based methods. However, the robustness provided by a *faithfulness measurable model* also makes some gradient-based methods more faithful.

To summarize, our contributions are:

- Introducing the concept of an *inherently faithfulness measurable model* (FMM).
- Proposing *masked fine-tuning* that enables masking to be in-distribution.
- Establishing a statistically grounded meta-validation for the faithfulness measurable model, using out-of-distribution detection.
- Making existing occlusion-based explanations more faithful, as they no longer cause out-of-distribution issues.
- Introducing signed variants of existing importance measures, which can separate between positive and negative contributing tokens.

2. Related Work

Much recent work in NLP has been devoted to investigating the faithfulness of importance measures. In this section, we categorize these faithfulness metrics according to their underlying principle and discuss their limitations. The limitations are annotated as (**a**) to (**f**) and refer to the desirables mentioned in the Introduction.

Note, we do not cover the faithfulness metrics that are specific to attention (Moradi et al., 2021; Wiegreffe & Pinter, 2019; Vashishth et al., 2019), as this paper presents a faithfulness metric for importance measures in general.

2.1. Correlating importance measures

One early idea was to compare two importance measures. The claim is that a correlation would be a very unlikely coincidence unless both explanations are faithful (Jain & Wallace, 2019). Both Jain & Wallace (2019) and Meister et al. (2021) find little correlation between attention, gradient, and leave-one-out; the explanations are therefore not faithful. Jain & Wallace (2019) do acknowledge the limitations of their approach, as it assumes each importance measure is faithful to begin with (**a**).

2.2. Known explanations in synthetic tasks

Arras et al. (2022) construct a purely synthetic task, where the true explanation is known, therefore the correlation can be applied appropriately. Unfortunately, this approach cannot be used on real datasets (**f**). Instead, Bastings et al. (2022) introduce spurious correlations into real datasets, creating partially synthetic tasks. They then evaluate if importance measures can detect these correlations. It is assumed that if an explanation fails this test, it is generally unfaithful. Bastings et al. (2022) conclude that faithfulness is both model and task-dependent.

Both methods are valid when measuring faithfulness on models trained on (partially) synthetic data. However, the model and task-dependent conclusion also means that we can't generalize the faithfulness findings to the models (**b**) and datasets of interest (**c**), thus limiting the applicability of this approach.

2.3. Similar inputs, similar explanation

Jacovi & Goldberg (2020) suggest that if similar inputs show similar explanations, then the explanation method is faithful. Zaman & Belinkov (2022) apply this idea using a multilingual dataset (**f**), where each language example is explained and aligned to the English example using a known alignment mapping. If the correlation between language pairs is high, this indicates faithfulness.

Besides being limited to multilingual datasets, the metric assumes the model behaves similarly among languages. However, languages may have different linguistic properties or spurious correlations. A faithful explanation would then yield different explanations for each language.

2.4. Removing important information should affect the prediction

Samek et al. (2017) introduce the erasure-metric: if information (input tokens) is truly important, then removing it should result in worse model performance compared to removing random information. However, Hooker et al. (2019) argue that removing tokens introduces an out-of-distribution (OOD) issue (**d**).

Hooker et al. (2019) solve the ODD issue by retraining the model. They point out a limitation: the method cannot separate an unfaithful explanation from "there exist dataset redundancies". However, this was later solved by Madsen et al. (2022a), who conclude, like Bastings et al. (2022), that faithfulness is both model and task-dependent.

Unfortunately, both methods are computationally expensive (**e**), due to the model being retrained. Additionally, retraining means that it is no longer possible to comment on the faithfulness of the deployed model (**b**). Finally, removing tokens and retraining can introduce redundancies which underestimates the faithfulness (Madsen et al., 2022a).

3. Inherently faithfulness measurable models (FMMs)

As an alternative to existing faithfulness methods, which all aim to work with any models, we propose creating *inherently faithfulness measurable models* (FMMs). These models provide the typical output (e.g., classification) for a given task and, by design, provide the means to measure the faithfulness of an explanation. Importantly, this allows measuring the faithfulness of a specific model, as there is

no need for proxy models, an important property in a real deployment setting.

An FMM does have the limitation that a specific model is required. However, our proposed method is very general, as it only requires a modified fine-tuning procedure applied to a masked language model.

3.1. Faithfulness of importance measures

In this paper, we look at importance measures (IMs), which are explanations that either score or rank how important each input token is for making a prediction. A faithfulness metric measures how much such an explanation reflects the true reasoning process of the model (Jacovi & Goldberg, 2020). Importantly, such a metric should work regardless of how the importance measure is calculated.

For importance measures, there are multiple definitions of truth. In this paper, we use the erasure-metric definition: *if information (tokens) is truly important, then masking them should result in worse model performance compared to masking random information (tokens)* (Samek et al., 2017; Hooker et al., 2019; DeYoung et al., 2020).

The challenge with an erasure-metric is that fine-tuned models do not support masking tokens. Even masked language models are usually only trained with 12% or 15% masking (Devlin et al., 2019; Liu et al., 2019; Wettig et al., 2023), and an erasure-metric use between 0% and 100% masking. Furthermore, catastrophic forgetting of the masking token is likely when fine-tuning.

Hooker et al. (2019) and Madsen et al. (2022a) solve this by retraining the model with partially masked inputs and call the approach ROAR (Remove and Retrain). Unfortunately, as discussed in Section 2, retraining has issues. It is computationally expensive, leaks the gold label, and measures a proxy model instead of the true model.

We find that the core issue is the need for retraining. Instead, if the fine-tuned model supports masking any permutation of tokens, then retraining would not be required, eliminating all issues. We propose a new fine-tuning procedure called *masked fine-tuning* to achieve this.

To evaluate faithfulness of an importance measure, we propose a three-step process, as visualized in Figure 1 (also see Appendix F for details):

1. Create a faithfulness measurable masked language model, using *masked fine-tuning*. – See Section 3.2 and Figure 1a.
2. Check for out-of-distribution (OOD) issues, by using a statistical in-distribution test. – See Section 3.3 and Figure 1b.
3. Measure the faithfulness of an explanation. – See Section 3.4 and Figure 1c.

3.2. Masked fine-tuning

To provide masking support in the fine-tuned model, we propose randomly masking the training dataset by uniformly sampling a masking rate between 0% and 100% for each observation and then randomly masking that ratio of tokens. However, half of the mini-batch remains unmasked to maintain the regular unmasked performance. This is analogous to multi-task learning, where one task is masking support, and the other is regular performance.

To include masking support in early stopping, the validation dataset is duplicated, where one copy is unmasked, and one copy is randomly masked.

3.3. In-distribution validation

Erasure-based metrics are only valid when the input is in-distribution, and previous works did not validate for this (Hooker et al., 2019; Madsen et al., 2022a). Additionally, in-distribution is the statistical null-hypothesis and can never be proven. However, we can validate this using an out-of-distribution (OOD) test, which we do in this paper.

We use the *MaSF* method by Matan et al. (2022) as the OOD test, which provides non-parametric p-values under the in-distribution null-hypothesis. While *MaSF* is developed and tested for computer vision, it is very general and naturally applies to NLP. The method works by developing empirical Cumulative Distribution Functions (CDFs) of the model's intermediate embeddings. In the case of RoBERTa we use the embeddings after the layer-normalization (Ba et al., 2016).

The validation dataset is used to develop the empirical CDFs. Because the validation dataset should be from the same distribution as the training dataset, the masked fine-tuning transformation is also used on the validation dataset. Once the CDFs are developed, a test observation can be tested against the CDFs which provides in-distribution p-values for each embedding, these are then aggregated using Fisher (Fisher, 1992) and Simes's (Simes, 1986) method. Finally, to a p-value for the entire masked test dataset being in-distribution we perform another Simes (Simes, 1986) aggregation. The details of the entire workflow and procedure are described in Appendix F.2.

3.4. Faithfulness metric

To measure faithfulness on a model trained using *masked fine-tuning*: the importance measure (IM) is computed for a given input, then $x\%$ (e.g., 10%) of the most important tokens are masked, then the IM is calculated on this masked input, finally an additional $x\%$ of the most important tokens are masked. This is repeated until 100% of the input is masked. The importance measure is re-calculated be-

cause otherwise, dataset redundancies will interfere with the metric, as shown by Madsen et al. (2022a).

At each iteration, the masked input is validated using MaSF and the performance is measured. Faithfulness is shown if and only if the performance is less than when masking random tokens. This procedure is identical to Recursive-ROAR by Madsen et al. (2022a), but without retraining and with in-distribution validation.

4. Importance measures (IMs)

Our proposed *inherently faithfulness measurable model* and the erasure-metric can be applied to a single observation, which is useful in practical settings but not for statistical conclusions. Therefore, we focus on importance measures, which are feasible to compute on the entire test dataset.

The importance measures used in this paper are all from existing literature; hence, the details are in the Appendix E, except we introduce signed and absolute variants of existing IMs. Additionally, we argue that occlusion-based IMs are only valid in combination with *masked fine-tuning*.

4.1. Gradient-based vs occlusion-based

A common idea is that if a small change in the input causes a large change in the output, then that indicates importance.

Gradient-based The relationship of change can be modeled by the gradient w.r.t. the input. Let $f(\mathbf{x})$ be the model with input[1] $\mathbf{x} \in \mathbb{R}^{T \times V}$, then the gradient explanation (**grad**) is $\nabla_{\mathbf{x}} f(\mathbf{x})_y \in \mathbb{R}^{T \times V}$ (Baehrens et al., 2010; Li et al., 2016a). Later work (Kindermans et al., 2016) have proposed using $\mathbf{x} \odot \nabla_{\mathbf{x}} f(\mathbf{x})_y$ instead (i.e. $\mathbf{x} \odot$ **grad**), or to sample multiple gradients (**IG**) (Sundararajan et al., 2017).

Occlusion-based An alternative to using gradients is to measure the change when removing or masking a token and measuring how it affects the output. One rarely addressed concern, is that removing or masking tokens is likely to cause out-of-distribution issues (Zhou & Shah, 2023). This would cause an otherwise sound method to become unfaithful. However, because our proposal *faithfulness measurable masked language model* supports masking, this should not be a concern. Importantly, this exemplifies how a *inherently faithfulness measurable model* may also produce more faithful explanations.

A simple occlusion-based explanation is leave-one-out (**LOO**) which tests the effect of each token, one-by-one (Li et al., 2016b). However, this does not account for when one token is masked another token may become more important. Instead, Zhou & Shah (2023) propose to optimize for the faithfulness metric itself, rather than using a heuristic. The idea is to use beam-search (**Beam**), where the generated sequence is the optimal masking order of tokens.

[1] V is vocabulary-size and T is sequence-length.

4.2. Signed and absolute variants

Most literature does not distinguish between positive and negative contributing tokens[2], which we categorize as an *absolute IM*. In the opposite case, when positive and negative are separated, we categorize them as a *signed IM*. In most cases, a signed IM can be transformed into an absolute IM using $abs(\cdot)$.

5. Experiments

We use RoBERTa in size `base` and `large`, with the default GLUE hyperparameters provided by Liu et al. (2019). We present results on 16 classification datasets in the appendix but only include BoolQ and MRPC in the main paper. These were chosen as they represent the general trends we observe, although we observe very consistent results across all datasets. The full dataset list is in Appendix C and model details are in Appendix D. The code is available at `https://github.com/AndreasMadsen/faithfulness-measurable-models`.

For each experiment, we use 5 seeds and present their means with their 95% confidence interval (error-bars or ribbons). The 95% confidence interval is computed using the bias-corrected and accelerated bootstrap method (Buckland et al., 1998; Michael R. Chernick & LaBudde, 2011). When relevant, each seed is presented as a plus (+).

5.1. Masked fine-tuning

There are two criteria for learning our proposed *faithfulness measurable masked language model*:

1. The usual performance metric, where no data is masked, should not decrease.
2. Masking any permutations of tokens should be in-distribution.

In Section 3.2, we propose *masked fine-tuning*, where one half of a mini-batch is uniformly masked between 0% and 100% and the other half is unmasked. Additionally, the validation dataset contains a masked copy and an unmasked copy.

Unmasked performance. To validate the first goal, Figure 2 presents an ablation study. It compares *masked fine-tuning* with using only unmasked data (*plain fine-tuning*), as is traditionally done, and using only uniformly masked

[2] Unfortunately, the details of this are often omitted in the literature and can often only be learned through reading the code. However, Meister et al. (2021) do provide the details.

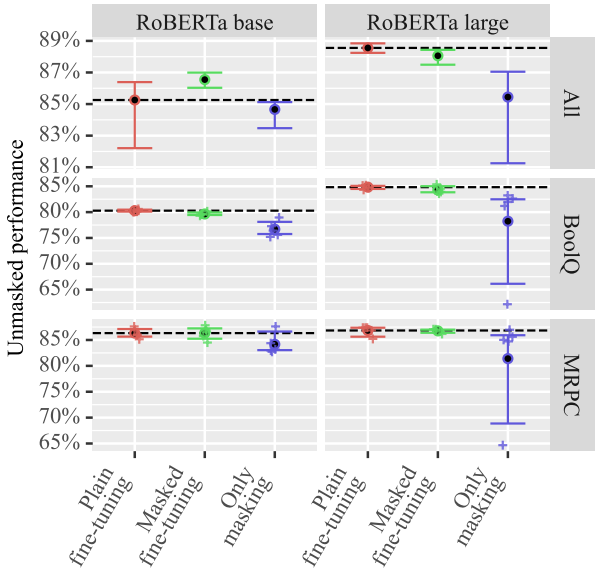


Figure 2. The unmasked performance for each fine-tuning strategy. *Plain fine-tuning* is the baseline (dashed line). We find that our *Masked fine-tuning* does not decrease performance. *All* is computed by taking the average of all datasets. More datasets and a more detailed ablation study can be found in Appendix H.

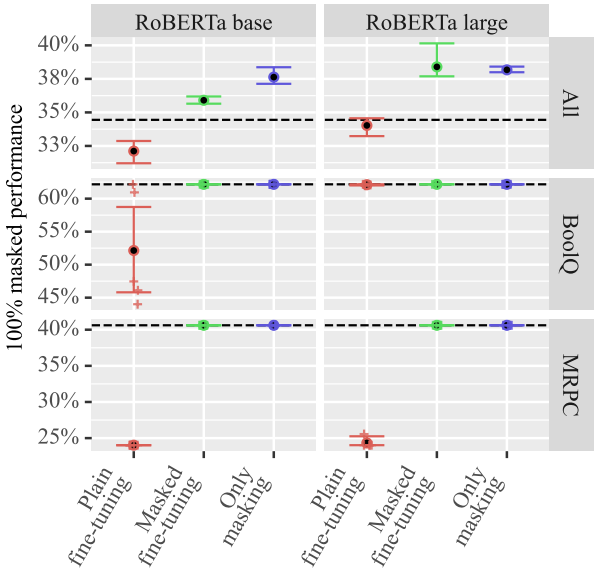


Figure 3. The 100% masked performance for each fine-tuning strategy. The dashed line represents the class-majority baseline. Results show that masking during training (either our *masked fine-tuning* or *only masking*) is necessary. More datasets and a more detailed ablation study can be found in Appendix H.

data (*only masking*). The unmasked performance is then measured (the usual benchmark).

We observe that no performance is loss when using our *masked fine-tuning*, some tasks even perform better likely because masking have a regualizing effect. However, when using *only masking* performance is lost unstable convergence is frequent. For bAbI-2&3, we also observe unstable convergence using *masked fine-tuning*. However, this is less frequent (worst case: 3/5) and only for RoBERTA-large (see Appendix H). Note the default RoBERTa hyperparameters are not meant for synthetic datasets like bAbI. Therefore, optimizing the hyperparameter would likely solve the stability issues with *masked fine-tuning*. Finally, when using *masked fine-tuning*, the models do need to be trained for slightly more epochs (twice more or less); see Appendix I. Again, tuning hyperparameter would likely help.

100% Masked performance. Measuring in-distribution support for masked data is challenging, as there is generally no known performance baseline. However, for 100% masked data, only the sequence length is left as information. Therefore, the performance of a model should be at least that of the class-majority baseline, where the most frequent class is "predicted" for all observations. We present an ablation study using this baseline in Figure 3. In Section 5.2, we perform a more in-depth validation.

From Figure 3, we observe that training with unmasked data (*Plain fine-tuning*) performs worse than the class-majority baseline, clearly showing an out-of-distribution issue. However, when using masked data, either *only masking* or *masked fine-tuning*, both effectively achieve in-distribution results for 100% masked data.

The best approach used in the following experiments. Appendix H contains a more detailed ablation study separation of the training and validation strategy. However, the conclusion is the same. *Masked fine-tuning* is the only method that achieves good results for both the unmasked and 100% masked cases.

For the following experiments in Section 5.2 and Section 5.3, the *masked fine-tuning* method is used. Additionally, we will only present results for RoBERTa-base for brevity. RoBERTa-large results are included in the appendix.

5.2. In-distribution validation

Because the expected performance for masked data is generally unknown, a statistical in-distribution test called *MaSF* (Matan et al., 2022) is used instead, as was briefly explained in Section 3.3, with details in Appendix F.2.

MaSF provides an in-distribution p-value for each observation. To test if all masked test observations are in-distribution, the p-values are aggregated using Simes's method (Simes, 1986). Because in-distribution is the null-hypothesis, we can never confirm in-distribution; we can only validate it. Rejecting the null hypothesis would mean

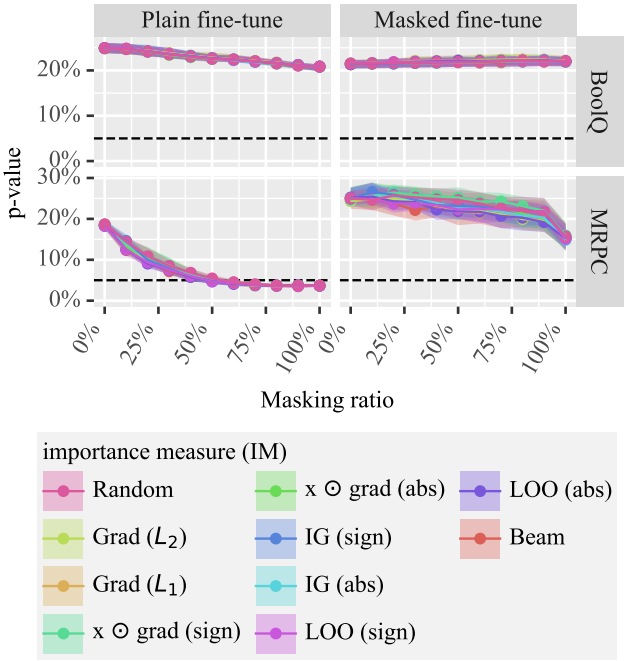


Figure 4. In-distribution p-values using MaSF, for RoBERTa-base with and without masked fine-tuning. The masked tokens are chosen according to an importance measure. P-values below the dashed line show out-of-distribution (OOD) results, given a 5% risk of a false positive. Results show that only when using *masked fine-tuning* is masking consistently not OOD. Because the results are highly consistent, the overlapping lines do not hide any important details. More datasets and models in Appendix J.

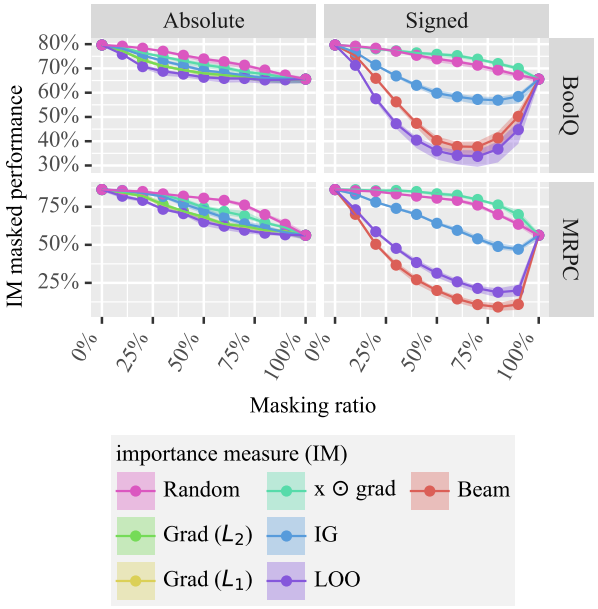


Figure 5. The performance given the masked datasets, where masking is done for the x% allegedly most important tokens according to the importance measure. If the performance for a given explanation is below the *"Random"* baseline, this shows faithfulness. Although faithfulness is not an absolute concept, so more is better. This plot is for RoBERTa-base and separates importance measures based on their signed and absolute variants. More datasets and models in Appendix K.

that some observation is out-of-distribution.

Because random uniform masking is not the same as strategically masking tokens, we validate in-distribution for each importance measure, where the masking is done according to the importance measure, identically to how the faithfulness metric is computed (Section 3.4).

Additionally, because MaSF does not consider the model's performance, it is necessary to consider these results in combination with regular performance metrics, see Section 5.1.

The results for when using *masked fine-tuning* and *plain fine-tuning* (no masking) are presented in Figure 4. The results show that masked datasets are consistently in-distribution only when using masked fine-tuning.

In the case of BoolQ, we suspect that because the training dataset is fairly small (7542 observations), the model does not completely forget the mask token. Additionally, a few datasets, such as bAbI-2, become out-of-distribution at 100% masking when using masked fine-tuning (see Appendix J). This contradicts the performance results for 100% masked data (Appendix H), which clearly show in-distribution performance. This is likely a limitation of MaSF because the empirical CDF in MaSF has very little 100% masked data, as the masking ratio is uniform between 0% and 100%. Fortunately, this is not a concern because Figure 3 shows in-distribution results for 100% masked data.

5.3. Faithfulness metric

Based on previous experiments, we can conclude that *masked fine-tuning* achieves both objectives: unaffected regular performance and support for masked inputs. Therefore, it is safe to apply the faithfulness metric to these models.

Briefly, the faithfulness metric works by showing that masking using an importance measure (IM) is more effective at removing important tokens than using a known false explanation, such as a random explanation. Therefore, if a curve is below the random baseline, the IM is faithful. Although faithfulness is not an absolute (Jacovi & Goldberg, 2020), so further below would indicate more faithful.

Figure 5 shows that most explanations are faithful, although some are much more faithful than others. In particular, occlusion-based importance measures (LOO, Beam) are the most faithful. This is expected, as they take advantage of the masking support that our *faithfulness measurable masked language model* offers.

Table 1. Faithfulness scores using Relative Area Between Curves (RACU) and the non-relative variant (ACU). The less relevant score is grayed out. Higher is better. Negative values indicate not-faithful. The comparison with Recursive-ROAR (Madsen et al., 2022a) is imperfect because Recursive-ROAR has limitations. See Table 11 for all datasets and Table 12 for RoBERTa-large. See Appendix A for additional result discussion.

Dataset	IM	Faithfulness [%]		
		Our		R-ROAR
		ACU	RACU	RACU
SST2	Grad (L_2)	$12.2^{+0.6}_{-0.7}$	$40.4^{+3.0}_{-1.7}$	$26.1^{+1.6}_{-2.2}$
	Grad (L_1)	$12.1^{+0.7}_{-0.7}$	$40.3^{+3.3}_{-1.8}$	–
	x $\odot$ grad (sign)	$-3.7^{+1.5}_{-1.6}$	$-12.2^{+4.5}_{-6.0}$	–
	x $\odot$ grad (abs)	$7.1^{+0.2}_{-0.2}$	$23.5^{+1.9}_{-1.1}$	$18.6^{+4.1}_{-4.6}$
	IG (sign)	$31.8^{+2.8}_{-2.2}$	$105.6^{+7.7}_{-7.7}$	–
	IG (abs)	$13.7^{+0.8}_{-0.8}$	$45.3^{+4.1}_{-2.8}$	$32.9^{+1.8}_{-1.5}$
	LOO (sign)	$51.6^{+1.4}_{-0.9}$	$171.3^{+5.8}_{-6.2}$	–
	LOO (abs)	$16.6^{+1.2}_{-1.0}$	$54.9^{+2.1}_{-1.5}$	–
	Beam	$56.4^{+0.5}_{-0.7}$	$187.3^{+8.1}_{-7.1}$	–
bAbI-2	Grad (L_2)	$28.5^{+0.8}_{-0.8}$	$96.3^{+6.8}_{-2.8}$	$57.8^{+2.0}_{-2.0}$
	Grad (L_1)	$28.5^{+0.9}_{-0.8}$	$96.3^{+6.8}_{-2.7}$	–
	x $\odot$ grad (sign)	$19.7^{+6.6}_{-8.1}$	$65.7^{+24.1}_{-26.3}$	–
	x $\odot$ grad (abs)	$27.3^{+1.7}_{-1.5}$	$92.0^{+2.5}_{-3.1}$	$48.1^{+3.2}_{-3.5}$
	IG (sign)	$40.3^{+0.9}_{-0.8}$	$136.3^{+4.4}_{-6.4}$	–
	IG (abs)	$29.1^{+1.0}_{-1.3}$	$98.3^{+5.5}_{-3.9}$	$42.0^{+3.8}_{-4.8}$
	LOO (sign)	$40.2^{+1.2}_{-0.8}$	$136.0^{+4.1}_{-6.5}$	–
	LOO (abs)	$28.5^{+0.9}_{-1.4}$	$96.3^{+9.2}_{-3.6}$	–
	Beam	$41.1^{+1.0}_{-0.7}$	$139.2^{+5.0}_{-7.3}$	–

Because signed importance measures can differentiate between positive and negative contributing tokens, while absolute tokens are not, it is to be expected that signed importance measures are more faithful. However, comparing them might not be fair because of this difference in capability. We let the reader decide this for themselves.

Relative Area Between Curves (RACU) Madsen et al. (2022a) propose to compute the area between the random curve and an explanation curve (RACU). This is then normalized by the theoretical optimal explanation, which would achieve the performance of 100% masking immediately. However, the normalization is only theoretically optimal for an absolute importance measure (IM). Signed IMs can trick the model into predicting the opposite label, thus achieving even lower performance. For this reason, we also show the un-normalized metric (ACU) in Table 1.

Note that comparing with Recursive ROAR (R-ROAR) (Madsen et al., 2022a) is troublesome because R-ROAR has issues, such as leaking the gold label. Additionally, while they also use RoBERTa-base it's not the same model because we use masked fine-tuning.

That said, our Faithfulness Measurable Masked Language Model drastically outperforms the R-ROAR approach on faithfulness. In Appendix A, we provide additional discussion on some less important observations.

6. Limitations

In this section, we discuss the most important limitations. Additionally, in the interest of completeness, Appendix B provides additional limitations.

No faithfulness ablation with regular fine-tuning We claim *masked fine-tuning* makes importance measures (IMs) more faithful. However, there is no ablation study where we measure faithfulness without *masked fine-tuning*. This is because, without *masked fine-tuning*, masking is out-of-distribution which makes the faithfulness measure invalid.

However, our argument for occlusion-based IMs has a theoretical foundation, as occlusion (i.e., masking) is only in-distribution because of *masked fine-tuning*. We also observe that occlusion-based IMs are consistently more faithful than gradient-based IMs. Finally, for gradient-based IMs, we compare with Recursive ROAR (Madsen et al., 2022a), and our approach provides more faithful explanations, although this comparison is imperfect as discussed in Section 5.3.

Uses masked language models (MLMs) Masked fine-tuning leverages pre-trained MLMs' partial support for token masking. Therefore, our approach does not immediately generalize to casual language models (CLM). However, despite CLMs' popularity for generative tasks, MLMs are still very relevant for classification tasks (Min et al., 2024) and for non-NLP tasks, such as analyzing biological sequences (genomes, proteins, etc.) (Zhang et al., 2023).

Additionally, it is possible to introduce the mask tokens to CLMs by masking random tokens in the input sequence while keeping the generation objective the same, similar to how unknown-word tokens are used. This approach could also be done in an additional pre-training step using existing pre-trained models. Regardless, masking support for CLMs is likely a more complex task and is left for future work.

Another direction useful for classification tasks, is to transform CLMs into MLMs, which has been shown to be quite straightforward (Muennighoff et al., 2024). It may also be possible to simply prompt an instruction-tuned CLM, such that it understands what masking means, for example Madsen et al. (2024) prompts with "The following content may contain redacted information marked with [REDACTED]".

In terms of supporting sequential outputs rather than just classification outputs, our methodology only requires a performance metric. Using sequential performance metrics such as ROUGE (Lin, 2004) or BLEU should therefore work perfectly well.

7. Conclusion

Using only a simple modified fine-tuning method, called *masked fine-tuning*, we are able to turn a typical general-purpose masked language model (RoBERTa) into an *inherently faithfulness measurable model* (FMM). Meaning that the model, by design, inherently provides a way to measure the faithfulness of importance measure (IM) explanations.

To the best of our knowledge, this is the first work that proposes creating a model designed to be faithfulness measurable. Importantly, our approach is very general, simple to apply, and stratifies critical desirables that previous measures didn't. The *masked fine-tuning* method does not decrease performance on all 16 tested datasets while also adding in-distribution support for token masking. This is verified using statistical OOD tests.

We find that occlusion-based IMs are consistently the most faithful. This is to be expected, as they take advantage of the masking support. Additionally, Beam uses beam-search to optimize towards faithfulness (Zhou & Shah, 2023), which our proposed faithfulness measurable masked language model makes computationally efficient to evaluate.

It is worth considering the significance of this. While our proposed model is not an *inherently explainable model* (Jacovi & Goldberg, 2020), it is *indirectly* inherently explainable because it provides a built-in way to measure faithfulness, which can then be optimized for. It does this without sacrificing the generality of the model, as it is still a RoBERTa model. As such, FMMs provide a new direction for interpretability, which bridges the gap between *post-hoc* (Madsen et al., 2022b) and *inherent* interpretability (Rudin, 2019). It does so by prioritizing faithfulness measures first and then the explanation, while previous directions have worked on explanation first and then measure faithfulness.

However, beam-search is just an approximative optimizer, and Leave-one-out does occasionally outperform Beam. Future work could look at better optimization methods to improve the faithfulness of such explanations.

Impact Statement

Interpretability is an essential component when considering the ethical deployment of a model. In this paper, we particularly consider deployment a key motivator in our work, as we specify that for a *faithfulness measurable model* the deployed model and measured model should be the same. Without this criterion, there is a dangerous potential for discrepancies between the analyzed model and the deployed model, as has been the case in previous work.

However, there is still a risk that the faithfulness measure can be wrong. This is problematic, as a false faithfulness measure would create unsupported confidence in the importance measure. A key assumption with our proposed metric is that the model provides in-distribution support for any permutation of masked tokens. We take extra care to validate this assumption using sound methodologies. Note that our work is not the only one with this assumption. In particular, ROAR-based metrics (Hooker et al., 2019; Madsen et al., 2022a) also assume in-distribution behavior but did not test for this, while we do test for it. Additionally, we attempt to provide complete transparency regarding the limitations of our work in Section 6 and Appendix B.

Use of anonymized medical data

The Diabetes and Anemia datasets used in this work are based on the anonymized open-access medical database MIMIC-III (Johnson et al., 2016). These datasets are used because they are common in the faithfulness literature (Jain & Wallace, 2019; Madsen et al., 2022a). In particular, they contain many redundancies and are long-sequence datasets. The specific authors of this paper who have worked with this data have undergone the necessary HIPAA certification in order to access this data and comply with HIPAA regulations.

Acknowledgements

Sarath Chandar is supported by the Canada CIFAR AI Chairs program, the Canada Research Chair in Lifelong Machine Learning, and the NSERC Discovery Grant.

Siva Reddy is supported by the Facebook CIFAR AI Chairs program and NSERC Discovery Grant.

Computing resources were provided by the Digital Research Alliance of Canada.

References

Arras, L., Horn, F., Montavon, G., Müller, K.-R., and Samek, W. "What is relevant in a text document?": An interpretable machine learning approach. *PLOS ONE*, 12(8):e0181142, 8 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0181142. URL https://dx.plos.org/10.1371/journal.pone.0181142.

Arras, L., Osman, A., and Samek, W. CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81:14–40, 5 2022. ISSN 15662535. doi: 10.1016/j.inffus.2021.11.008. URL https://linkinghub.elsevier.com/retrieve/pii/S1566253521002335.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer Normalization. *Arxiv*, 2016. URL http://arxiv.org/abs/1607.06450.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M.,

Hansen, K., and Müller, K. R. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11:1803–1831, 12 2010. ISSN 15324435. URL http://arxiv.org/abs/0912.1128.

Bansal, N., Agarwal, C., and Nguyen, A. SAM: The Sensitivity of Attribution Methods to Hyperparameters. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 11–21. IEEE, 6 2020. ISBN 978-1-7281-9360-1. doi: 10.1109/CVPRW50498.2020.00009. URL https://ieeexplore.ieee.org/document/9150607/.

Bastings, J., Ebert, S., Zablotskaia, P., Sandholm, A., and Filippova, K. "Will You Find These Shortcuts?" A Protocol for Evaluating the Faithfulness of Input Salience Methods for Text Classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 976–991, Stroudsburg, PA, USA, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.64. URL https://aclanthology.org/2022.emnlp-main.64.

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M. F., and Eckersley, P. Explainable Machine Learning in Deployment. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 648–657, 9 2019. doi: 10.1145/3351095.3375624. URL https://dl.acm.org/doi/10.1145/3351095.3375624.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. 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, Stroudsburg, PA, USA, 2015. Association for Computational Linguistics. ISBN 9781941643327. doi: 10.18653/v1/D15-1075. URL http://aclweb.org/anthology/D15-1075.

Buckland, S. T., Davison, A. C., and Hinkley, D. V. Bootstrap Methods and Their Application. *Biometrics*, 54(2): 795, 6 1998. ISSN 0006341X. doi: 10.2307/3109789.

Clark, C., Lee, K., Chang, M. W., Kwiatkowski, T., Collins, M., and Toutanova, K. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, volume 1, pp. 2924–2936, 2019. ISBN 9781950737130.

Dagan, I., Glickman, O., and Magnini, B. The PASCAL Recognising Textual Entailment Challenge. In Quiñonero-Candela, J., Dagan, I., Magnini, B., and D'Alché-Buc, F. (eds.), *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pp. 177–190. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. ISBN 978-3-540-33428-6. doi: 10.1007/11736790{_}9. URL http://link.springer.com/10.1007/11736790_9.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, volume 1, pp. 4171–4186. Association for Computational Linguistics (ACL), 10 2019. ISBN 9781950737130. URL http://arxiv.org/abs/1810.04805.

DeYoung, J., Jain, S., Rajani, N. F., Lehman, E., Xiong, C., Socher, R., and Wallace, B. C. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, Stroudsburg, PA, USA, 11 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.408. URL https://www.aclweb.org/anthology/2020.acl-main.408.

Dolan, W. B. and Brockett, C. Automatically Constructing a Corpus of Sentential Paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, pp. 9–16, 2005. URL https://research.microsoft.com/apps/pubs/default.aspx?id=101076.

Doshi-Velez, F. and Kim, B. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv*, 2 2017. URL http://arxiv.org/abs/1702.08608.

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S. J., O'Brien, D., Shieber, S., Waldo, J., Weinberger, D., and Wood, A. Accountability of AI Under the Law: The Role of Explanation. *SSRN Electronic Journal*, Online, 11 2017. ISSN 1556-5068. doi: 10.2139/ssrn.3064761. URL https://www.ssrn.com/abstract=3064761.

Dziedzic, A., Rabanser, S., Yaghini, M., Ale, A., Erdogdu, M. A., and Papernot, N. p-DkNN: Out-of-Distribution Detection Through Statistical Testing of Deep Representations. *arXiv*, 7 2022. URL http://arxiv.org/abs/2207.12545.

Ebrahimi, J., Rao, A., Lowd, D., and Dou, D. HotFlip: White-Box Adversarial Examples for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 2, pp. 31–36, Stroudsburg, PA, USA, 2018. Association for

Computational Linguistics. ISBN 9781948087346. doi: 10.18653/v1/P18-2006. URL http://aclweb.org/anthology/P18-2006.

Fisher, R. A. Statistical Methods for Research Workers. In Kotz, S. and Johnson, N. L. (eds.), *Breakthroughs in Statistics: Methodology and Distribution*, pp. 66–70. Springer New York, New York, NY, 1992. ISBN 978-1-4612-4380-9. doi: 10.1007/978-1-4612-4380-9{_}6. URL http://link.springer.com/10.1007/978-1-4612-4380-9_6.

Goodman, B. and Flaxman, S. European union regulations on algorithmic decision making and a "right to explanation". *AI Magazine*, 38(3):50–57, 2017. ISSN 07384602. doi: 10.1609/aimag.v38i3.2741.

Hooker, S., Erhan, D., Kindermans, P.-J. J., and Kim, B. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 6 2019. URL http://arxiv.org/abs/1806.10758.

Iyer, S., Dandekar, N., and Csernai, K. First Quora Dataset Release: Question Pairs, 2017. URL https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs.

Jacovi, A. and Goldberg, Y. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, Stroudsburg, PA, USA, 4 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL https://www.aclweb.org/anthology/2020.acl-main.386.

Jain, S. and Wallace, B. C. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North*, volume 1, pp. 3543–3556, Stroudsburg, PA, USA, 2 2019. Association for Computational Linguistics. ISBN 9781950737130. doi: 10.18653/v1/N19-1357. URL http://aclweb.org/anthology/N19-1357.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H. W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035, 12 2016. ISSN 20524463. doi: 10.1038/sdata.2016.35. URL http://www.nature.com/articles/sdata201635.

Kindermans, P.-J., Schütt, K., Müller, K.-R., and Dähne, S. Investigating the influence of noise and distractors on the interpretation of neural networks. In *NIPS Workshop on Interpretable Machine Learning in Complex Systems*, 2016. URL http://arxiv.org/abs/1611.07270.

Kwon, G., Prabhushankar, M., Temel, D., and AlRegib, G. Backpropagated Gradient Representations for Anomaly Detection. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12366 LNCS: 206–226, 7 2020. ISSN 16113349. doi: 10.1007/978-3-030-58589-1{_}13. URL http://arxiv.org/abs/2007.09507.

Li, J., Chen, X., Hovy, E., and Jurafsky, D. Visualizing and Understanding Neural Models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 681–691, Stroudsburg, PA, USA, 2016a. Association for Computational Linguistics. ISBN 9781941643914. doi: 10.18653/v1/N16-1082. URL http://aclweb.org/anthology/N16-1082.

Li, J., Monroe, W., and Jurafsky, D. Understanding Neural Networks through Representation Erasure. *arXiv*, 2016b. URL http://arxiv.org/abs/1612.08220.

Lin, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, 7 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv*, 7 2019. ISSN 23318422. URL http://arxiv.org/abs/1907.11692.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pp. 142–150, Portland, Oregon, USA, 6 2011. Association for Computational Linguistics. ISBN 9781932432879. URL https://www.aclweb.org/anthology/P11-1015.

Madsen, A., Meade, N., Adlakha, V., and Reddy, S. Evaluating the Faithfulness of Importance Measures in NLP by Recursively Masking Allegedly Important Tokens and Retraining. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1731–1751, Abu Dhabi, United Arab Emirates, 12 2022a. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.125.

Madsen, A., Reddy, S., and Chandar, S. Post-hoc Interpretability for Neural NLP: A Survey. *ACM Computing Surveys*, 55(8):1–42, 8 2022b. ISSN 0360-

0300. doi: 10.1145/3546577. URL https://dl.acm.org/doi/10.1145/3546577.

Madsen, A., Chandar, S., and Reddy, S. Are self-explanations from Large Language Models faithful? *arXiv*, 1 2024. URL http://arxiv.org/abs/2401.07927.

Marneffe, M.-C. d., Simons, M., and Tonhauser, J. The CommitmentBank: Investigating projection in naturally occurring discourse. *Proceedings of Sinn und Bedeutung*, 23(2):107–124, 2019. ISSN 2629-6055. URL https://ojs.ub.uni-konstanz.de/sub/index.php/sub/article/view/601.

Matan, H., Frostig, T., Heller, R., and Soudry, D. A Statistical Framework for Efficient Out of Distribution Detection in Deep Neural Networks. *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Oy9WeuZD51.

Meister, C., Lazov, S., Augenstein, I., and Cotterell, R. Is Sparse Attention more Interpretable? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 122–129, Stroudsburg, PA, USA, 8 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.17. URL http://arxiv.org/abs/2106.01087https://aclanthology.org/2021.acl-short.17.

Michael R. Chernick and LaBudde, R. A. *An introduction to bootstrap methods with applications to R*. John Wiley & Sons, 2011. ISBN 9780470467046.

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., and Roth, D. Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey. *ACM Computing Surveys*, 56(2):1–40, 2 2024. ISSN 0360-0300. doi: 10.1145/3605943. URL https://dl.acm.org/doi/10.1145/3605943.

Moradi, P., Kambhatla, N., and Sarkar, A. Measuring and improving faithfulness of attention in neural machine translation. *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 2791–2802, 2021. doi: 10.18653/v1/2021.eacl-main.243.

Mudrakarta, P. K., Taly, A., Sundararajan, M., and Dhamdhere, K. Did the model understand the question? In *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, volume 1, pp. 1896–1906, 5 2018. ISBN 9781948087322. doi: 10.18653/v1/p18-1176. URL https://www.aclweb.org/anthology/P18-1176/.

Muennighoff, N., Su, H., Wang, L., Yang, N., Wei, F., Yu, T., Singh, A., and Kiela, D. Generative Representational Instruction Tuning. *arXiv*, 2024. URL http://arxiv.org/abs/2402.09906.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. SQuad: 100,000+ questions for machine comprehension of text. *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pp. 2383–2392, 2016. doi: 10.18653/v1/d16-1264.

Ross, A., Marasović, A., and Peters, M. Explaining NLP Models via Minimal Contrastive Editing (MiCE). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3840–3852, Stroudsburg, PA, USA, 12 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.336. URL https://aclanthology.org/2021.findings-acl.336.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0048-x. URL http://www.nature.com/articles/s42256-019-0048-x.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., and Muller, K.-R. Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673, 11 2017. ISSN 2162-237X. doi: 10.1109/TNNLS.2016.2599820. URL https://ieeexplore.ieee.org/document/7552539/.

Simes, R. J. An Improved Bonferroni Procedure for Multiple Tests of Significance. *Biometrika*, 73(3):751, 12 1986. ISSN 00063444. doi: 10.2307/2336545. URL https://www.jstor.org/stable/2336545?origin=crossref.

Socher, R., Bauer, J., Manning, C. D., and Ng, A. Y. Parsing with compositional vector grammars. In *ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, volume 1, pp. 455–465. Association for Computational Linguistics, 2013. ISBN 9781937284503. URL https://aclanthology.org/P13-1045/.

Sun, R. and Lampert, C. H. KS(conf): A Light-Weight Test if a Multiclass Classifier Operates Outside of Its Specifications. *International Journal of Computer Vision*, 128(4):970–995, 4 2020. ISSN 0920-5691. doi: 10.1007/s11263-019-01232-x.

URL http://link.springer.com/10.1007/s11263-019-01232-x.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *34th International Conference on Machine Learning, ICML 2017*, volume 7, pp. 5109–5118, 3 2017. ISBN 9781510855144. URL http://arxiv.org/abs/1703.01365.

Vashishth, S., Upadhyay, S., Tomar, G. S., and Faruqui, M. Attention Interpretability Across NLP Tasks. *arXiv*, 9 2019. URL http://arxiv.org/abs/1909.11218.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*, 32(July):1–30, 2019a. ISSN 10495258.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *International Conference on Learning Representations*, 2019b. URL https://openreview.net/forum?id=rJ4km2R5t7.

Warstadt, A., Singh, A., and Bowman, S. R. Neural Network Acceptability Judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 11 2019. ISSN 2307-387X. doi: 10.1162/tacl{_}a{_}00290. URL https://direct.mit.edu/tacl/article/43528.

Weston, J., Bordes, A., Chopra, S., Rush, A. M., Van Merriënboer, B., Joulin, A., and Mikolov, T. Towards AI-complete question answering: A set of prerequisite toy tasks. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, 2 2016. URL http://arxiv.org/abs/1502.05698.

Wettig, A., Gao, T., Zhong, Z., and Chen, D. Should You Mask 15% in Masked Language Modeling? *EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 2977–2992, 2 2023. doi: 10.18653/v1/2023.eacl-main.217. URL http://arxiv.org/abs/2202.08005.

Wiegreffe, S. and Pinter, Y. Attention is not not Explanation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, 8 2019. doi: 10.18653/v1/D19-1002. URL https://www.aclweb.org/anthology/D19-1002.

Williams, A., Nangia, N., and Bowman, S. 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)*, volume 1, pp. 1112–1122, Stroudsburg, PA, USA, 2018. Association for Computational Linguistics. ISBN 9781948087278. doi: 10.18653/v1/N18-1101. URL http://aclweb.org/anthology/N18-1101.

Yang, J., Zhou, K., Li, Y., and Liu, Z. Generalized Out-of-Distribution Detection: A Survey. *arXiv*, 10 2021. URL http://arxiv.org/abs/2110.11334.

Yin, K. and Neubig, G. Interpreting Language Models with Contrastive Explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 184–198, Stroudsburg, PA, USA, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.14. URL https://github.com/kayoyin/interpret-lm.https://aclanthology.org/2022.emnlp-main.14.

Zaman, K. and Belinkov, Y. A Multilingual Perspective Towards the Evaluation of Attribution Methods in Natural Language Inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, pp. 1556–1576, 4 2022. URL http://arxiv.org/abs/2204.05428.

Zhang, S., Fan, R., Liu, Y., Chen, S., Liu, Q., and Zeng, W. Applications of transformer-based language models in bioinformatics: a survey. *Bioinformatics Advances*, 3(1), 1 2023. ISSN 2635-0041. doi: 10.1093/bioadv/vbad001. URL https://academic.oup.com/bioinformaticsadvances/article/doi/10.1093/bioadv/vbad001/6984737.

Zhou, Y. and Shah, J. The Solvability of Interpretability Evaluation Metrics. In *Findings of the Association for Computational Linguistics: EACL*, 2023. URL http://arxiv.org/abs/2205.08696.

A. Additional discussion

In this section, we discuss additional observations that can be made from the main paper results. These observations are valuable but are not necessary for the main message in the paper.

Table 1 show that the RACU scores have a lower variance (confidence interval) using our methodology compared to Recursive ROAR. This is likely because Recursive ROAR leaks the gold label (Madsen et al., 2022a), which causes oscillation in the faithfulness curve.

We also observe that gradient-based explanations are more faithful when using our model. We suspect this is partially also because there is no leakage issue. However, previous work has also shown that gradient-based methods behave more favorably on robust models in computer vision (Bansal et al., 2020). Using masked fine-tuning can be seen as a robustness objective, as the model becomes robust to missing information.

Surprisingly, input times gradient (x $\odot$ grad) appears to be the worst explanation. This is a curious result because both *gradient* (Grad) and *integrated gradient* (IG) perform much better. Recall that *integrated gradient* is essentially $x \odot \int \text{grad}$ (Sundararajan et al., 2017). Perhaps a version of *integrated gradient* where the x-hadamard product (x $\odot$) is not used might improve the faithfulness.

Finally, Diabetes and bAbI-2 get 90% RACU faithfulness (using absolute importance measures). This suggests that there are only a few words that are critical for the prediction of these tasks. This may indicate a problem with their efficacy as benchmarking datasets. However, for Diabetes, which is purely a diagnostic dataset (Jain & Wallace, 2019; Johnson et al., 2016), this may be fine if those tokens should be critical to the task.

B. Additional limitations

B.1. Not a post-hoc method

While this work solves existing limitations with previous methods, it introduces the significant limitation that it, by definition, requires a *faithfulness measurable model*. As such, the question of faithfulness needs to be considered ahead of time when developing a model. It can not be an afterthought, which is often how interpretability is approached (Bhatt et al., 2019; Madsen et al., 2022b).

While this is a significant limitation, considering explanation ahead of deployment is increasingly becoming a legal requirement (Doshi-Velez et al., 2017). Currently, the European Union provides a "right to explanation" regarding automatic decisions, which includes NLP models (Goodman & Flaxman, 2017).

B.2. In-distribution is impossible to prove

Because in-distribution is always the null hypothesis, it is impossible to statistically show that inputs are truly in-distribution. The typical approach to similar statistical questions[3] is to keep validating in-distribution using various methods. Unfortunately, the literature on this topic in deep learning is extremely limited (Yang et al., 2021; Sun & Lampert, 2020; Matan et al., 2022; Dziedzic et al., 2022; Kwon et al., 2020).

Therefore, we would advocate for more work on identifying out-of-distribution inputs using non-parametric methods that primarily consider the model's internal state. Using parametric methods or works that use axillary models is more well-explored but not useful for our purpose.

B.3. Requires repeated measures on the test dataset

Because datasets have redundancies, it is necessary to reevaluate the importance measures (Hooker et al., 2019; Madsen et al., 2022a). This leads to an increased computational cost.

However, unlike previous work (Madsen et al., 2022a), our method only requires reevaluation of the test dataset, which is often quite small. Additionally, some IMs, such as the beam-search method (Zhou & Shah, 2023), take dataset redundancies into account and therefore do not require reevaluation. Reevaluation could be done if desired but would result in the exact same results.

B.4. Measures only faithfulness

There are other important aspects to an importance measure, such as if the explanation is useful to humans (Doshi-Velez & Kim, 2017), but our method does not measure this. We consider this a separate topic, as it is an HCI question that requires experimental studies with humans. The topic of faithfulness can not be measured by humans (Jacovi & Goldberg, 2020), as neural networks are too complicated for humans to manually evaluate if an explanation is true.

C. Datasets

The datasets used in this work are all public and listed below. They are all used for their intended use, which is measuring classification performance. The Diabetes and Anemia datasets are from MIMIC-III which requires a HIPPA certification for data analysis (Johnson et al., 2016). The first author complies with this and has not shared the data with any, including other authors.

Note that when computing the importance measure on paired-sequence tasks, only the first sequence is consid-

[3] A similar well-explored statistical question is how to show that the error in a linear model is normally distributed.

Table 2. Datasets used, all datasets are either single-sequence or sequence-pair datasets. All datasets are sourced from GLUE (Wang et al., 2019b), SuperGLUE (Wang et al., 2019a), MIMIC-III (Johnson et al., 2016), or bAbI (Weston et al., 2016). The decisions regarding which metrics are used are also from these sources. The class-majority baseline is when the most frequent class is always selected.

Type	Dataset	Size			Inputs		Performance		Citation
		Train	Validation	Test	masked	auxilary	metric	class-majority	
NLI	RTE	1992	498	277	`sentence1`	`sentence2`	Accuracy	47%	Dagan et al., 2006
	SNLI	549367	9842	9824	`premise`	`hypothesis`	Macro F1	34%	Bowman et al., 2015
	MNLI	314162	78540	9815	`premise`	`hypothesis`	Accuracy	35%	Williams et al., 2018
	QNLI	83794	20949	5463	`sentence`	`question`	Accuracy	51%	Rajpurkar et al., 2016
	CB	200	50	56	`premise`	`hypothesis`	Macro F1	22%	Marneffe et al., 2019
Paraphrase	MRPC	2934	734	408	`sentence1`	`sentence2`	Macro F1	41%	Dolan et al., 2005
	QQP	291077	72769	40430	`question1`	`question2`	Macro F1	39%	Iyer et al., 2017
Sentiment	SST2	53879	13470	872	`sentence`	–	Accuracy	51%	Socher et al., 2013
	IMDB	20000	5000	25000	`text`	–	Macro F1	33%	Maas et al., 2011
Diagnosis	Anemia	4262	729	1243	`text`	–	Marco F1	39%	Johnson et al., 2016
	Diabetese	8066	1573	1729	`text`	–	Marco F1	45%	Johnson et al., 2016
Acceptability	CoLA	6841	1710	1043	`sentence`	–	Matthew	0%	Warstadt et al., 2019
QA	BoolQ	7542	1885	3270	`passage`	`question`	Accuracy	62%	Clark et al., 2019
	bAbI-1	8000	2000	1000	`paragraph`	`question`	Micro F1	15%	Weston et al., 2016
	bAbI-2	8000	2000	1000	`paragraph`	`question`	Micro F1	19%	Weston et al., 2016
	bAbI-3	8000	2000	1000	`paragraph`	`question`	Micro F1	18%	Weston et al., 2016

ered. This is to stay consistent with previous work (Jain & Wallace, 2019; Madsen et al., 2022a) and because for tasks like document-based Q&A (e.g., bAbI), it does not make sense to mask the question.

The essential statistics for each dataset, and which part is masked and auxiliary, are specified in Table 2.

D. Models

In this paper, we use the RoBERTa model (Liu et al., 2019), although any masked language model of similar size or larger is likely to work. We choose RoBERTa model, because converges consistently and reasonable hyperparameters are well established. This should make reproducing the results in this paper easier. We use both the `base` (125M parameters) and `large` size (355M parameters).

The hyperparameters are defined by Liu et al. (2019, Appendix C, GLUE). Although these hyperparameters are for the GLUE tasks, we use them for all tasks. The one exception, is that the maximum number of epoch is higher. This is because when *masked fine-tuning* require more epochs. In Table 3 we specify the max epoch parameter. However, when using early stopping with the validation dataset, the optimization is not sensitive to the specific number of epochs, lower numbers are only used to reduce the compute time.

Table 3. Dataset statistics. Performance metrics are the mean with a 95% confidence interval.

Dataset	max epoch	Performance	
		RoBERTa-base	RoBERTa-large
BoolQ	15	$80\%^{+0.2}_{-0.2}$	$85\%^{+0.3}_{-0.3}$
CB	50	$65\%^{+17.6}_{-47.9}$	$87\%^{+3.1}_{-8.2}$
CoLA	15	$59\%^{+1.3}_{-1.1}$	$66\%^{+0.8}_{-0.8}$
IMDB	10	$95\%^{+0.2}_{-0.2}$	$96\%^{+0.2}_{-0.4}$
Anemia	20	$84\%^{+0.8}_{-0.7}$	$84\%^{+0.5}_{-0.8}$
Diabetes	20	$76\%^{+0.9}_{-0.9}$	$77\%^{+0.6}_{-1.6}$
MNLI	10	$87\%^{+0.4}_{-0.2}$	$90\%^{+0.3}_{-0.2}$
MRPC	20	$86\%^{+0.8}_{-0.7}$	$87\%^{+0.5}_{-1.2}$
QNLI	20	$92\%^{+0.1}_{-0.1}$	$94\%^{+0.1}_{-0.2}$
QQP	10	$90\%^{+0.1}_{-0.1}$	$91\%^{+0.0}_{-0.1}$
RTE	30	$75\%^{+1.4}_{-2.9}$	$83\%^{+1.3}_{-1.4}$
SNLI	10	$91\%^{+0.1}_{-0.2}$	$92\%^{+0.1}_{-0.2}$
SST2	10	$94\%^{+0.2}_{-0.2}$	$96\%^{+0.2}_{-0.2}$
bAbI-1	20	$100\%^{+0.0}_{-0.1}$	$100\%^{+0.0}_{-0.0}$
bAbI-2	20	$99\%^{+0.1}_{-0.1}$	$100\%^{+0.1}_{-0.1}$
bAbI-3	20	$90\%^{+0.2}_{-0.3}$	$90\%^{+0.5}_{-0.5}$

E. Importance measure details

This section provides additional details on the importance measures, which were only briefly described in Section 4. In particular, not all importance measures have both signed and absolute variants (Table 4), this section should clarify why.

Table 4. Overview of the importance measures used in this paper. Note that not all importance measures exist in both signed and absolute variants due to their mathematical construction.

Category	Full name	Short name	Variants
Gradient	Gradient w.r.t. input	Grad	absolute
	Input times gradient	x ⊙ grad	both
	Integrated gradient	IG	both
Occlusion	Leave-on-out	LOO	both
	Beam-search	Beam	signed

E.1. Gradient-based

The common idea is that if a small change in the input causes a large change in the output, then that indicates importance.

Gradient (Grad) measures the mentioned relationship using the gradient, which is a linear approximation. Let $f(\mathbf{x})$ be the model with input $\mathbf{x} \in \mathbb{R}^{T \times V}$ (V is vocabulary-size and T is sequence-length), then the gradient is $\nabla_{\mathbf{x}} f(\mathbf{x})_y \in \mathbb{R}^{T \times V}$ (Baehrens et al., 2010; Li et al., 2016a). Because the desire is an importance measure for each token (i.e., $\mathbb{R}^T$), the vocabulary dimension is typically[4] reduced using either L_1 or L_2 norm. In this paper, we consider both.

Input times Gradient ($\mathbf{x} \odot$ grad) multiplies the gradient with the input, which some argue to be better (Kindermans et al., 2016) i.e., $\mathbf{x} \odot \nabla_{\mathbf{x}} f(\mathbf{x}) \in \mathbb{R}^{T \times V}$. Because $\mathbf{x}$ is a one-hot-encoding, using any norm function is just the absolute value of the non-zero element, which is what is typically used. However, we observe that for NLP, one could also consider the signed version of this, where the non-zero element is simply picked out. In this paper, we consider both the signed and absolute variants.

Integrated Gradient (IG) by Sundararajan et al. (2017) is a very popular explanation method. This can be seen as an extension of the *input times gradient* method, and therefore we also consider a signed and absolute variant for this. The method works by sampling gradients between a baseline $f(\mathbf{b})$ and the input $f(\mathbf{x})$. We use a zero-vector baseline and 20 samples, as is commonly done in NLP literature (Mudrakarta et al., 2018)

$$\begin{aligned} \text{IG}(\mathbf{x}) &= (\mathbf{x} - \mathbf{b}) \odot \frac{1}{k} \sum_{i=1}^{k} \nabla_{\tilde{\mathbf{x}}_i} f(\tilde{\mathbf{x}}_i)_c \\ \tilde{\mathbf{x}}_i &= \mathbf{b} + \frac{i}{k}(\mathbf{x} - \mathbf{b}). \end{aligned} \tag{1}$$

[4]The topic of vocabulary-dimension reduction is rarely discussed in papers.

E.2. Occlusion-based

Rather than linear-approximating the relation between input and output using gradients, the relationship can also be approximated by removing or masking each token, one by one, and measuring how it affects the output.

One rarely addressed concern is that removing/masking tokens is likely to be out-of-distribution (Zhou & Shah, 2023). This would cause an otherwise sound method to become unfaithful. However, because the proposed *faithfulness measurable masked language model* supports masking, this should not be a concern. Importantly, this exemplifies how a *faithfulness measurable model* may also produce more faithful explanations.

Leave-on-out (LOO) directly computes the difference between the model output difference between the unmasked input and the input with the i'th token masked, i.e. the importance is $\{f(\hat{\mathbf{x}})_y - f(\tilde{\mathbf{x}}_i)_y\}_{i=1}^{T} \in \mathbb{R}^T$ (Li et al., 2016b). Even though this importance measure does not have a vocabulary dimension, it is common to take its absolute (Meister et al., 2021). However, we consider both absolute and signed variants.

Optimizing for faithfulness (Beam). Zhou & Shah (2023) propose that we can optimize for the faithfulness metric itself rather than using heuristics. The central idea is to use beam-search, where the generated sequence is the optimal masking order of tokens. Each iteration of the beam-search masks one additional token, where the token is selected by testing every possibility and maximizing the faithfulness metric. This could be reframed as a recursive version of leave-on-out. They propose several optimization targets, but since our faithfulness metric is analog to comprehensiveness, we use this variation.

The number of forward passes is $\mathcal{O}(B \cdot T^2)$; it is not exactly $B \cdot T^2$ because many of them become redundant. This is quite computationally costly, although one advantage is that this explanation is inherently recursive, hence it is not necessary to reevaluate the importance measure in each iteration of the faithfulness metric. However, for long sequence datasets, such as IMDB, BaBi-3, Anemia, and Diabetes, it is not feasible to apply this explanation. In our experiments, we use a beam-size of $B = 10$.

F. Workflow algorithms

This section provides the workflow details and algorithms, which were mentioned in Section 3.

F.1. Masked fine-tuning

Masked fine-tuning is a multi-task learning method, where one task is the typical unmasked performance and the other

task is masking support. We achieve this by uniformly masking half of a mini-batch, at between 0% and 100% masking, and do not modify the other half. Note that this is slightly different from some multi-task learning methods, which may sample randomly between the two tasks, where we split deterministically. Other methods may also switch between the two tasks in each step, we don't do this as it can create unstable oscillations. Instead, both tasks are included in the same mini-batch.

There are many approaches to implementing masked fine-tuning with identical results, however Algorithm 1 presents our implementation.

Algorithm 1 Creates the mini-batches used in masked fine-tuning.

Require: B is a mini-batch with N randomly sampled observations from the training dataset. $[\mathcal{M}]$ is the masking token.

```
function MiniBatch(B)
  M ← ∅ {Stores new mini-batch}
  for i ← 1 to N do
    if i is even then
      r ← SampleUniform(0, 1)
      M_i ← MaskTokens(r, B_i) {Masks r% randomly selected tokens in B_i.}
    else
      M_i ← B_i
    end if
  end for
  return M
end function

function MaskTokens(x, r)
  x̃ ← x
  for t ← 1 to T do
    s ← SampleUniform(0, 1)
    if s < r then
      x_t ← [M] {Masks token t.}
    end if
  end for
  return x̃
end function
```

F.2. In-distribution validation (MaSF)

MaSF is a statistical in-distribution test developed by Matan et al. (2022). *MaSF* is an acronym that stands for Max-Simes-Fisher, which is the order of aggregation functions it uses. Matan et al. (2022) presented some other combinations and orders of aggregation functions but found this to be the best. However, many combinations had similar performance in their benchmark, so we do not consider this particular choice to be important and assume it generalizes well to our case. In Section 5.2, we verified this with an ablation study.

At its core, *MaSF* is an aggregation of many in-distribution p-values, where each p-value is from an in-distribution test of a latent embedding. That is, given a history of embedding observations, which presents a distribution, what is the probability of observing the new embedding or something more extreme? For example, if that probability is less than 5%, it could be classified as out-of-distribution at a 5% risk of a false-positive.

However, a model has many internal embeddings, and thus, there will be many p-values. If each were tested independently, there would be many false positives. This is known as p-hacking. To prevent this, the p-values are aggregated using the Simes and Fisher methods, which are aggregation methods for p-values that prevent this issue. Once the p-values are aggregated, it becomes a "global null-test". This means the aggregated statistical test checks if any of the embeddings are out-of-distribution.

Emperical CDF. Each p-value is computed using an empirical communicative density function (CDF). A nice property of an empirical CDF is that it doesn't assume any distribution, a property called non-paramatic. It is however still a model, only if an infinite amount of data was available would it represent the true distribution.

A CDF measures the probability of observing z or less than z, i.e., $\mathbb{P}(Z \leq z)$. The empirical version simply counts how many embeddings were historically less than the tested embedding, as shown in (2). However, as we are also interested in cases where the embedding is abnormally large, hence we also use $\mathbb{P}(Z > z) = 1 - \mathbb{P}(Z \leq z)$. We are then interested in the most unlikely case, which is known as the two-sided p-value, i.e. $\min(\mathbb{P}(Z \leq z), 1 - \mathbb{P}(Z \leq z))$.

$$\mathbb{P}(Z \leq z) = \frac{1}{|Z_{\text{emp}}|} \sum_{i=1}^{|Z_{\text{emp}}|} 1[Z_{\text{emp},i} < z] \tag{2}$$

The historical embeddings are collected by running the model on the validation dataset. Note that for this to be accurate, the validation dataset should be i.i.d. with the training dataset. This can easily be accomplished by randomly splitting the datasets, which is common practice, and transforming the validation dataset the same way as the training dataset.

Algorithm. In the case of *MaSF*, the embeddings are first aggregated along the sequence dimension using the max operation. Matan et al. (2022) only applied *MaSF* to computer vision, in which case it was the width and height dimensions. However, we generalize this to NLP by swapping width and height with the sequence dimension.

The max-aggregated embeddings from the validation dataset

provide the historical data for the empirical CDFs. If a network has L layers, each with H latent dimensions, there will be $L \cdot H$ CDFs. The same max-aggregated embeddings are then transformed into p-values using those CDFs. Next, the p-values are aggregated using Simes's method (Simes, 1986) along the latent dimension, which provides another set of CDFs and p-values, one for each layer (L CDFs and p-values). Finally, those p-values are aggregated using Fisher's method (Fisher, 1992), providing one CDF and one p-value for each observation.

The algorithm for *MaSF* can be found in Algorithm 2. While this algorithm does work, a practical implementation is in our experience non-trivial, as for an entire test dataset ($\mathcal{D}_T$) are $\mathcal{O}(|\mathcal{D}_T| \cdot H \cdot L)$ CDFs evaluations, each involving $\mathcal{O}(\mathcal{D}_V)$ comparisons. While this is computationally trivial on a GPU, it can require a lot of memory usage when done in parallel. Therefore, we found that a practical implementation must batch over both the test and validation datasets.

Algorithm 2 MaSF algorithm, which provides p-values under the in-distribution null-hypothesis.

Require: x is the input. $f_e(x) \in \mathbb{R}^{T \times H \times L}$ provides the model embeddings, where T is sequence-length, H is the hidden-size, and L is the number of layers. $\mathbb{P}$ are the empirical CDFs; these are collected by running the MaSF algorithm on a validation dataset.

function MaSF(x, $\mathbb{P}$)
  $e \leftarrow f_e(x)$ {Get embeddings}
  for $l \leftarrow 1$ **to** L **do**
    for $h \leftarrow 1$ **to** H **do**
      $z_{l,h}^{(1)} \leftarrow \max_{t=1}^{T} e_{l,h,t}$ {Ma-step}
      $\tilde{p}_{l,h}^{(1)} \leftarrow \mathbb{P}_{l,h}^{(1)}(Z < z_{l,h}^{(1)})$
      $p_{l,h}^{(1)} \leftarrow \min(\tilde{p}_{l,h}^{(1)}, 1 - \tilde{p}_{l,h}^{(1)})$
    end for
    $z_l^{(2)} \leftarrow \text{Simes}(p_{l,:}^{(1)})$ {S-step}
    $\tilde{p}_l^{(2)} \leftarrow \mathbb{P}_l^{(2)}(Z < z_l^{(2)})$
    $p_l^{(2)} \leftarrow \min(\tilde{p}_l^{(2)}, 1 - \tilde{p}_l^{(2)})$
  end for
  $z^{(3)} \leftarrow \text{Fisher}(p^{(2)})$ {F-step}
  $\tilde{p}^{(3)} \leftarrow \mathbb{P}^{(3)}(Z < z^{(3)})$
  $p^{(3)} \leftarrow 1 - \tilde{p}^{(3)}$
  return $p^{(3)}$
end function

function Simes(p)
  $q \leftarrow \text{SortAscending}(p)$
  return $\min_{i=1}^{N} q_i \frac{N}{i}$
end function

function Fisher(p)
  return $-2 \sum_{i=1}^{N} \log(p_i)$
end function

F.3. Faithfulness metric

Faithfulness is measured by masking 10% the tokens according to an importance measure, such that the most important tokens are masked. This new 10%-masked observation is then used as a new input, and the model prediction is explained. This is then repeated until 100% masking. This process is identical to the Recursive ROAR approach by (Madsen et al., 2022a), except without re-training. For an explanation, we define the evaluation procedure in Algorithm 3. However, in terms of Figure 6, Algorithm 3 only provides the "importance measure" curve. To get the random baseline, simply apply Algorithm 3 again with a random explanation.

Algorithm 3 Measures the masked model performance given an explanation.

Require: $\text{IM}(f, x, y) \in \mathbb{R}^T$ explains the model f for the input x and label y, with sequence-length T. δ is the iterations step-size (e.g. 10%) and $[\mathcal{M}]$ is the masking token.

function RecursiveEval(IM, f, x, y, δ)
  $\tilde{x}_0 \leftarrow x$
  $p_0 \leftarrow \text{PerformanceMetric}(f(x), y)$
  for $i \leftarrow 1$ **to** $1/\delta$ **do**
    $e_i \leftarrow \text{IM}(f, \tilde{x}_{i-1}, y)$
    $\tilde{x}_i \leftarrow \text{AddMask}(e_i, \tilde{x}_{i-1}, \delta)$ {Mask $\delta \cdot T$ more tokens in $\tilde{x}_{i-1}$ using scores e_i.}
    $p_i \leftarrow \text{PerformanceMetric}(f(\tilde{x}_i), y)$
  end for
  return p
end function

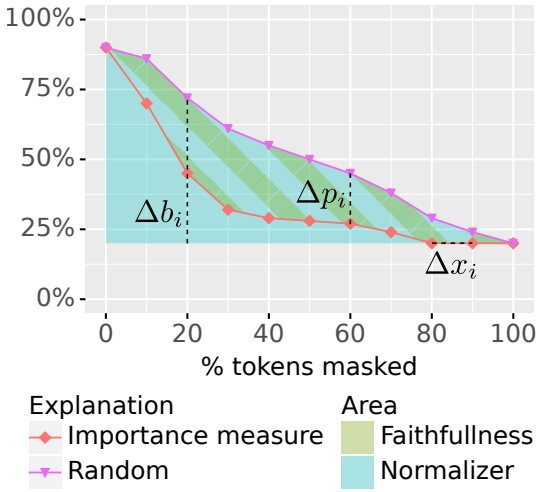


Figure 6. Visualization of the faithfulness calculation done in (3). The *faithfulness* area is the numerator in (3), while the *normalizer* area is the denominator. – Figure by Madsen et al. (2022a), included with permission.

Relative Area between Curves (RACU) ACU is the "faithfulness" area between the "importance measure" curve and the "random" curve as shown in Figure 6. This was not explicitly defined by Madsen et al. (2022a), who defined RACU.

RACU is a normalization of ACU, normalized by the theoretically optimal explanation, which archives the performance of 100% masking immediately. However, as argued in Section 5.3 this normalization is only valid for absolute importance measures. Therefore, we also include the unnormalized score (AUC) in this paper. Both metrics are defined in Equation (3).

Unfortunately, because different models (e.g., RoBERTa-base versus RoBERTa-large) have different performance characteristics, it is not possible to compare unnormalized score (AUC) between different models (Arras et al., 2017). For this reason, RACU may still be useful for signed metrics, even though it can not be interpreted as a 0%-100% scale.

$$
\begin{aligned}
\text{ACU} &= \sum_{i=0}^{I-1} \frac{1}{2} \Delta x_i (\Delta p_i + \Delta p_{i+1}) \\
\text{RACU} &= \frac{\text{ACU}}{\sum_{i=0}^{I-1} \frac{1}{2} \Delta x_i (\Delta b_i + \Delta b_{i+1})} \\
\text{where } \Delta x_i &= x_{i+1} - x_i \quad \textit{step size} \\
\Delta p_i &= b_i - p_i \quad \textit{performance delta} \\
\Delta b_i &= b_i - b_I \quad \textit{baseline delta}
\end{aligned}
\tag{3}
$$

G. Compute

This section reports the compute resources and requirements. The compute hardware specifications are in Table 5 and were the same for all experiments. All computing was performed using 99% hydroelectric power.

Table 5. The computing hardware used. Note, that a shared user system were used, only the allocated resources are reported.

CPU	12 cores, Intel Silver 4216 Cascade Lake @ 2.1GHz
GPU	1x NVidia V100 (32G HBM2 memory)
Memory	24 GB

Note that the importance measures computed are the same for both the faithfulness results and the out-of-distribution results. Hence, these do not need to be computed twice. Additionally, the beam-search method is in itself recursive, so this was only computed for 0% masking.

G.1. Implementation

We use the HuggingFace implementation of RoBERTa and the TensorFlow framework. The code is available at https://github.com/AndreasMadsen/faithfulness-measurable-models.

G.2. Walltimes

We here include the walltimes for all experiments.

- Table 6 shows wall-times for the masked fine-tuning.
- Table 7 shows wall times for the in-distribution validation, not including importance measures.
- Table 8 shows wall-times for the faithfulness evaluation, not including importance measures.
- Table 9 shows wall-times for the importance measures.

Table 6. Walltime for fine-tuning. Masked fine-tuning does not affect the total wall time in our setup.

Dataset	Walltime [hh:mm]	
	RoBERTa-base	RoBERTa-large
BoolQ	00:51	02:03
CB	00:06	00:12
CoLA	00:17	00:33
IMDB	01:44	04:02
Anemia	00:48	02:04
Diabetes	01:34	04:04
MNLI	06:39	14:47
MRPC	00:12	00:27
QNLI	04:03	09:12
QQP	05:13	11:52
RTE	00:18	00:43
SNLI	04:57	10:38
SST2	01:19	02:44
bAbI-1	00:27	01:01
bAbI-2	00:50	02:05
bAbI-3	01:43	04:28
sum	01:43	04:28
x5 seeds	08:37	22:21

Table 7. Walltime for in-distribution validation. This does not include importance measure calculations. See Table 9.

Dataset	Walltime [hh:mm]	
	RoBERTa-base	RoBERTa-large
BoolQ	00:04	00:09
CB	00:01	00:02
CoLA	00:02	00:04
IMDB	00:15	00:44
Anemia	00:01	00:03
Diabetes	00:01	00:04
MNLI	00:20	00:57
MRPC	00:02	00:03
QNLI	00:05	00:12
QQP	00:47	02:13
RTE	00:02	00:04
SNLI	00:04	00:09
SST2	00:09	00:25
bAbI-1	00:01	00:03
bAbI-2	00:02	00:05
bAbI-3	00:01	00:03
sum	02:06	05:25
x5 seeds	10:30	27:09

Table 8. Walltime for faithfulness evaluation. This does not include importance measure calculations. See Table 9.

Dataset	Walltime [hh:mm]	
	RoBERTa-base	RoBERTa-large
BoolQ	00:02	00:05
CB	00:00	00:01
CoLA	00:01	00:02
IMDB	00:13	00:39
Anemia	00:01	00:02
Diabetes	00:01	00:03
MNLI	00:02	00:05
MRPC	00:01	00:02
QNLI	00:01	00:04
QQP	00:05	00:13
RTE	00:01	00:02
SNLI	00:01	00:03
SST2	00:01	00:01
bAbI-1	00:00	00:01
bAbI-2	00:01	00:02
bAbI-3	00:00	00:02
sum	00:38	01:34
x5 seeds	03:13	07:51

Table 9. Walltime for importance measures. Note that because the beam-search method (Beam) scales quadratic with the sequence-length, it is not feasible to compute for all datasets.

Dataset	IM	Walltime [hh:mm]	
		RoBERTa-base	RoBERTa-large
bAbI-1	Beam	00:54	02:24
	Grad (L_1)	00:01	00:04
	Grad (L_2)	00:02	00:04
	x $\odot$ grad (abs)	00:01	00:04
	x $\odot$ grad (sign)	00:01	00:04
	IG (abs)	00:04	00:12
	IG (sign)	00:04	00:11
	LOO (abs)	00:24	00:49
	LOO (sign)	00:24	00:49
	Random	00:00	00:00
bAbI-2	Beam	20:56	61:24
	Grad (L_1)	00:02	00:06
	Grad (L_2)	00:02	00:05
	x $\odot$ grad (abs)	00:02	00:05
	x $\odot$ grad (sign)	00:02	00:05
	IG (abs)	00:10	00:29
	IG (sign)	00:10	00:29
	LOO (abs)	00:39	01:26
	LOO (sign)	00:39	01:26
	Random	00:00	00:00
bAbI-3	Beam	–	–
	Grad (L_1)	00:02	00:05
	Grad (L_2)	00:02	00:05
	x $\odot$ grad (abs)	00:01	00:04
	x $\odot$ grad (sign)	00:01	00:04
	IG (abs)	00:18	00:54
	IG (sign)	00:18	00:53
	LOO (abs)	01:09	03:18
	LOO (sign)	01:09	03:18
	Random	00:00	00:00
BoolQ	Beam	00:33	01:22
	Grad (L_1)	00:05	00:11
	Grad (L_2)	00:05	00:11
	x $\odot$ grad (abs)	00:04	00:10
	x $\odot$ grad (sign)	00:04	00:10
	IG (abs)	00:39	01:48
	IG (sign)	00:39	01:49
	LOO (abs)	00:16	00:38
	LOO (sign)	00:16	00:38
	Random	00:00	00:00
CB	Beam	00:45	02:09
	Grad (L_1)	00:01	00:03
	Grad (L_2)	00:01	00:03
	x $\odot$ grad (abs)	00:01	00:03
	x $\odot$ grad (sign)	00:01	00:03
	IG (abs)	00:01	00:04
	IG (sign)	00:01	00:04
	LOO (abs)	00:09	00:19
	LOO (sign)	00:09	00:19
	Random	00:00	00:00
CoLA	Beam	00:11	00:19
	Grad (L_1)	00:02	00:05
	Grad (L_2)	00:02	00:04
	x $\odot$ grad (abs)	00:02	00:04
	x $\odot$ grad (sign)	00:02	00:04
	IG (abs)	00:04	00:09
	IG (sign)	00:04	00:09
	LOO (abs)	00:09	00:18
	LOO (sign)	00:09	00:18
	Random	00:00	00:00
Anemia	Beam	–	–
	Grad (L_1)	00:02	00:06
	Grad (L_2)	00:02	00:06
	x $\odot$ grad (abs)	00:01	00:05
	x $\odot$ grad (sign)	00:01	00:05
	IG (abs)	00:23	01:08
	IG (sign)	00:23	01:08
	LOO (abs)	02:23	06:58
	LOO (sign)	02:23	07:01
	Random	00:00	00:00
Diabetes	Beam	–	–
	Grad (L_1)	00:03	00:07
	Grad (L_2)	00:03	00:07
	x $\odot$ grad (abs)	00:02	00:06
	x $\odot$ grad (sign)	00:02	00:06
	IG (abs)	00:32	01:34
	IG (sign)	00:32	01:34
	LOO (abs)	03:19	09:44
	LOO (sign)	03:17	09:45
	Random	00:00	00:00
MRPC	Beam	00:14	00:33
	Grad (L_1)	00:02	00:04
	Grad (L_2)	00:02	00:04
	x $\odot$ grad (abs)	00:02	00:04
	x $\odot$ grad (sign)	00:02	00:04
	IG (abs)	00:03	00:07
	IG (sign)	00:03	00:07
	LOO (abs)	00:08	00:17
	LOO (sign)	00:08	00:17
	Random	00:00	00:00
RTE	Beam	01:32	04:26
	Grad (L_1)	00:02	00:04
	Grad (L_2)	00:02	00:04
	x $\odot$ grad (abs)	00:02	00:04
	x $\odot$ grad (sign)	00:02	00:04
	IG (abs)	00:04	00:09
	IG (sign)	00:04	00:09
	LOO (abs)	00:10	00:22
	LOO (sign)	00:10	00:22
	Random	00:00	00:00
SST2	Beam	00:18	00:43
	Grad (L_1)	00:02	00:04
	Grad (L_2)	00:02	00:04
	x $\odot$ grad (abs)	00:02	00:04
	x $\odot$ grad (sign)	00:02	00:04
	IG (abs)	00:04	00:09
	IG (sign)	00:04	00:09
	LOO (abs)	00:09	00:19
	LOO (sign)	00:10	00:19
	Random	00:00	00:00
SNLI	Beam	01:10	02:38
	Grad (L_1)	00:05	00:07
	Grad (L_2)	00:06	00:07
	x $\odot$ grad (abs)	00:05	00:06
	x $\odot$ grad (sign)	00:04	00:06
	IG (abs)	00:21	00:57
	IG (sign)	00:21	00:56
	LOO (abs)	00:12	00:26
	LOO (sign)	00:12	00:26
	Random	00:01	00:00
IMDB	Beam	–	–
	Grad (L_1)	00:34	01:18
	Grad (L_2)	00:34	01:17
	x $\odot$ grad (abs)	00:22	01:03
	x $\odot$ grad (sign)	00:22	01:03
	IG (abs)	06:49	20:08
	IG (sign)	06:54	20:09
	LOO (abs)	25:02	73:17
	LOO (sign)	24:48	72:55
	Random	00:01	00:01
MNLI	Beam	05:44	15:34
	Grad (L_1)	00:05	00:11
	Grad (L_2)	00:05	00:11
	x $\odot$ grad (abs)	00:04	00:09
	x $\odot$ grad (sign)	00:04	00:09
	IG (abs)	00:35	01:35
	IG (sign)	00:35	01:34
	LOO (abs)	00:19	00:46
	LOO (sign)	00:19	00:46
	Random	00:00	00:00
QNLI	Beam	06:39	18:51
	Grad (L_1)	00:04	00:08
	Grad (L_2)	00:04	00:08
	x $\odot$ grad (abs)	00:03	00:07
	x $\odot$ grad (sign)	00:03	00:08
	IG (abs)	00:23	01:03
	IG (sign)	00:23	01:04
	LOO (abs)	00:17	00:43
	LOO (sign)	00:17	00:43
	Random	00:00	00:00
QQP	Beam	04:44	11:12
	Grad (L_1)	00:12	00:26
	Grad (L_2)	00:12	00:26
	x $\odot$ grad (abs)	00:10	00:22
	x $\odot$ grad (sign)	00:10	00:22
	IG (abs)	01:48	04:57
	IG (sign)	01:48	04:59
	LOO (abs)	00:36	01:24
	LOO (sign)	00:36	01:23
	Random	00:01	00:01
	sum	145:00	406:58
	x5 seeds	725:02	2034:53

H. Masked fine-tuning

In Section 5.1, we show selected results for unmasked performance and 100% masked performance. In this appendix, we extend those results to all 16 datasets. In addition to this, this appendix contains a more detailed ablation study, where the training strategy and validation strategy are considered separate. As such, the results in Section 5.1 are a strict subset of these detailed results. In Table 10 we show how the terminologies relate.

Table 10. This table relates terminologies between the fine-tuning strategies mentioned in Section 5.1 and the training strategy and validation strategy terms.

Section 5.1	Training strategy	Validation strategy
Masked fine-tuning	Use 50/50	Use both
Plain fine-tuning	No masking	No masking
Only masking	Masking	Masking

Training strategy The training strategy applies to the training dataset during fine-tuning.

No masking No masking is applied to the training dataset. This is what is ordinarily done in the literature.

Masking Masking is applied to every observation. The masking is uniformly sampled, at a masking rate between 0% and 100%.

Use 50/50 Half of the mini-batch using the *No masking* strategy and the other half use the *Masking* strategy.

Validation strategy The validation dataset is used to select the optional epoch. This is similar to early stopping, but rather than stopping immediately. The training continues, and the best epoch is chosen at the end of the training.

The validation strategy applies to the validation dataset during fine-tuning.

No masking No masking is applied to the validation dataset. This is what is ordinarily done in the literature.

Masking Masking is applied to every observation. The masking is uniformly sampled, at a masking rate between 0% and 100%.

Use both A copy of the validation dataset has the *No masking* strategy applied to it. Another copy of the validation dataset has the *Masking* strategy applied to it. As such, the validation dataset is twice as long, but it does not add additional observations or inforamtion.

H.1. Findings

We generally find that the choice of validation strategy when using the *Use 50/50* training strategy is not important. Interestingly, *Masking* for the validation dataset and *No masking* for the training dataset often works too.

However, because *Use 50/50* for training strategy and *Use both* for validation strategy, i.e. masked fine-tuning, work well in all cases and is theoretically sound, this is the approach we recommend and use throughout the paper.

H.2. All datasets aggregation

In Section 5.1 we also include an *All* "dataset". This is a simple arithmetic mean over all the performance of all 16 datasets. This is similar to how the GLUE benchmark (Wang et al., 2019b) works. To compute the confidence interval, a dataset-aggregation is done for each seed, such that the all-observation are i.i.d..

Because some seeds do not converge for some datasets, such as bAbI-2 and bAbI-3 (as mentioned in Section 5.1), those outliers and not included in the aggregation, also hyperparameter optimization will likely help. For complete transparency, we do include them in the statistics for the individual datasets and show all individual performances with a (+) symbol.

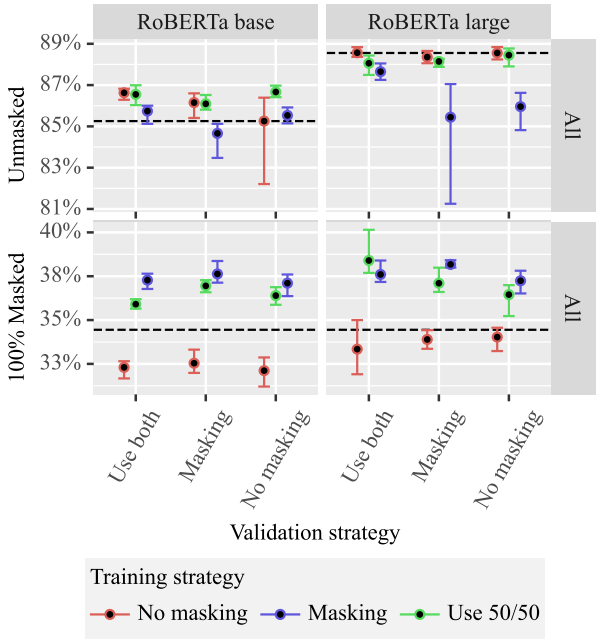


Figure 7. The all aggregation for the 100% masked performance and unmasked performance. The baseline (dashed line) for 100% masked performance is the class-majority baseline. Unmasked performance is when using no masking for both validation and training.

H.3. Test dataset

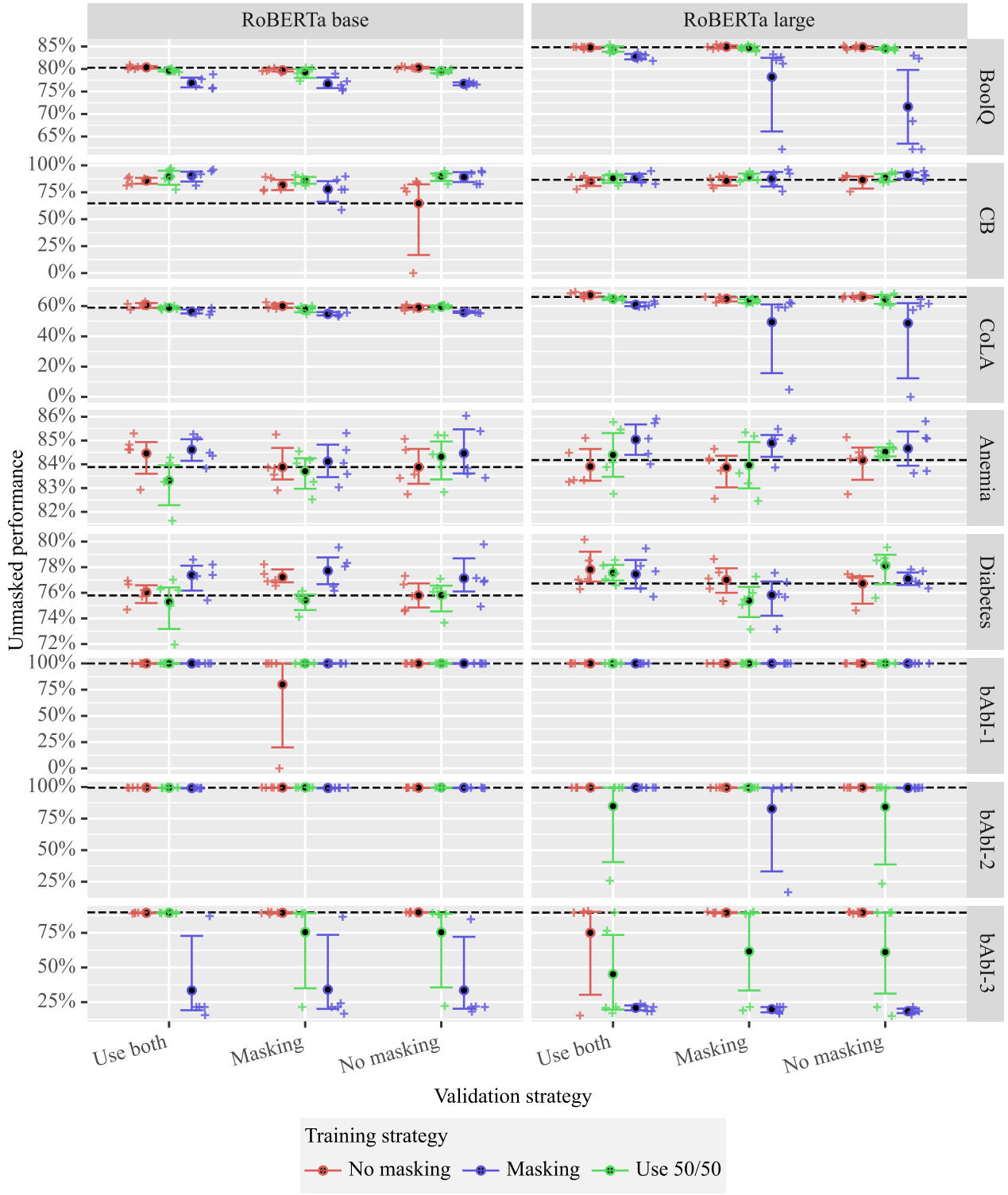


Figure 8. The unmasked performance for each validation and training strategy, using the test dataset. Not that *"No masking"* as a *training strategy* is not a valid option only a baseline, as it creates OOD issues. We find that the multi-task *training strategy "Use 50/50"* works best. This plot is **page-1**. Corresponding main results in Figure 2.

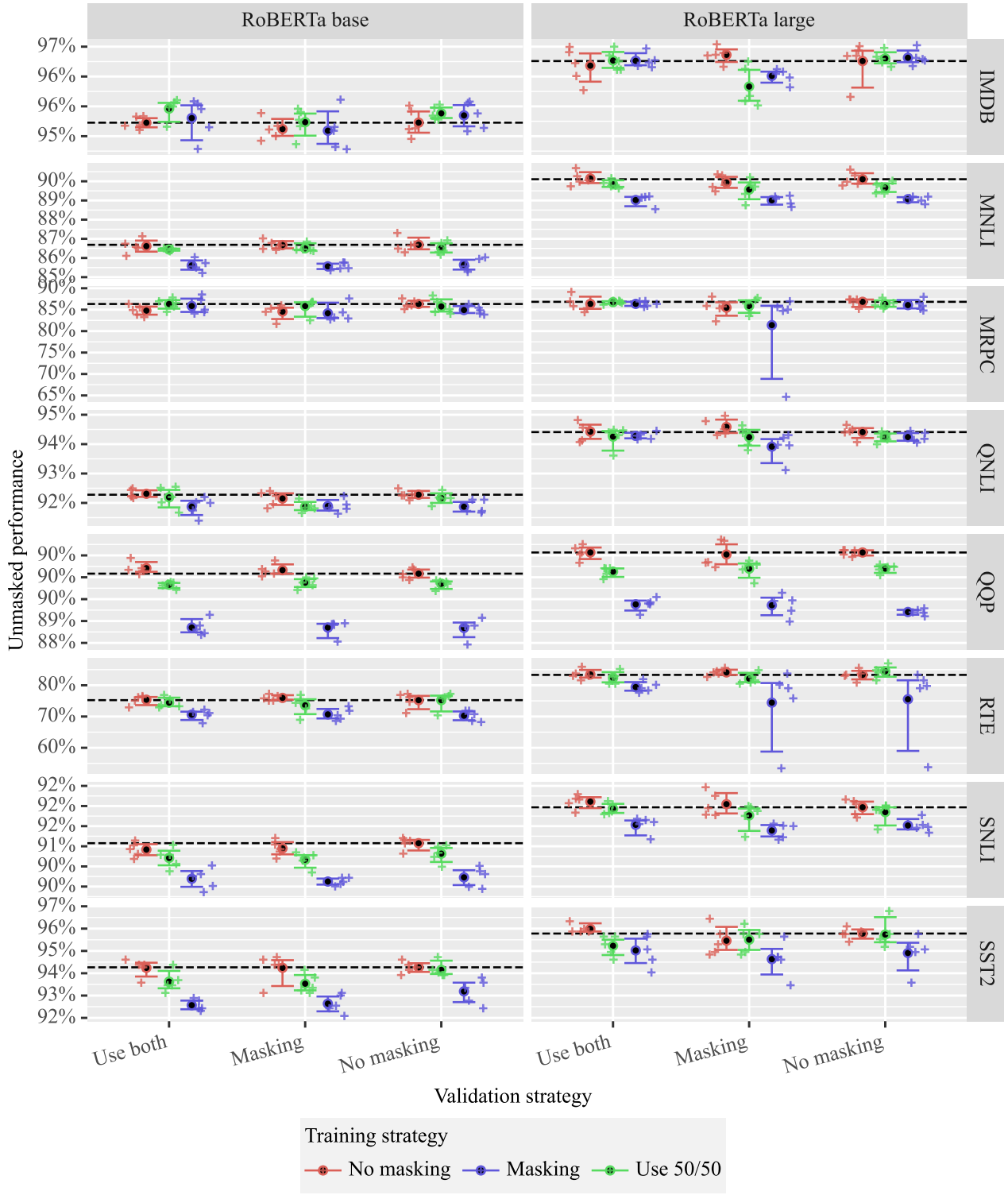


Figure 9. The unmasked performance for each validation and training strategy, using the test dataset. Not that *"No masking"* as a *training strategy* is not a valid option only a baseline, as it creates OOD issues. We find that the multi-task *training strategy "Use 50/50"* works best. This plot is **page-2**. Corresponding main results in Figure 2.

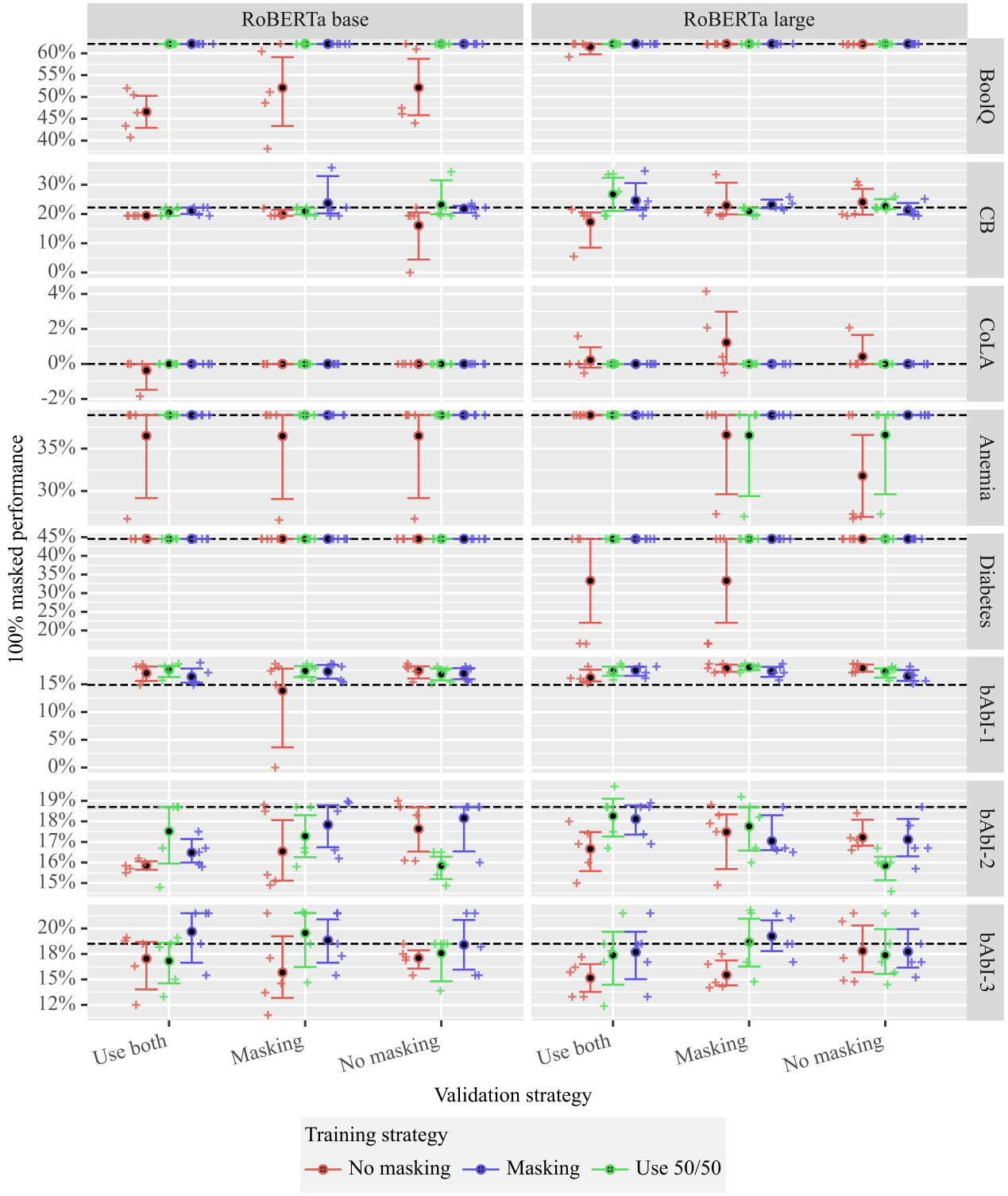


Figure 10. The 100% masked performance, using the test dataset. The dashed line represents the class-majority classifier baseline. Results show that masking during training (*"Masking"* or *"Use 50/50"*) is necessary. This plot is **page-1**. Corresponding main paper results in Figure 3.

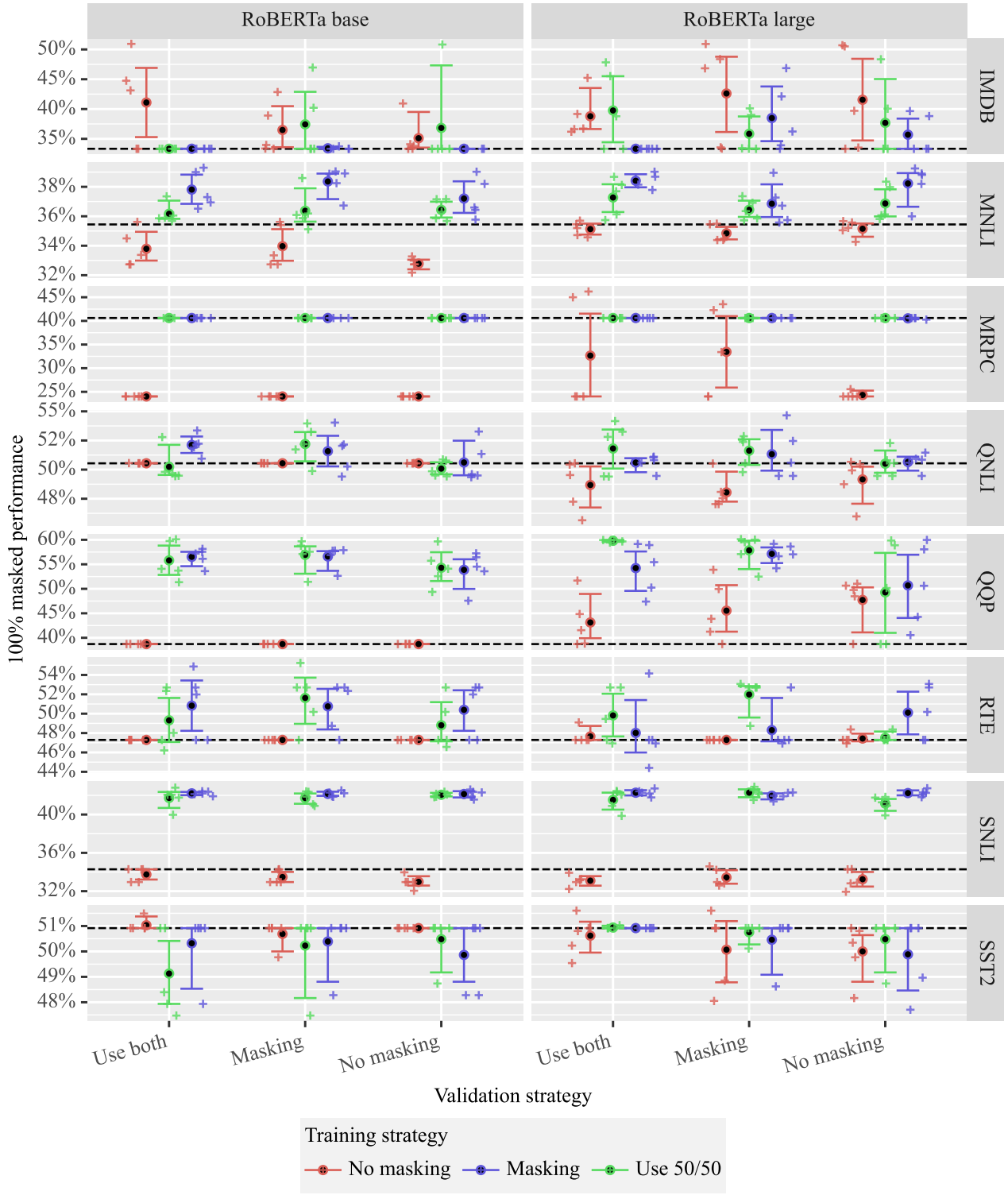


Figure 11. The 100% masked performance, using the test dataset. The dashed line represents the class-majority classifier baseline. Results show that masking during training (*"Masking"* or *"Use 50/50"*) is necessary. This plot is **page-2**. Corresponding main paper results in Figure 3.

H.4. Validation dataset

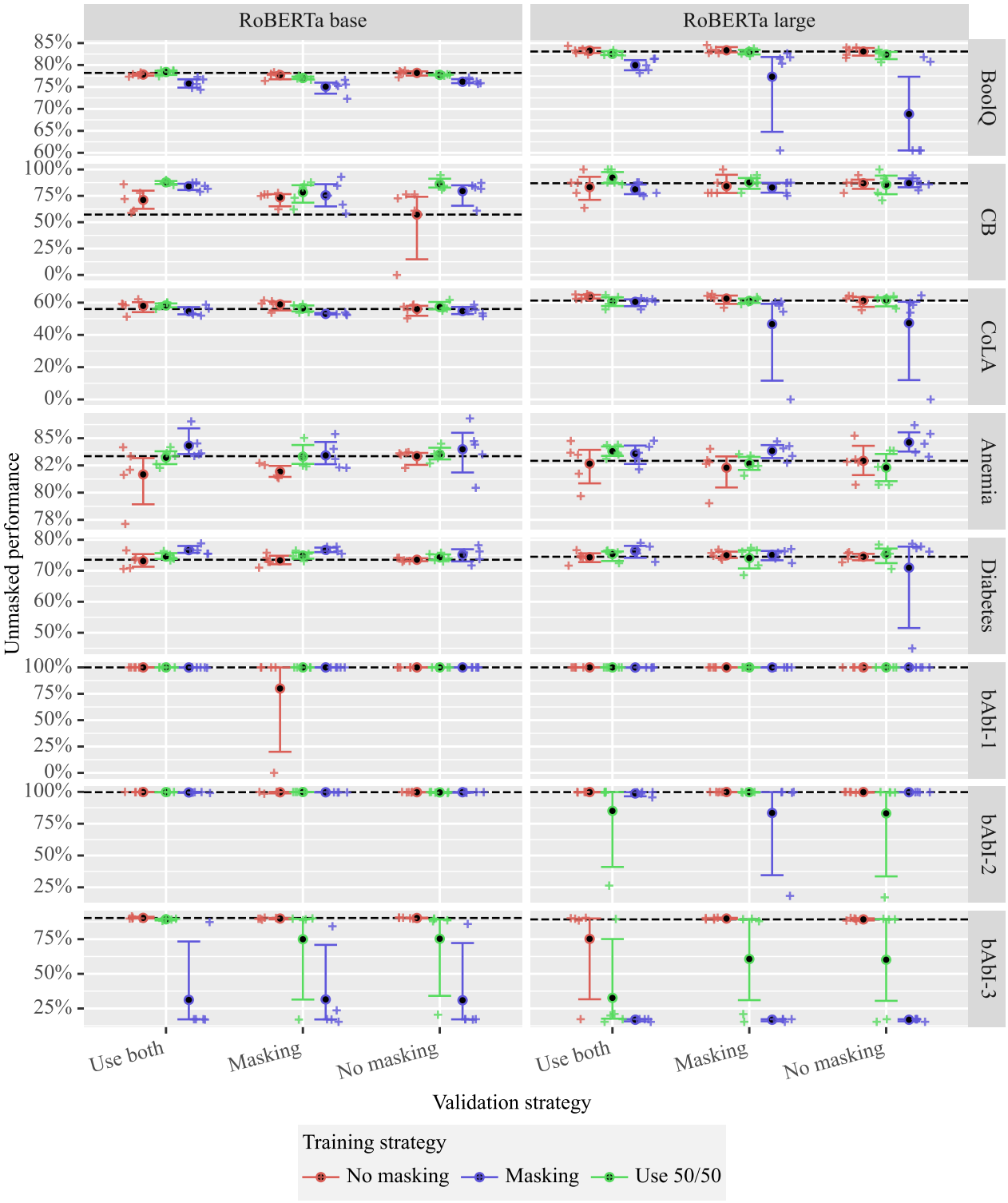


Figure 12. The unmasked performance for each validation and training strategy, using the validation dataset. Not that *"No masking"* as a *training strategy* is not a valid option only a baseline, as it creates OOD issues. We find that the multi-task *training strategy "Use 50/50"* works best. This plot is **page-1**.

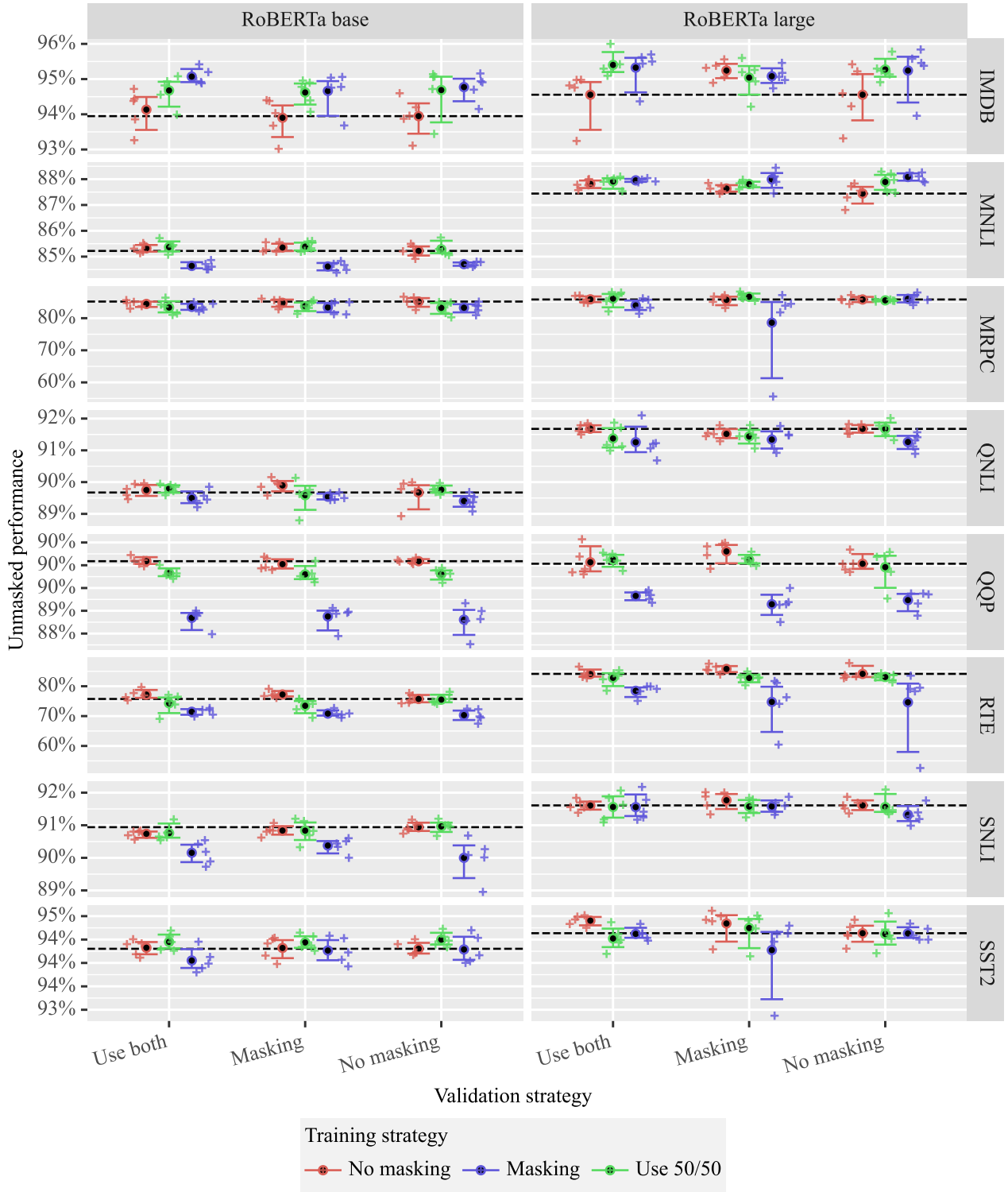


Figure 13. The unmasked performance for each validation and training strategy, using the validation dataset. Not that *"No masking"* as a *training strategy* is not a valid option only a baseline, as it creates OOD issues. We find that the multi-task *training strategy "Use 50/50"* works best. This plot is **page-2**.

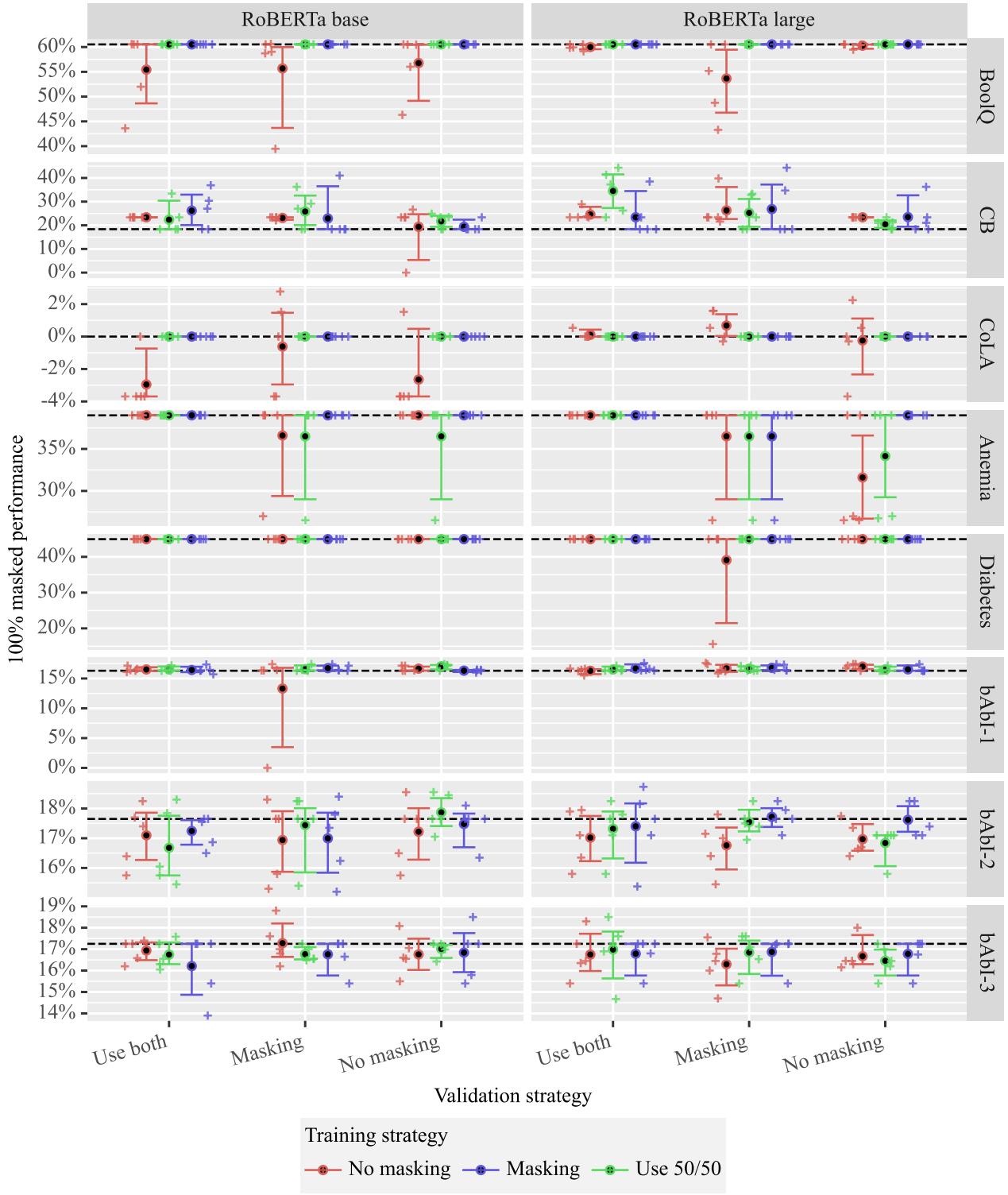


Figure 14. The 100% masked performance, using the validation dataset. The dashed line represents the class-majority classifier baseline. Results show that masking during training (*"Masking"* or *"Use 50/50"*) is necessary. This plot is **page-1**.

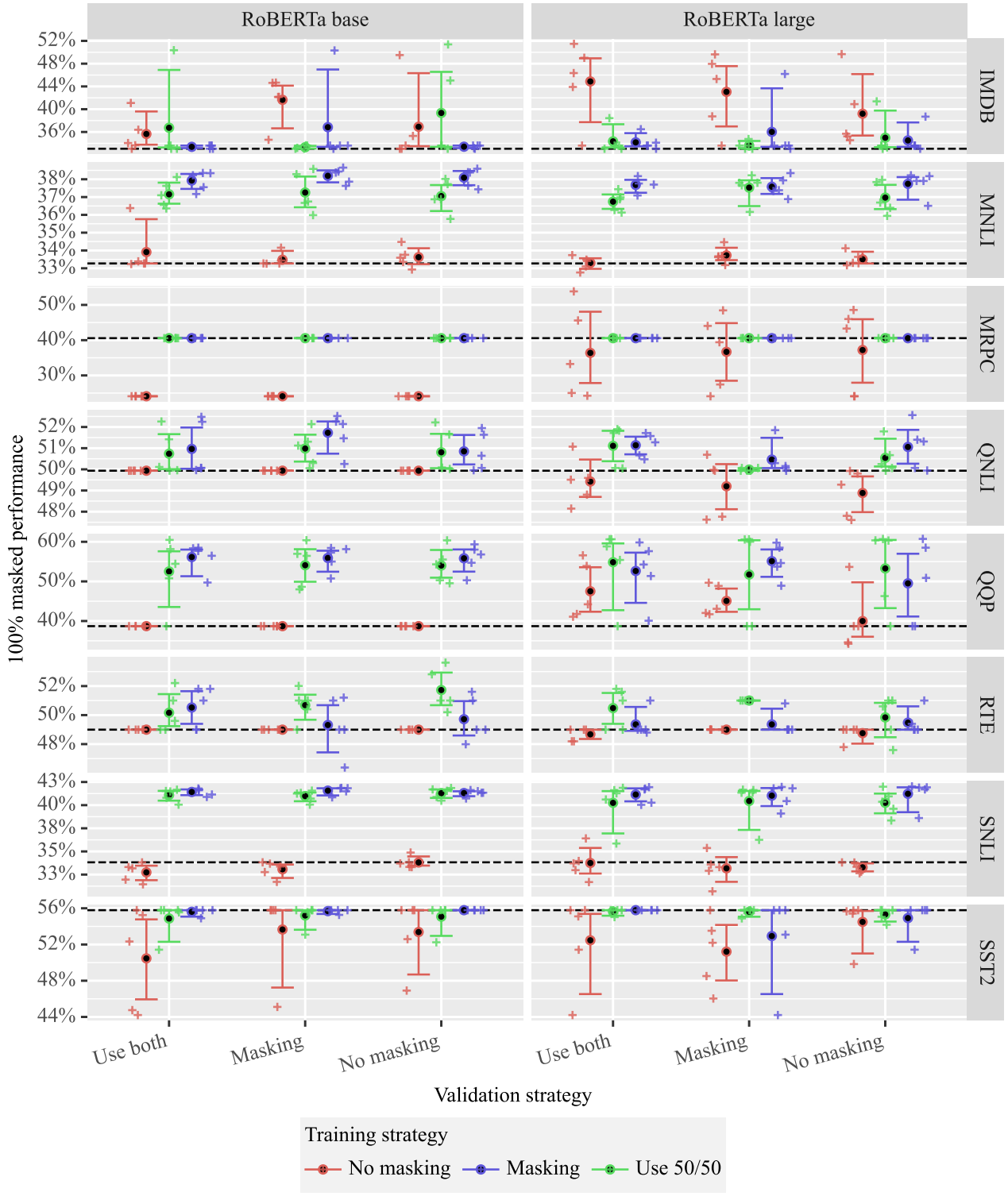


Figure 15. The 100% masked performance, using the validation dataset. The dashed line represents the class-majority classifier baseline. Results show that masking during training (*"Masking"* or *"Use 50/50"*) is necessary. This plot is **page-2**.

I. Convergence speed

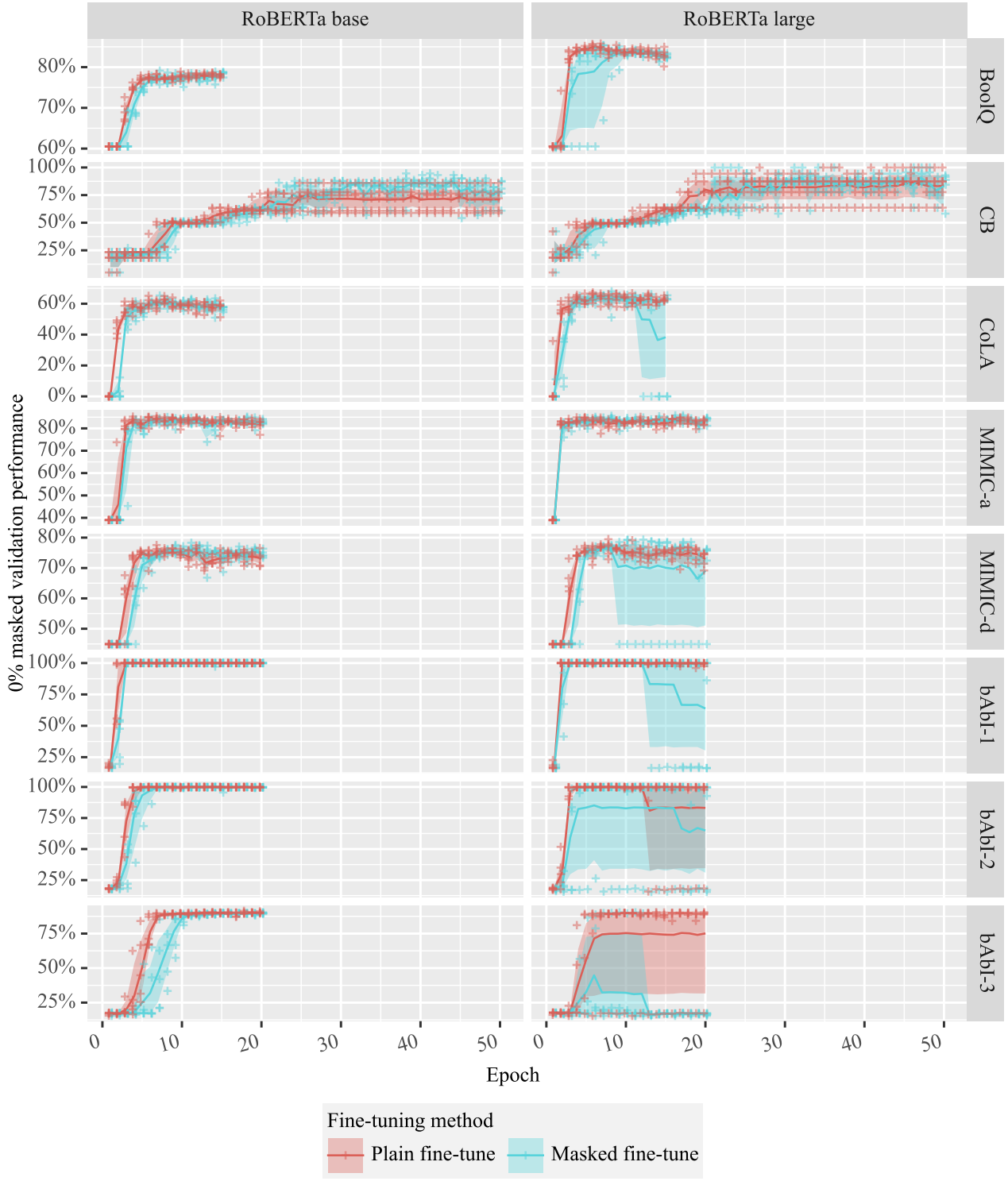


Figure 16. The validation performance for each epoch. Note that the max number of epochs vary depending on the dataset. This is only to limit the compute requirements when fine-tuning. The best epoch is selected by the "early-stopping" dataset, which has one copy with no masking and one copy with uniformly sampled masking ratios. This plot is **page-1**.

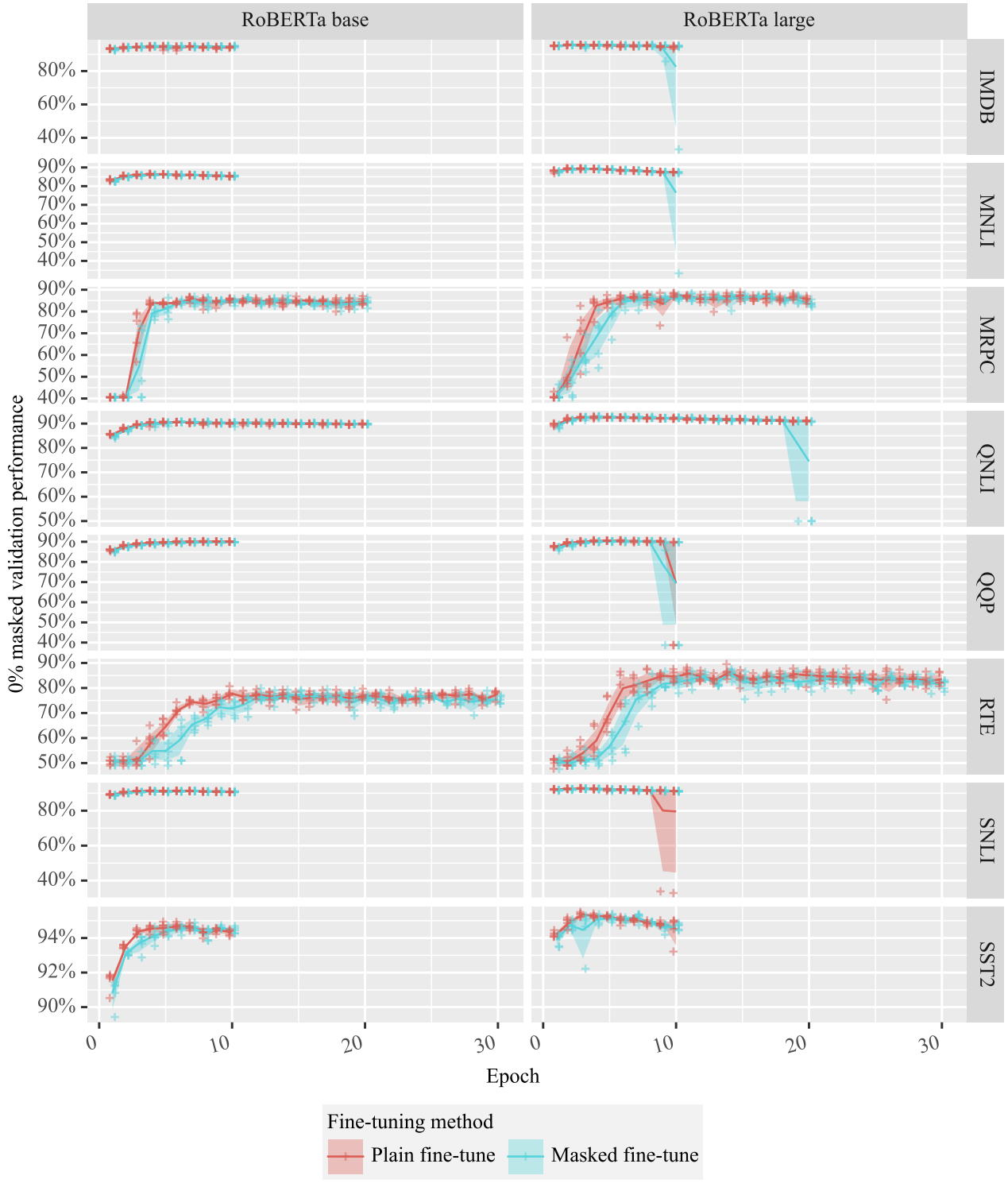


Figure 17. The validation performance for each epoch. Note that the max number of epochs vary depending on the dataset. This is only to limit the compute requirements when fine-tuning. The best epoch is selected by the "early-stopping" dataset, which has one copy with no masking and one copy with uniformly sampled masking ratios. This plot is **page-2**.

J. In-distribution validation

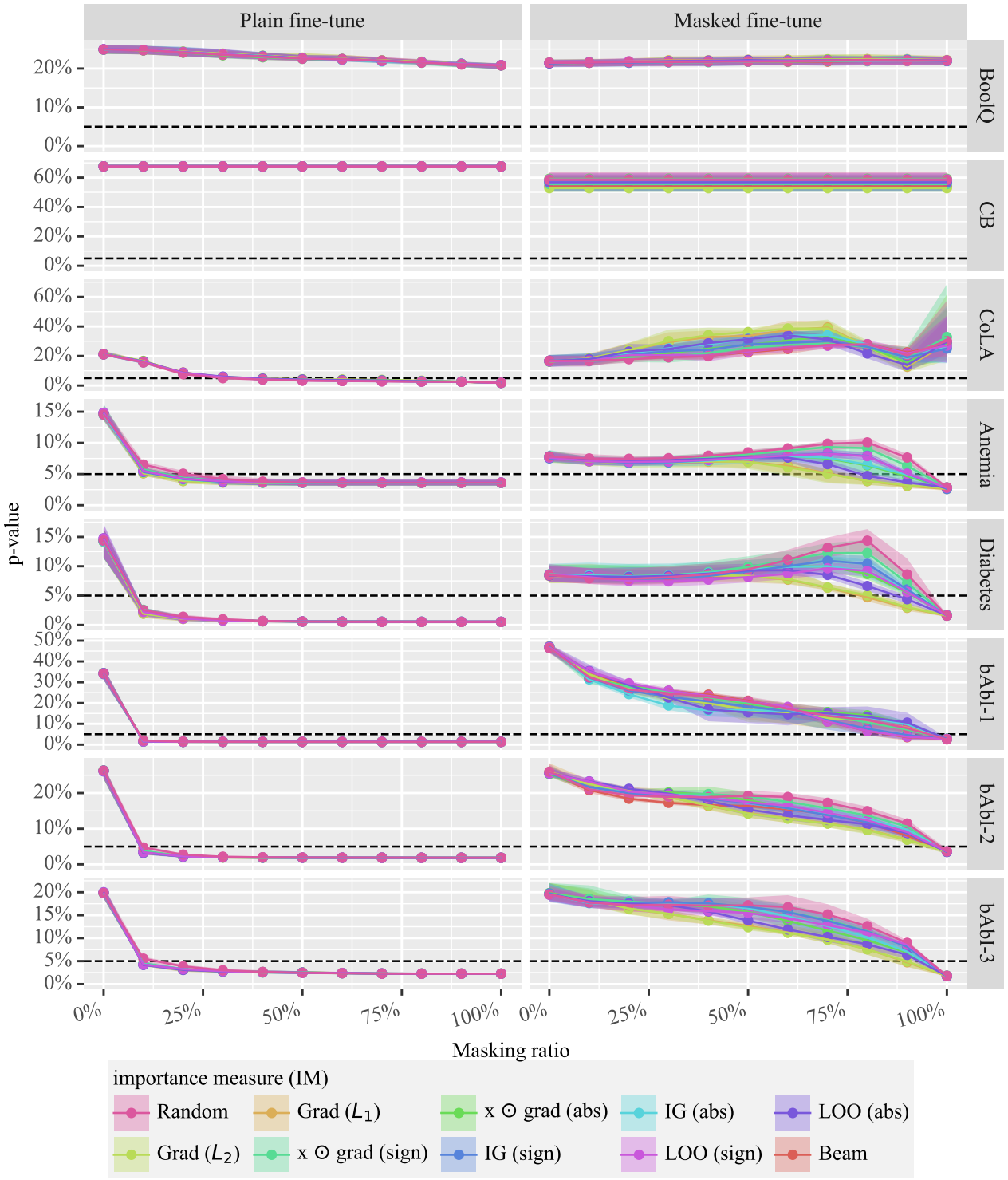


Figure 18. In-distribution p-values using MaSF, for **RoBERTa-base** with and without masked fine-tuning, **page-1**. The masked tokens are chosen according to an importance measure. P-values below the dashed line show out-of-distribution (OOD) results, given a 5% risk of a false positive. Results show that only when using masked fine-tuning, masked data is consistently not OOD. Corresponding main paper results are in Figure 4.

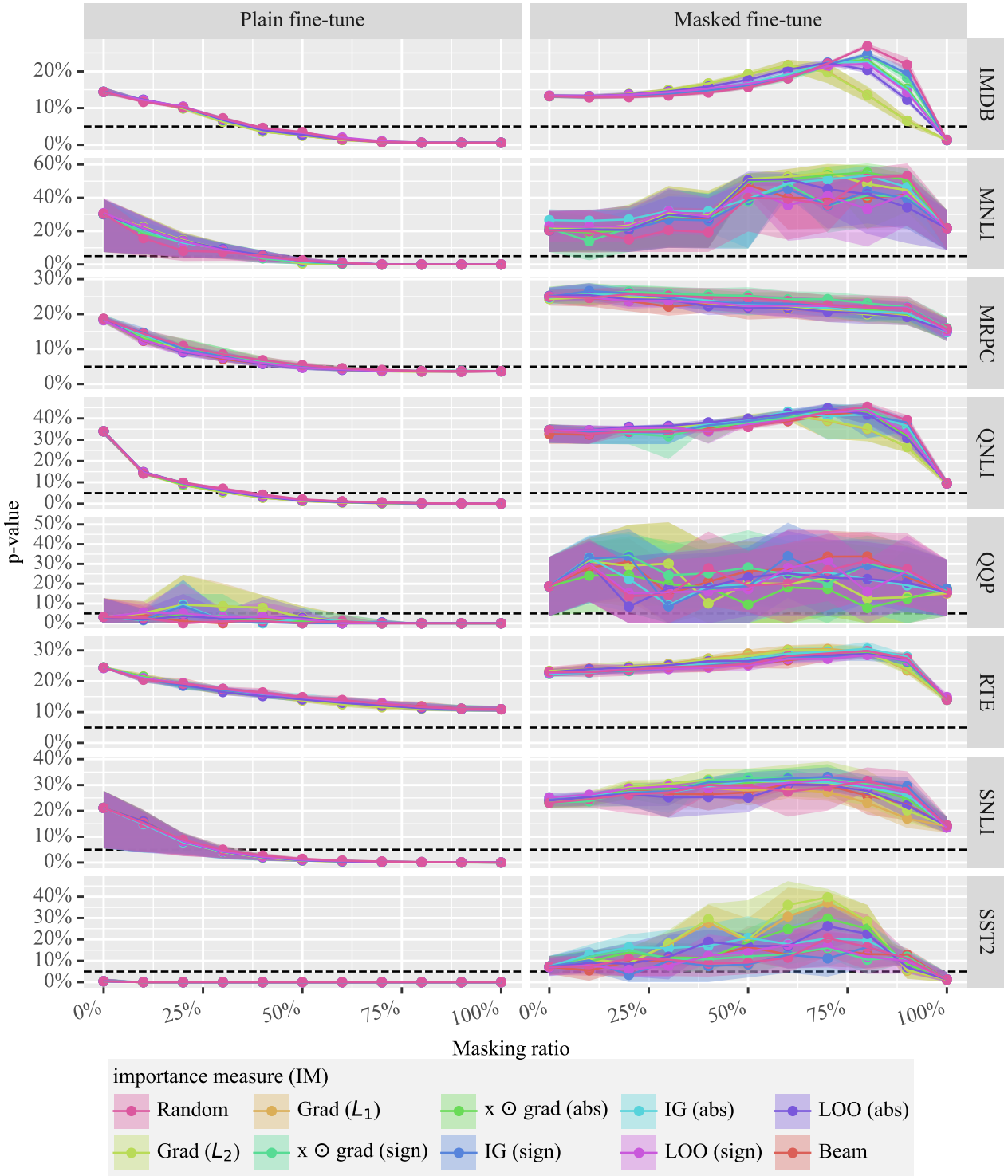


Figure 19. In-distribution p-values using MaSF, for **RoBERTa-base** with and without masked fine-tuning, **page-2**. The masked tokens are chosen according to an importance measure. P-values below the dashed line show out-of-distribution (OOD) results, given a 5% risk of a false positive. Results show that only when using masked fine-tuning, masked data is consistently not OOD. Corresponding main paper results are in Figure 4.

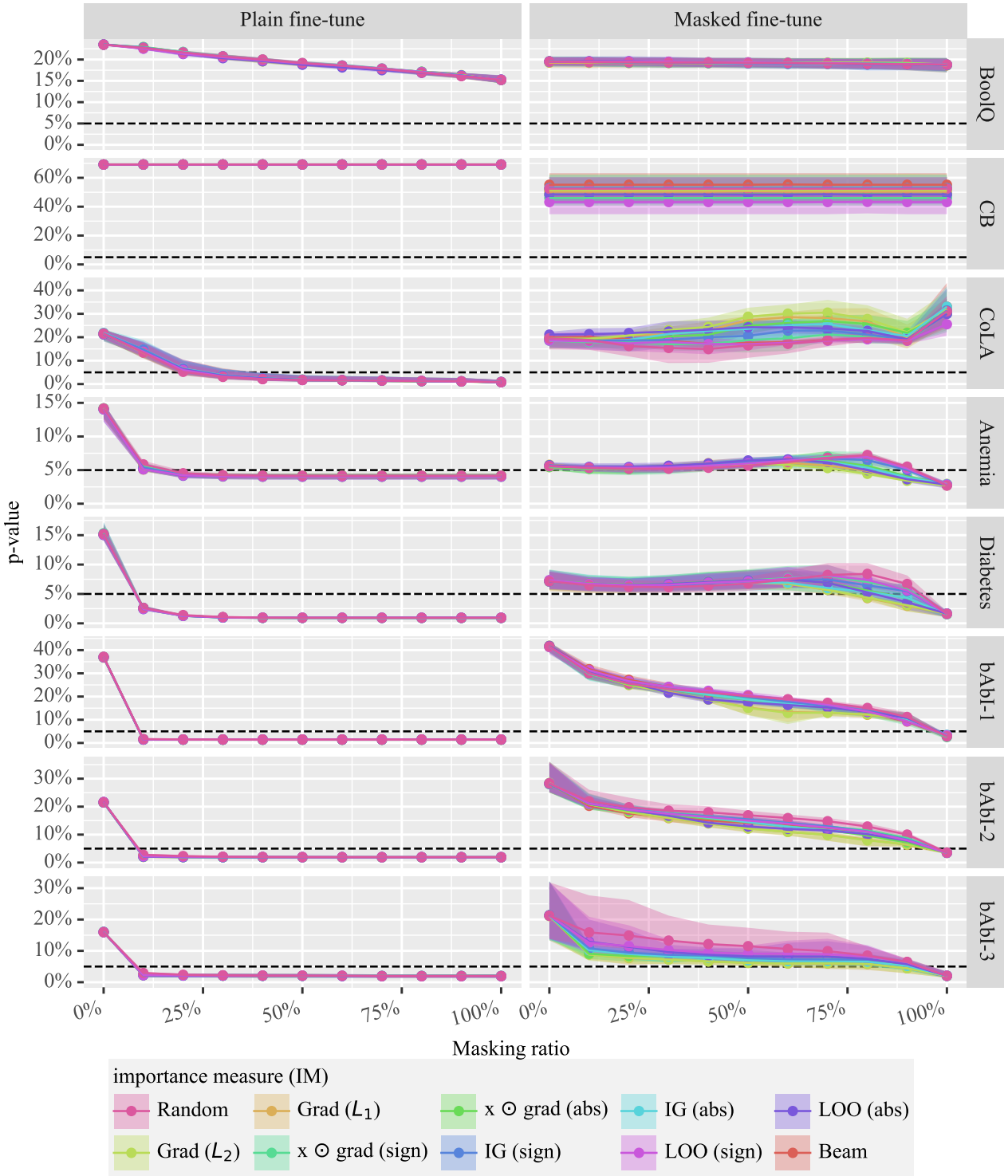


Figure 20. In-distribution p-values using MaSF, for **RoBERTa-large** with and without masked fine-tuning, **page-1**. The masked tokens are chosen according to an importance measure. P-values below the dashed line show out-of-distribution (OOD) results, given a 5% risk of a false positive. Results show that only when using masked fine-tuning, masked data is consistently not OOD. Corresponding main paper results are in Figure 4.

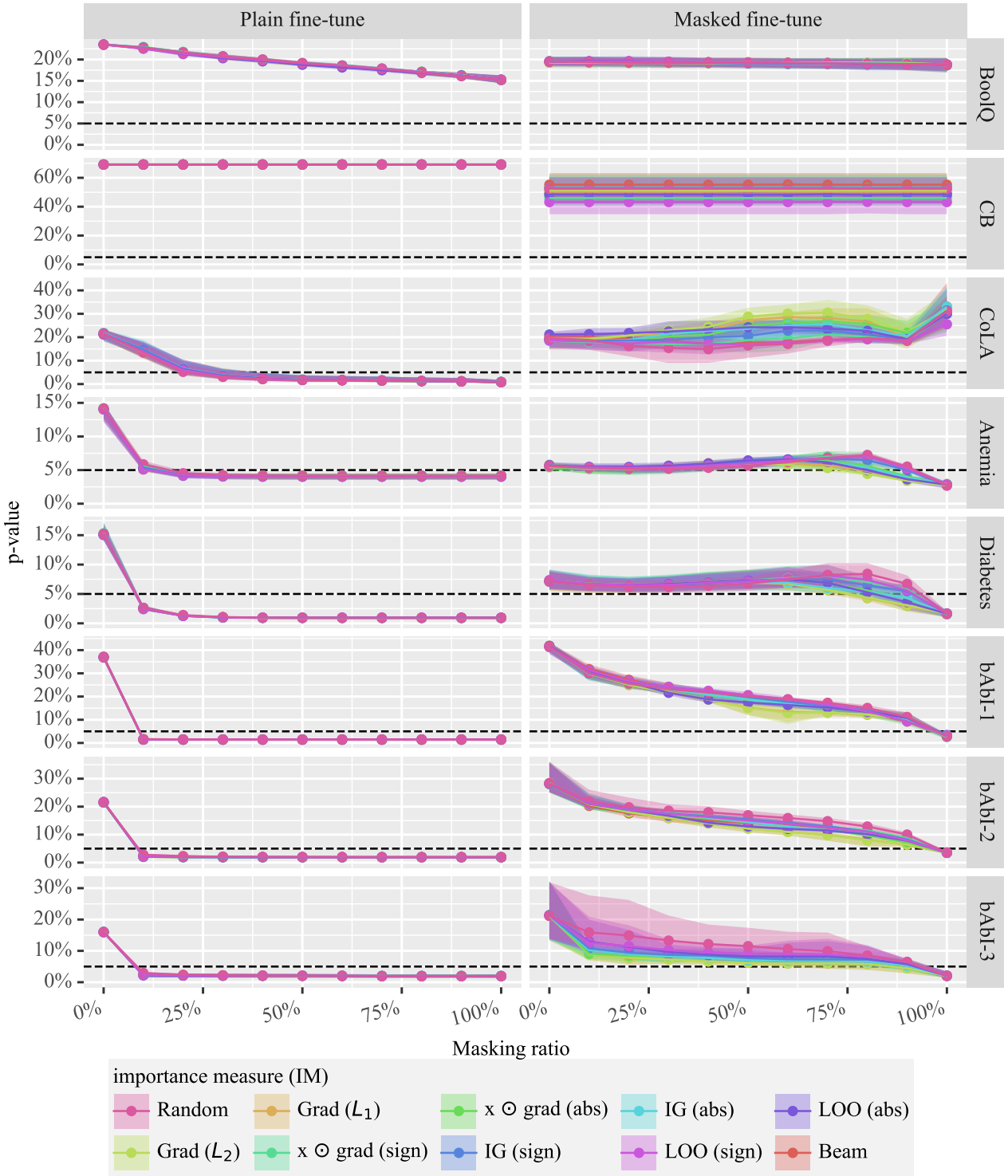


Figure 21. In-distribution p-values using MaSF, for **RoBERTa-large** with and without masked fine-tuning, **page-2**. The masked tokens are chosen according to an importance measure. P-values below the dashed line show out-of-distribution (OOD) results, given a 5% risk of a false positive. Results show that only when using masked fine-tuning, masked data is consistently not OOD. Corresponding main paper results are in Figure 4.

K. Faithfulness metrics

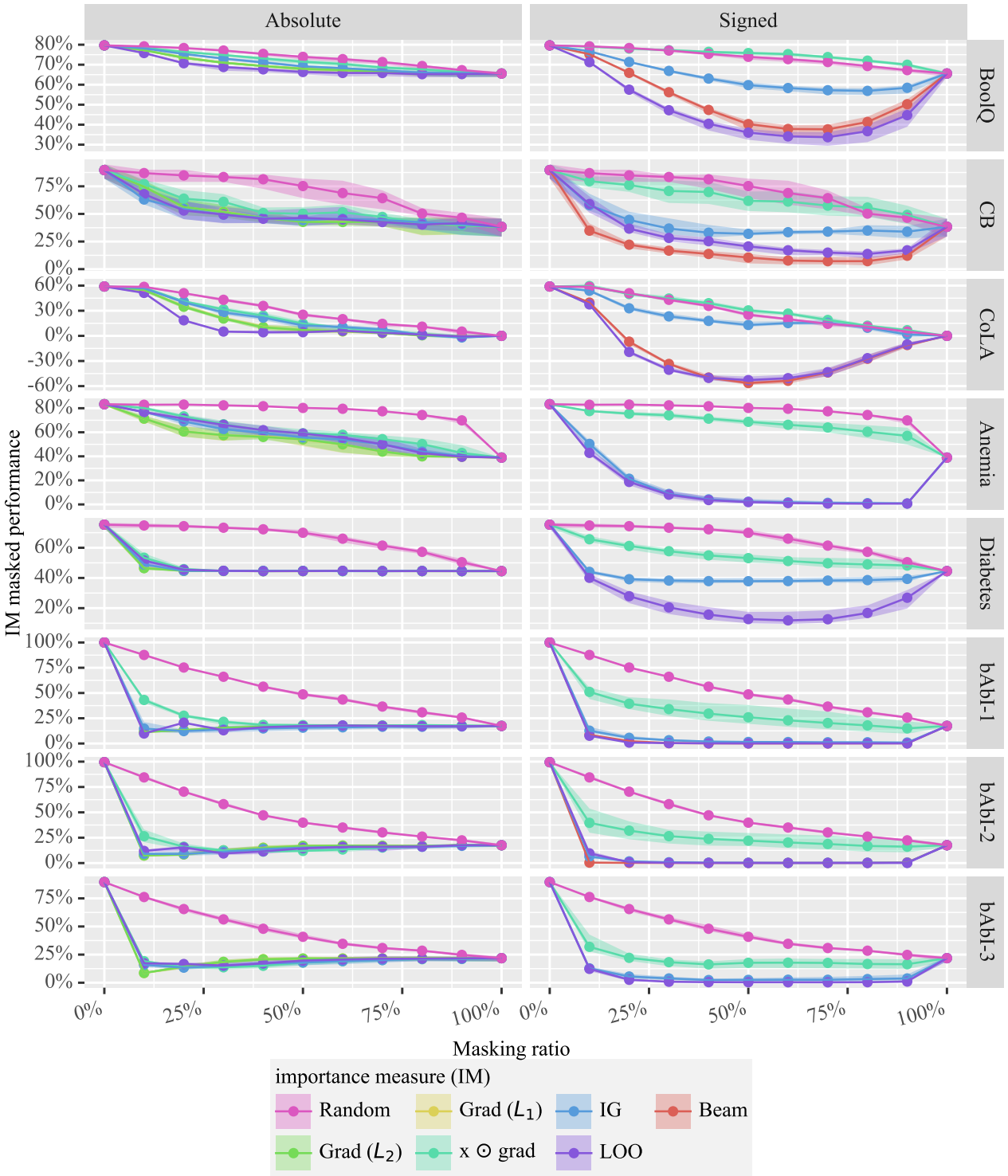


Figure 22. The performance given the masked datasets, where masking is done for the x% allegedly most important tokens according to the importance measure. If the performance for a given explanation is below the *"Random"* baseline, this shows faithfulness. Although, faithfulness is not an absolute concept, so more is better. This plot is **page-1** for **RoBERTa-base**. Corresponding main paper results in Section 5.3.

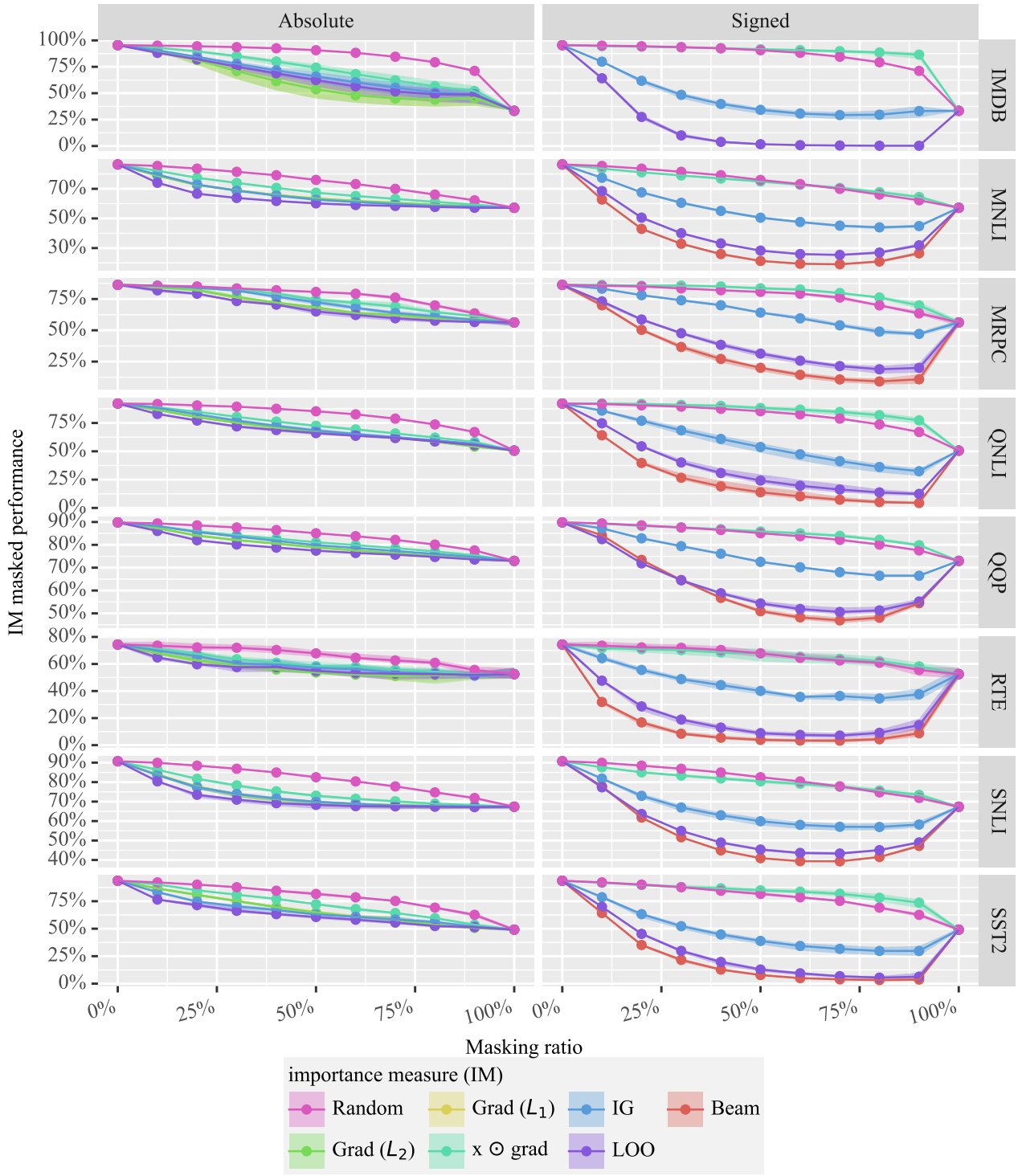


Figure 23. The performance given the masked datasets, where masking is done for the x% allegedly most important tokens according to the importance measure. If the performance for a given explanation is below the *"Random"* baseline, this shows faithfulness. Although, faithfulness is not an absolute concept, so more is better. This plot is **page-2** for **RoBERTa-base**. Corresponding main paper results in Section 5.3.

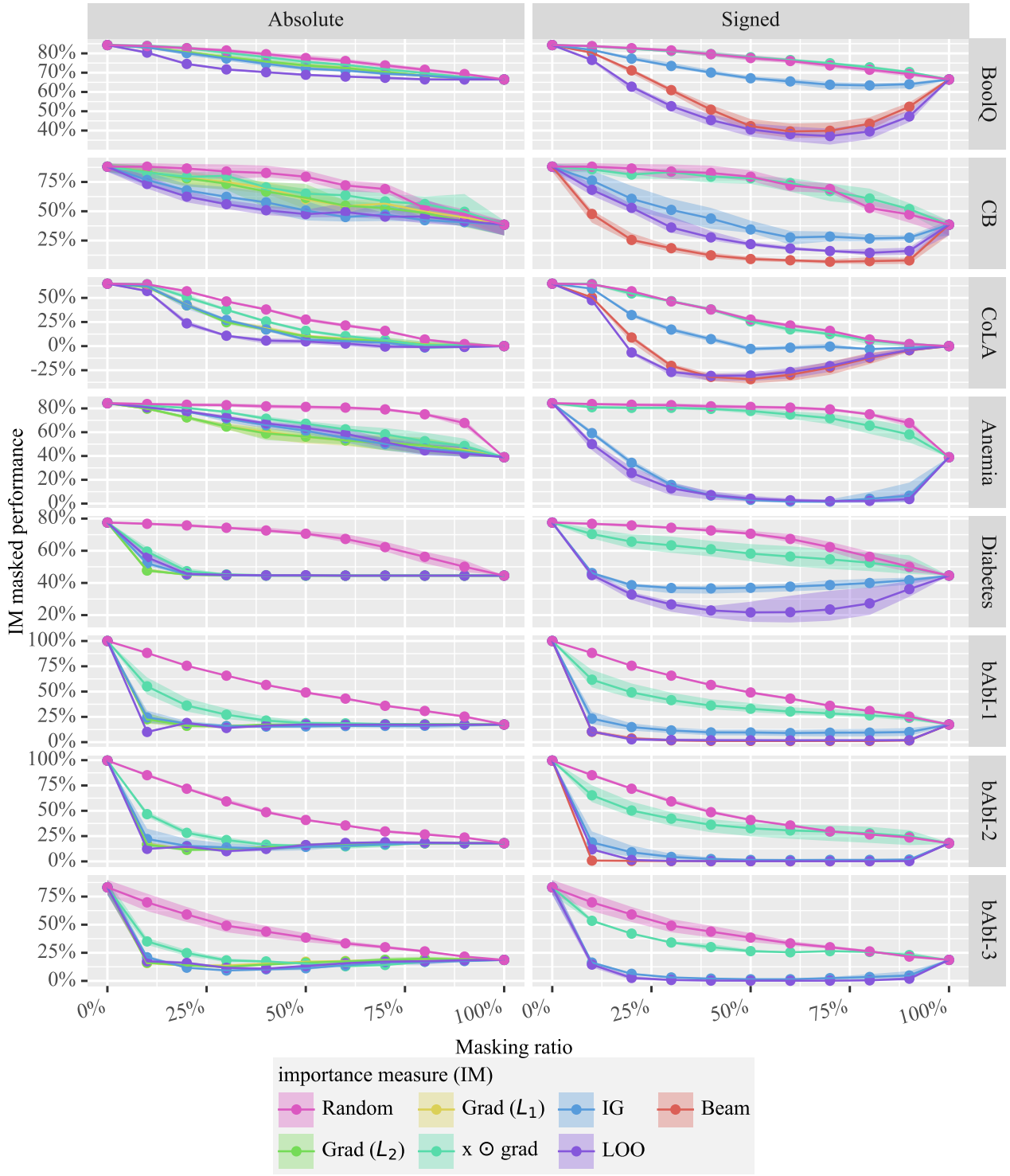


Figure 24. The performance given the masked datasets, where masking is done for the x% allegedly most important tokens according to the importance measure. If the performance for a given explanation is below the *"Random"* baseline, this shows faithfulness. Although, faithfulness is not an absolute concept, so more is better. This plot is **page-1** for **RoBERTa-large**. Corresponding main paper results in Section 5.3.

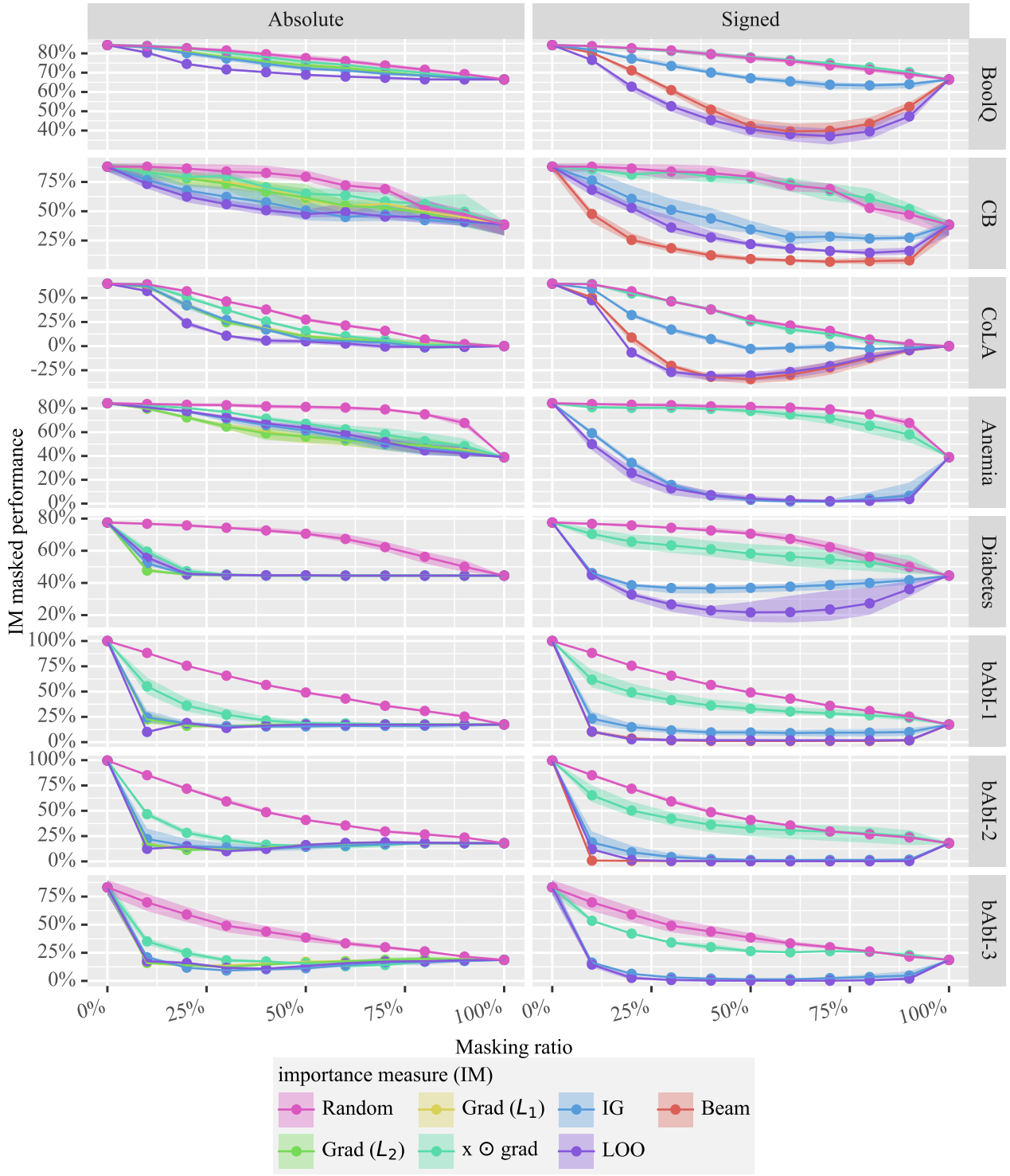


Figure 25. The performance given the masked datasets, where masking is done for the x% allegedly most important tokens according to the importance measure. If the performance for a given explanation is below the *"Random"* baseline, this shows faithfulness. Although, faithfulness is not an absolute concept, so more is better. This plot is **page-2** for **RoBERTa-large**. Corresponding main paper results in Section 5.3.

K.1. Relative Area between Curves (RACU)

Table 11. Faithfulness scores for **RoBERTa-base**. Shows Relative Area Between Curves (RACU) and the non-relative variant (ACU), defined by Madsen et al. (2022a). Also compares with Recursive-ROAR by Madsen et al. (2022a).

Dataset	IM	Faithfulness [%]		
		Our		R-ROAR
		ACU	RACU	RACU
bAbI-1	Grad (L_2)	$32.6^{+1.2}_{-1.1}$	$91.9^{+2.0}_{-3.2}$	$64.2^{+2.6}_{-2.6}$
	Grad (L_1)	$32.7^{+1.3}_{-0.9}$	$92.1^{+2.0}_{-3.1}$	–
	x ⊙ grad (sign)	$21.4^{+4.9}_{-7.3}$	$60.5^{+13.7}_{-20.7}$	–
	x ⊙ grad (abs)	$27.0^{+1.1}_{-1.4}$	$76.0^{+3.0}_{-1.6}$	$52.1^{+1.8}_{-3.7}$
	IG (sign)	$44.1^{+0.8}_{-0.8}$	$124.2^{+3.3}_{-4.5}$	–
	IG (abs)	$33.3^{+2.4}_{-2.4}$	$93.7^{+5.5}_{-4.1}$	$48.2^{+4.1}_{-5.7}$
	LOO (sign)	$46.1^{+0.7}_{-0.7}$	$129.9^{+3.4}_{-3.5}$	–
	LOO (abs)	$32.4^{+1.2}_{-1.2}$	$91.2^{+1.2}_{-0.7}$	–
	Beam	$45.9^{+0.7}_{-0.7}$	$129.2^{+3.1}_{-3.2}$	–
bAbI-2	Grad (L_2)	$28.5^{+0.8}_{-0.8}$	$96.3^{+6.8}_{-2.8}$	$57.8^{+2.0}_{-2.0}$
	Grad (L_1)	$28.5^{+0.9}_{-0.8}$	$96.3^{+6.8}_{-2.7}$	–
	x ⊙ grad (sign)	$19.7^{+6.6}_{-8.1}$	$65.7^{+24.1}_{-26.3}$	–
	x ⊙ grad (abs)	$27.3^{+1.7}_{-1.5}$	$92.0^{+2.5}_{-3.1}$	$48.1^{+3.2}_{-3.5}$
	IG (sign)	$40.3^{+0.9}_{-0.8}$	$136.3^{+4.4}_{-6.4}$	–
	IG (abs)	$29.1^{+1.0}_{-1.3}$	$98.3^{+5.5}_{-3.9}$	$42.0^{+3.8}_{-4.8}$
	LOO (sign)	$40.2^{+1.2}_{-0.8}$	$136.0^{+4.1}_{-6.5}$	–
	LOO (abs)	$28.5^{+0.9}_{-1.4}$	$96.3^{+9.2}_{-3.6}$	–
	Beam	$41.1^{+1.0}_{-0.7}$	$139.2^{+5.0}_{-7.3}$	–
bAbI-3	Grad (L_2)	$23.5^{+0.9}_{-1.5}$	$97.3^{+3.5}_{-4.2}$	$34.0^{+14.6}_{-15.1}$
	Grad (L_1)	$23.5^{+0.9}_{-1.3}$	$97.2^{+3.6}_{-4.3}$	–
	x ⊙ grad (sign)	$23.1^{+2.4}_{-2.5}$	$96.7^{+21.8}_{-14.9}$	–
	x ⊙ grad (abs)	$24.4^{+0.7}_{-1.2}$	$101.2^{+4.3}_{-5.4}$	$22.4^{+15.9}_{-12.4}$
	IG (sign)	$36.6^{+1.3}_{-1.4}$	$152.5^{+14.7}_{-17.4}$	–
	IG (abs)	$24.4^{+1.0}_{-1.3}$	$100.9^{+4.0}_{-3.9}$	$-27.9^{+18.0}_{-49.1}$
	LOO (sign)	$38.7^{+0.9}_{-0.7}$	$160.5^{+13.8}_{-13.2}$	–
	LOO (abs)	$23.4^{+0.7}_{-1.1}$	$97.0^{+4.5}_{-4.6}$	–
	Beam	–	–	–
BoolQ	Grad (L_2)	$4.1^{+0.2}_{-0.3}$	$50.1^{+3.6}_{-3.2}$	–
	Grad (L_1)	$4.1^{+0.2}_{-0.3}$	$50.1^{+4.0}_{-3.1}$	–
	x ⊙ grad (sign)	$-1.3^{+0.7}_{-0.4}$	$-16.0^{+8.2}_{-4.5}$	–
	x ⊙ grad (abs)	$1.7^{+0.2}_{-0.3}$	$21.3^{+2.7}_{-3.3}$	–
	IG (sign)	$9.6^{+0.6}_{-0.6}$	$118.3^{+10.3}_{-7.9}$	–
	IG (abs)	$3.0^{+0.2}_{-0.4}$	$37.8^{+4.7}_{-5.3}$	–
	LOO (sign)	$26.3^{+1.9}_{-1.4}$	$323.1^{+14.4}_{-15.2}$	–
	LOO (abs)	$5.3^{+0.3}_{-0.2}$	$65.2^{+3.3}_{-2.2}$	–
	Beam	$21.2^{+1.4}_{-1.1}$	$261.4^{+10.7}_{-14.5}$	–
CB	Grad (L_2)	$20.3^{+1.3}_{-1.3}$	$64.7^{+9.9}_{-9.9}$	–
	Grad (L_1)	$20.2^{+1.4}_{-1.3}$	$64.7^{+10.7}_{-10.7}$	–
	x ⊙ grad (sign)	$6.0^{+4.0}_{-3.6}$	$17.8^{+7.0}_{-11.0}$	–
	x ⊙ grad (abs)	$15.9^{+2.2}_{-3.2}$	$52.0^{+15.7}_{-14.3}$	–
	IG (sign)	$30.1^{+5.7}_{-2.7}$	$97.6^{+22.7}_{-28.3}$	–
	IG (abs)	$21.1^{+2.3}_{-2.3}$	$67.5^{+10.2}_{-13.2}$	–
	LOO (sign)	$40.9^{+5.8}_{-5.5}$	$131.4^{+26.7}_{-27.3}$	–
	LOO (abs)	$21.0^{+1.7}_{-2.4}$	$67.4^{+12.8}_{-12.8}$	–
	Beam	$51.0^{+6.4}_{-6.8}$	$163.5^{+32.7}_{-33.8}$	–
CoLA	Grad (L_2)	$12.7^{+0.9}_{-1.1}$	$43.4^{+4.0}_{-3.2}$	–
	Grad (L_1)	$12.7^{+0.8}_{-1.2}$	$43.2^{+3.7}_{-3.4}$	–
	x ⊙ grad (sign)	$-2.4^{+1.9}_{-1.1}$	$-8.3^{+6.0}_{-3.7}$	–
	x ⊙ grad (abs)	$8.0^{+1.0}_{-0.8}$	$27.1^{+2.9}_{-2.9}$	–
	IG (sign)	$8.1^{+1.1}_{-0.8}$	$27.5^{+2.8}_{-2.6}$	–
	IG (abs)	$8.7^{+1.7}_{-0.8}$	$29.8^{+6.6}_{-3.0}$	–
	LOO (sign)	$52.0^{+1.7}_{-2.9}$	$177.2^{+7.2}_{-7.8}$	–
	LOO (abs)	$17.0^{+0.4}_{-0.4}$	$57.9^{+1.2}_{-0.9}$	–
	Beam	$50.6^{+1.3}_{-2.3}$	$172.7^{+5.9}_{-7.2}$	–
Anemia	Grad (L_2)	$23.8^{+0.6}_{-0.5}$	$62.1^{+1.4}_{-1.7}$	$18.2^{+11.8}_{-13.8}$
	Grad (L_1)	$23.8^{+0.6}_{-0.6}$	$62.2^{+1.4}_{-2.1}$	–
	x ⊙ grad (sign)	$9.7^{+2.6}_{-2.5}$	$25.1^{+6.5}_{-6.2}$	–
	x ⊙ grad (abs)	$16.6^{+1.3}_{-1.3}$	$43.2^{+3.0}_{-3.7}$	$8.8^{+22.7}_{-22.8}$
	IG (sign)	$62.0^{+1.6}_{-1.5}$	$161.8^{+2.7}_{-2.9}$	–
	IG (abs)	$20.0^{+0.9}_{-1.6}$	$52.1^{+2.5}_{-4.5}$	$12.5^{+11.3}_{-7.0}$
	LOO (sign)	$63.3^{+1.4}_{-1.6}$	$165.2^{+2.9}_{-3.2}$	–
	LOO (abs)	$18.9^{+1.0}_{-1.3}$	$49.2^{+2.5}_{-3.8}$	–
	Beam	–	–	–
Diabetes	Grad (L_2)	$19.7^{+1.2}_{-0.7}$	$91.8^{+0.6}_{-0.9}$	$57.9^{+14.4}_{-19.8}$
	Grad (L_1)	$19.6^{+1.0}_{-0.7}$	$91.6^{+0.5}_{-0.9}$	–
	x ⊙ grad (sign)	$10.9^{+1.9}_{-1.3}$	$51.1^{+9.1}_{-7.8}$	–
	x ⊙ grad (abs)	$18.8^{+1.3}_{-0.7}$	$87.9^{+1.4}_{-2.0}$	$53.4^{+23.2}_{-29.3}$
	IG (sign)	$24.8^{+1.5}_{-2.1}$	$115.8^{+2.8}_{-7.8}$	–
	IG (abs)	$19.4^{+1.0}_{-0.6}$	$90.5^{+0.6}_{-1.2}$	$26.1^{+12.0}_{-25.1}$
	LOO (sign)	$41.5^{+2.3}_{-5.9}$	$193.4^{+8.9}_{-17.4}$	–
	LOO (abs)	$19.1^{+1.1}_{-0.6}$	$89.0^{+0.4}_{-0.6}$	–
	Beam	–	–	–

Dataset	IM	Faithfulness [%]		
		Our		R-ROAR
		ACU	RACU	RACU
MRPC	Grad (L_2)	$7.8^{+0.5}_{-0.6}$	$36.8^{+4.3}_{-6.1}$	–
	Grad (L_1)	$7.9^{+0.5}_{-0.7}$	$37.4^{+4.2}_{-6.3}$	–
	x ⊙ grad (sign)	$-2.9^{+0.5}_{-0.6}$	$-13.7^{+2.9}_{-2.9}$	–
	x ⊙ grad (abs)	$3.3^{+0.6}_{-0.8}$	$15.6^{+3.4}_{-4.9}$	–
	IG (sign)	$12.7^{+0.8}_{-1.0}$	$59.9^{+6.2}_{-10.6}$	–
	IG (abs)	$5.4^{+1.0}_{-0.6}$	$25.4^{+7.0}_{-3.9}$	–
	LOO (sign)	$37.1^{+2.4}_{-1.3}$	$174.4^{+12.4}_{-18.9}$	–
	LOO (abs)	$9.9^{+0.8}_{-0.9}$	$47.0^{+5.5}_{-8.9}$	–
	Beam	$45.6^{+2.6}_{-1.5}$	$214.8^{+13.3}_{-25.1}$	–
RTE	Grad (L_2)	$9.4^{+0.8}_{-1.0}$	$73.9^{+23.6}_{-19.3}$	–
	Grad (L_1)	$9.4^{+0.8}_{-1.0}$	$73.5^{+23.3}_{-19.1}$	–
	x ⊙ grad (sign)	$0.1^{+2.3}_{-1.7}$	$-3.5^{+17.3}_{-13.9}$	–
	x ⊙ grad (abs)	$5.7^{+0.5}_{-0.5}$	$44.3^{+11.8}_{-8.5}$	–
	IG (sign)	$20.2^{+1.3}_{-2.2}$	$156.6^{+45.5}_{-28.4}$	–
	IG (abs)	$7.3^{+1.6}_{-0.9}$	$55.9^{+14.8}_{-13.7}$	–
	LOO (sign)	$44.4^{+1.6}_{-1.9}$	$344.8^{+92.4}_{-75.5}$	–
	LOO (abs)	$9.4^{+0.9}_{-0.6}$	$73.2^{+21.1}_{-15.2}$	–
	Beam	$51.3^{+2.0}_{-1.7}$	$399.7^{+109.9}_{-89.9}$	–
SST2	Grad (L_2)	$12.2^{+0.6}_{-0.7}$	$40.4^{+3.0}_{-1.7}$	$26.1^{+1.6}_{-2.2}$
	Grad (L_1)	$12.1^{+0.7}_{-0.7}$	$40.3^{+3.3}_{-1.8}$	–
	x ⊙ grad (sign)	$-3.7^{+1.5}_{-1.6}$	$-12.2^{+4.5}_{-6.0}$	–
	x ⊙ grad (abs)	$7.1^{+0.2}_{-0.2}$	$23.5^{+1.9}_{-1.1}$	$18.6^{+4.1}_{-4.6}$
	IG (sign)	$31.8^{+2.8}_{-2.2}$	$105.6^{+7.7}_{-7.7}$	–
	IG (abs)	$13.7^{+0.8}_{-0.8}$	$45.3^{+4.1}_{-2.8}$	$32.9^{+1.8}_{-1.5}$
	LOO (sign)	$51.6^{+1.4}_{-0.9}$	$171.3^{+5.8}_{-6.2}$	–
	LOO (abs)	$16.6^{+1.2}_{-1.0}$	$54.9^{+2.1}_{-1.5}$	–
	Beam	$56.4^{+0.5}_{-0.7}$	$187.3^{+8.1}_{-7.1}$	–
SNLI	Grad (L_2)	$8.9^{+1.0}_{-0.4}$	$62.2^{+2.7}_{-1.7}$	$50.7^{+1.1}_{-0.8}$
	Grad (L_1)	$8.9^{+1.0}_{-0.4}$	$62.2^{+2.6}_{-1.8}$	–
	x ⊙ grad (sign)	$1.3^{+0.7}_{-0.8}$	$9.2^{+5.2}_{-4.9}$	–
	x ⊙ grad (abs)	$6.4^{+0.5}_{-0.3}$	$44.8^{+1.6}_{-1.6}$	$41.0^{+0.4}_{-0.5}$
	IG (sign)	$16.3^{+1.3}_{-1.3}$	$113.5^{+7.6}_{-6.4}$	–
	IG (abs)	$8.9^{+0.8}_{-0.5}$	$62.3^{+2.5}_{-2.1}$	$56.7^{+1.0}_{-1.1}$
	LOO (sign)	$26.6^{+0.4}_{-0.4}$	$186.6^{+9.1}_{-12.4}$	–
	LOO (abs)	$10.5^{+0.8}_{-0.5}$	$73.6^{+2.0}_{-1.9}$	–
	Beam	$29.3^{+0.2}_{-0.3}$	$205.2^{+8.1}_{-8.7}$	–
IMDB	Grad (L_2)	$24.9^{+5.9}_{-4.3}$	$47.8^{+10.9}_{-8.5}$	$25.4^{+3.1}_{-2.0}$
	Grad (L_1)	$24.9^{+5.7}_{-4.4}$	$47.8^{+10.9}_{-8.3}$	–
	x ⊙ grad (sign)	$-3.2^{+1.3}_{-1.0}$	$-6.2^{+2.5}_{-1.9}$	–
	x ⊙ grad (abs)	$12.9^{+4.5}_{-2.6}$	$24.7^{+8.7}_{-5.1}$	$16.9^{+1.1}_{-3.0}$
	IG (sign)	$40.3^{+3.3}_{-1.9}$	$77.3^{+6.3}_{-3.6}$	–
	IG (abs)	$18.4^{+4.2}_{-2.8}$	$35.4^{+8.0}_{-5.9}$	$35.1^{+2.4}_{-1.7}$
	LOO (sign)	$68.0^{+0.9}_{-0.9}$	$130.7^{+1.3}_{-1.5}$	–
	LOO (abs)	$20.7^{+4.2}_{-3.2}$	$39.8^{+9.5}_{-6.0}$	–
	Beam	–	–	–
MNLI	Grad (L_2)	$8.8^{+0.3}_{-0.2}$	$49.6^{+1.1}_{-1.2}$	–
	Grad (L_1)	$8.8^{+0.3}_{-0.2}$	$49.7^{+1.1}_{-1.2}$	–
	x ⊙ grad (sign)	$0.6^{+0.6}_{-0.7}$	$3.5^{+3.7}_{-3.8}$	–
	x ⊙ grad (abs)	$5.7^{+0.2}_{-0.2}$	$32.0^{+0.7}_{-0.8}$	–
	IG (sign)	$18.5^{+0.6}_{-0.7}$	$104.3^{+5.9}_{-5.6}$	–
	IG (abs)	$9.1^{+0.2}_{-0.2}$	$51.3^{+1.0}_{-2.0}$	–
	LOO (sign)	$34.7^{+0.7}_{-0.5}$	$195.7^{+3.2}_{-2.8}$	–
	LOO (abs)	$11.8^{+0.3}_{-0.3}$	$66.7^{+0.9}_{-1.1}$	–
	Beam	$40.5^{+0.8}_{-0.4}$	$228.9^{+4.5}_{-4.5}$	–
QNLI	Grad (L_2)	$12.7^{+0.6}_{-0.7}$	$40.8^{+2.4}_{-2.0}$	–
	Grad (L_1)	$12.7^{+0.6}_{-0.7}$	$40.8^{+2.3}_{-1.9}$	–
	x ⊙ grad (sign)	$-3.7^{+1.9}_{-1.1}$	$-11.9^{+6.0}_{-3.4}$	–
	x ⊙ grad (abs)	$8.8^{+0.6}_{-0.6}$	$28.3^{+2.1}_{-1.7}$	–
	IG (sign)	$24.4^{+4.3}_{-2.0}$	$78.0^{+10.4}_{-7.4}$	–
	IG (abs)	$11.4^{+0.7}_{-0.6}$	$36.5^{+2.0}_{-2.6}$	–
	LOO (sign)	$46.1^{+1.8}_{-2.6}$	$147.7^{+4.7}_{-8.7}$	–
	LOO (abs)	$14.0^{+0.6}_{-0.4}$	$44.9^{+3.2}_{-1.3}$	–
	Beam	$55.7^{+1.5}_{-2.2}$	$178.4^{+4.8}_{-8.6}$	–
QQP	Grad (L_2)	$4.5^{+0.1}_{-0.1}$	$40.0^{+0.4}_{-0.8}$	–
	Grad (L_1)	$4.5^{+0.1}_{-0.1}$	$40.0^{+0.4}_{-0.7}$	–
	x ⊙ grad (sign)	$-0.8^{+0.3}_{-0.6}$	$-7.3^{+3.0}_{-5.5}$	–
	x ⊙ grad (abs)	$2.8^{+0.2}_{-0.2}$	$24.5^{+1.4}_{-1.5}$	–
	IG (sign)	$9.1^{+0.4}_{-0.3}$	$81.3^{+2.6}_{-2.3}$	–
	IG (abs)	$3.5^{+0.2}_{-0.2}$	$31.1^{+2.3}_{-1.3}$	–
	LOO (sign)	$22.0^{+0.6}_{-0.9}$	$195.4^{+6.1}_{-7.7}$	–
	LOO (abs)	$5.6^{+0.1}_{-0.2}$	$49.6^{+0.8}_{-1.8}$	–
	Beam	$23.3^{+0.5}_{-0.6}$	$207.3^{+6.6}_{-5.7}$	–

Table 12. Faithfulness scores for **RoBERTa-large**. Shows Relative Area Between Curves (RACU) and the non-relative variant (ACU), defined by Madsen et al. (2022a). Note that Madsen et al. (2022a) does not report results for Recursive-ROAR with RoBERTa-large.

Dataset	IM	Faithfulness [%]		
		Our		R-ROAR
		ACU	RACU	RACU
bAbI-1	Grad (L_2)	$31.1^{+1.0}_{-1.5}$	$87.6^{+1.6}_{-2.3}$	–
	Grad (L_1)	$31.1^{+1.1}_{-1.6}$	$87.7^{+1.6}_{-2.5}$	–
	x ⊙ grad (sign)	$13.9^{+3.1}_{-6.5}$	$39.1^{+8.8}_{-16.0}$	–
	x ⊙ grad (abs)	$24.0^{+2.6}_{-3.2}$	$67.5^{+5.7}_{-8.3}$	–
	IG (sign)	$36.3^{+2.9}_{-3.1}$	$102.4^{+7.7}_{-6.9}$	–
	IG (abs)	$31.6^{+1.3}_{-1.5}$	$88.9^{+2.5}_{-2.7}$	–
	LOO (sign)	$44.4^{+0.9}_{-1.0}$	$125.2^{+1.7}_{-1.7}$	–
	LOO (abs)	$32.4^{+0.4}_{-0.6}$	$91.4^{+0.4}_{-0.9}$	–
	Beam	$44.7^{+0.8}_{-1.0}$	$126.0^{+1.1}_{-1.1}$	–
bAbI-2	Grad (L_2)	$28.2^{+0.9}_{-1.5}$	$94.0^{+6.3}_{-3.4}$	–
	Grad (L_1)	$28.0^{+0.9}_{-1.4}$	$93.5^{+5.3}_{-2.8}$	–
	x ⊙ grad (sign)	$8.2^{+7.0}_{-4.9}$	$26.9^{+20.4}_{-19.4}$	–
	x ⊙ grad (abs)	$22.6^{+1.5}_{-1.5}$	$75.5^{+7.7}_{-5.2}$	–
	IG (sign)	$37.9^{+1.3}_{-1.3}$	$126.6^{+5.8}_{-9.7}$	–
	IG (abs)	$27.4^{+1.7}_{-1.7}$	$91.7^{+6.8}_{-9.5}$	–
	LOO (sign)	$40.6^{+1.9}_{-0.7}$	$135.4^{+4.0}_{-4.0}$	–
	LOO (abs)	$28.1^{+0.9}_{-0.9}$	$93.9^{+2.8}_{-1.8}$	–
	Beam	$41.7^{+1.8}_{-0.7}$	$139.2^{+4.3}_{-4.3}$	–
bAbI-3	Grad (L_2)	$22.4^{+3.8}_{-3.8}$	$94.7^{+0.2}_{-0.2}$	–
	Grad (L_1)	$22.2^{+3.6}_{-3.6}$	$94.1^{+0.4}_{-0.4}$	–
	x ⊙ grad (sign)	$8.5^{+3.5}_{-3.5}$	$34.4^{+9.0}_{-9.0}$	–
	x ⊙ grad (abs)	$19.9^{+4.6}_{-4.6}$	$83.2^{+5.5}_{-5.5}$	–
	IG (sign)	$33.0^{+3.8}_{-3.8}$	$141.3^{+7.7}_{-7.7}$	–
	IG (abs)	$24.3^{+3.7}_{-3.7}$	$103.2^{+1.5}_{-1.5}$	–
	LOO (sign)	$35.0^{+3.0}_{-3.0}$	$150.5^{+12.4}_{-12.4}$	–
	LOO (abs)	$23.3^{+4.2}_{-4.2}$	$98.7^{+1.2}_{-1.2}$	–
	Beam	–	–	–
BoolQ	Grad (L_2)	$2.6^{+0.1}_{-0.3}$	$24.8^{+1.7}_{-1.9}$	–
	Grad (L_1)	$2.7^{+0.2}_{-0.3}$	$25.3^{+2.3}_{-1.9}$	–
	x ⊙ grad (sign)	$-0.4^{+0.2}_{-0.1}$	$-3.6^{+1.6}_{-1.5}$	–
	x ⊙ grad (abs)	$1.2^{+0.3}_{-0.3}$	$10.8^{+2.0}_{-2.0}$	–
	IG (sign)	$6.9^{+1.2}_{-0.8}$	$65.6^{+15.7}_{-6.9}$	–
	IG (abs)	$3.2^{+0.4}_{-0.8}$	$30.3^{+4.6}_{-5.0}$	–
	LOO (sign)	$25.6^{+1.7}_{-2.7}$	$242.9^{+39.2}_{-26.4}$	–
	LOO (abs)	$6.2^{+0.6}_{-0.5}$	$58.0^{+3.3}_{-1.2}$	–
	Beam	$21.5^{+1.3}_{-2.4}$	$204.0^{+30.8}_{-21.7}$	–
CB	Grad (L_2)	$10.1^{+4.4}_{-4.2}$	$30.8^{+15.2}_{-12.9}$	–
	Grad (L_1)	$8.9^{+3.9}_{-4.3}$	$27.3^{+12.6}_{-13.2}$	–
	x ⊙ grad (sign)	$-0.1^{+2.7}_{-3.3}$	$0.6^{+9.5}_{-8.1}$	–
	x ⊙ grad (abs)	$5.6^{+1.8}_{-2.9}$	$17.4^{+6.1}_{-9.8}$	–
	IG (sign)	$28.5^{+3.9}_{-3.4}$	$85.3^{+19.0}_{-14.0}$	–
	IG (abs)	$17.1^{+3.2}_{-2.2}$	$51.3^{+10.5}_{-9.0}$	–
	LOO (sign)	$39.0^{+2.9}_{-2.4}$	$116.1^{+9.5}_{-15.8}$	–
	LOO (abs)	$19.0^{+4.0}_{-2.4}$	$56.6^{+8.8}_{-12.8}$	–
	Beam	$51.8^{+3.0}_{-3.0}$	$154.6^{+14.0}_{-22.6}$	–
CoLA	Grad (L_2)	$10.9^{+0.9}_{-0.7}$	$35.0^{+2.9}_{-2.5}$	–
	Grad (L_1)	$10.4^{+0.9}_{-0.6}$	$33.3^{+3.1}_{-2.2}$	–
	x ⊙ grad (sign)	$1.2^{+0.2}_{-0.3}$	$3.8^{+0.5}_{-0.9}$	–
	x ⊙ grad (abs)	$6.6^{+0.9}_{-0.9}$	$21.3^{+2.9}_{-2.9}$	–
	IG (sign)	$17.3^{+1.1}_{-1.1}$	$55.4^{+3.6}_{-3.2}$	–
	IG (abs)	$11.7^{+0.4}_{-0.3}$	$37.5^{+1.0}_{-0.9}$	–
	LOO (sign)	$38.9^{+4.1}_{-2.0}$	$124.9^{+11.4}_{-6.0}$	–
	LOO (abs)	$17.6^{+0.9}_{-1.3}$	$56.7^{+3.2}_{-3.4}$	–
	Beam	$37.5^{+4.7}_{-2.7}$	$120.3^{+13.4}_{-8.2}$	–
Anemia	Grad (L_2)	$18.6^{+1.7}_{-1.5}$	$47.9^{+3.7}_{-3.6}$	–
	Grad (L_1)	$18.4^{+1.8}_{-1.8}$	$47.4^{+4.3}_{-4.3}$	–
	x ⊙ grad (sign)	$4.6^{+1.8}_{-1.6}$	$11.9^{+4.6}_{-4.2}$	–
	x ⊙ grad (abs)	$11.6^{+2.1}_{-2.1}$	$29.8^{+5.5}_{-5.5}$	–
	IG (sign)	$58.2^{+2.9}_{-4.0}$	$150.1^{+4.2}_{-5.0}$	–
	IG (abs)	$16.4^{+1.2}_{-1.9}$	$42.3^{+2.4}_{-5.7}$	–
	LOO (sign)	$60.5^{+3.3}_{-2.8}$	$156.0^{+7.4}_{-5.7}$	–
	LOO (abs)	$15.6^{+1.6}_{-1.7}$	$40.2^{+3.5}_{-3.7}$	–
	Beam	–	–	–
Diabetes	Grad (L_2)	$20.1^{+1.8}_{-1.1}$	$90.7^{+0.5}_{-0.4}$	–
	Grad (L_1)	$20.1^{+1.9}_{-1.1}$	$90.7^{+0.6}_{-0.4}$	–
	x ⊙ grad (sign)	$7.4^{+3.1}_{-4.0}$	$34.2^{+16.6}_{-19.0}$	–
	x ⊙ grad (abs)	$18.6^{+1.5}_{-0.9}$	$84.2^{+0.6}_{-0.8}$	–
	IG (sign)	$25.3^{+4.0}_{-2.3}$	$113.5^{+7.9}_{-5.1}$	–
	IG (abs)	$19.6^{+1.4}_{-1.0}$	$88.7^{+0.5}_{-1.2}$	–
	LOO (sign)	$34.8^{+6.7}_{-4.9}$	$156.0^{+12.2}_{-18.3}$	–
	LOO (abs)	$19.2^{+1.4}_{-1.0}$	$86.6^{+0.6}_{-0.9}$	–
	Beam	–	–	–

Dataset	IM	Faithfulness [%]		
		Our		R-ROAR
		ACU	RACU	RACU
MRPC	Grad (L_2)	$6.6^{+2.2}_{-1.1}$	$22.9^{+3.1}_{-4.3}$	–
	Grad (L_1)	$6.6^{+1.9}_{-1.1}$	$22.7^{+1.5}_{-3.7}$	–
	x ⊙ grad (sign)	$-0.9^{+1.0}_{-1.1}$	$-3.4^{+3.2}_{-2.2}$	–
	x ⊙ grad (abs)	$4.4^{+1.2}_{-1.0}$	$15.7^{+2.6}_{-4.8}$	–
	IG (sign)	$15.7^{+1.6}_{-2.0}$	$55.9^{+9.5}_{-8.8}$	–
	IG (abs)	$8.1^{+1.3}_{-1.1}$	$28.8^{+9.2}_{-3.7}$	–
	LOO (sign)	$29.9^{+0.8}_{-1.3}$	$110.0^{+30.9}_{-24.2}$	–
	LOO (abs)	$10.2^{+1.6}_{-1.6}$	$36.1^{+8.8}_{-5.2}$	–
	Beam	$40.1^{+1.4}_{-2.3}$	$146.9^{+41.5}_{-30.6}$	–
RTE	Grad (L_2)	$7.1^{+0.8}_{-1.4}$	$38.8^{+5.9}_{-5.9}$	–
	Grad (L_1)	$7.4^{+0.9}_{-1.2}$	$40.4^{+5.1}_{-7.6}$	–
	x ⊙ grad (sign)	$-0.1^{+1.0}_{-1.3}$	$-1.7^{+4.8}_{-7.2}$	–
	x ⊙ grad (abs)	$5.0^{+1.3}_{-1.0}$	$27.1^{+6.8}_{-3.0}$	–
	IG (sign)	$22.6^{+2.3}_{-2.0}$	$127.1^{+28.9}_{-31.6}$	–
	IG (abs)	$7.7^{+1.2}_{-1.3}$	$43.0^{+8.8}_{-10.7}$	–
	LOO (sign)	$38.5^{+1.9}_{-3.6}$	$213.6^{+37.1}_{-40.5}$	–
	LOO (abs)	$9.9^{+0.4}_{-0.4}$	$55.4^{+11.9}_{-9.9}$	–
	Beam	$50.3^{+1.2}_{-2.9}$	$280.1^{+57.9}_{-45.7}$	–
SST2	Grad (L_2)	$9.8^{+1.1}_{-1.0}$	$32.4^{+3.5}_{-3.0}$	–
	Grad (L_1)	$9.7^{+1.1}_{-0.9}$	$32.0^{+3.2}_{-2.9}$	–
	x ⊙ grad (sign)	$-3.4^{+0.7}_{-0.6}$	$-11.4^{+2.4}_{-2.2}$	–
	x ⊙ grad (abs)	$5.4^{+1.6}_{-1.1}$	$18.0^{+5.2}_{-3.5}$	–
	IG (sign)	$40.1^{+3.2}_{-1.9}$	$133.1^{+10.6}_{-6.5}$	–
	IG (abs)	$15.6^{+0.9}_{-0.7}$	$51.6^{+3.4}_{-1.5}$	–
	LOO (sign)	$49.4^{+1.3}_{-1.4}$	$164.0^{+1.7}_{-2.9}$	–
	LOO (abs)	$17.2^{+1.0}_{-0.7}$	$57.1^{+3.6}_{-2.2}$	–
	Beam	$55.6^{+1.0}_{-0.5}$	$184.5^{+2.1}_{-2.6}$	–
SNLI	Grad (L_2)	$8.2^{+0.3}_{-0.5}$	$53.8^{+1.3}_{-2.0}$	–
	Grad (L_1)	$8.2^{+0.4}_{-0.5}$	$53.3^{+1.6}_{-2.1}$	–
	x ⊙ grad (sign)	$-0.3^{+0.3}_{-0.3}$	$-2.2^{+1.7}_{-2.2}$	–
	x ⊙ grad (abs)	$5.6^{+0.3}_{-0.4}$	$36.5^{+1.7}_{-2.0}$	–
	IG (sign)	$14.0^{+0.3}_{-0.3}$	$91.2^{+1.0}_{-1.0}$	–
	IG (abs)	$8.0^{+0.4}_{-0.5}$	$52.6^{+1.4}_{-1.7}$	–
	LOO (sign)	$26.3^{+0.4}_{-0.5}$	$172.3^{+4.3}_{-7.3}$	–
	LOO (abs)	$11.0^{+0.5}_{-0.3}$	$71.8^{+1.7}_{-0.9}$	–
	Beam	$29.6^{+0.3}_{-0.3}$	$193.7^{+5.2}_{-4.5}$	–
IMDB	Grad (L_2)	$13.9^{+3.5}_{-1.9}$	$29.4^{+6.0}_{-2.4}$	–
	Grad (L_1)	$13.7^{+4.1}_{-1.9}$	$28.9^{+6.0}_{-2.5}$	–
	x ⊙ grad (sign)	$-2.9^{+0.4}_{-0.4}$	$-6.4^{+1.3}_{-1.2}$	–
	x ⊙ grad (abs)	$7.7^{+2.9}_{-1.2}$	$16.3^{+3.9}_{-2.8}$	–
	IG (sign)	$53.2^{+3.4}_{-4.1}$	$114.2^{+12.8}_{-11.5}$	–
	IG (abs)	$18.9^{+3.7}_{-4.2}$	$40.3^{+6.2}_{-9.8}$	–
	LOO (sign)	$60.5^{+1.1}_{-1.0}$	$130.1^{+13.0}_{-13.0}$	–
	LOO (abs)	$16.7^{+5.0}_{-1.9}$	$35.5^{+5.6}_{-3.8}$	–
	Beam	–	–	–
MNLI	Grad (L_2)	$7.9^{+0.1}_{-0.2}$	$38.7^{+0.9}_{-1.3}$	–
	Grad (L_1)	$7.8^{+0.2}_{-0.2}$	$38.3^{+1.0}_{-1.5}$	–
	x ⊙ grad (sign)	$-0.5^{+0.5}_{-0.2}$	$-2.3^{+2.6}_{-1.1}$	–
	x ⊙ grad (abs)	$5.2^{+0.1}_{-0.1}$	$25.4^{+0.8}_{-1.1}$	–
	IG (sign)	$18.6^{+0.8}_{-1.1}$	$91.1^{+4.3}_{-4.6}$	–
	IG (abs)	$9.0^{+0.3}_{-0.2}$	$44.2^{+1.5}_{-0.7}$	–
	LOO (sign)	$33.0^{+0.9}_{-1.0}$	$161.9^{+6.1}_{-8.4}$	–
	LOO (abs)	$12.4^{+0.1}_{-0.2}$	$60.6^{+1.4}_{-0.9}$	–
	Beam	$41.2^{+0.9}_{-1.2}$	$201.7^{+6.5}_{-9.1}$	–
QNLI	Grad (L_2)	$9.5^{+0.5}_{-0.4}$	$28.3^{+1.4}_{-0.9}$	–
	Grad (L_1)	$9.4^{+0.3}_{-0.3}$	$28.0^{+1.3}_{-0.5}$	–
	x ⊙ grad (sign)	$-1.4^{+0.4}_{-0.5}$	$-4.1^{+1.4}_{-1.3}$	–
	x ⊙ grad (abs)	$6.5^{+0.3}_{-0.2}$	$19.1^{+0.7}_{-0.6}$	–
	IG (sign)	$25.0^{+3.3}_{-5.1}$	$74.0^{+9.6}_{-13.7}$	–
	IG (abs)	$9.7^{+1.2}_{-1.8}$	$28.7^{+3.2}_{-5.0}$	–
	LOO (sign)	$40.4^{+1.5}_{-2.1}$	$119.8^{+4.0}_{-5.0}$	–
	LOO (abs)	$12.5^{+0.2}_{-0.3}$	$37.0^{+0.3}_{-0.4}$	–
	Beam	$53.8^{+1.7}_{-2.1}$	$159.5^{+4.6}_{-4.4}$	–
QQP	Grad (L_2)	$4.0^{+0.3}_{-0.2}$	$33.5^{+1.4}_{-1.9}$	–
	Grad (L_1)	$4.0^{+0.3}_{-0.2}$	$33.0^{+1.3}_{-1.9}$	–
	x ⊙ grad (sign)	$-0.4^{+0.3}_{-0.2}$	$-3.3^{+2.3}_{-2.0}$	–
	x ⊙ grad (abs)	$2.5^{+0.3}_{-0.2}$	$20.7^{+1.3}_{-2.2}$	–
	IG (sign)	$8.9^{+1.0}_{-0.6}$	$73.7^{+4.8}_{-2.6}$	–
	IG (abs)	$3.8^{+0.7}_{-0.3}$	$31.6^{+2.5}_{-1.6}$	–
	LOO (sign)	$20.4^{+0.7}_{-0.4}$	$169.8^{+8.6}_{-11.1}$	–
	LOO (abs)	$5.7^{+0.2}_{-0.3}$	$47.3^{+2.3}_{-2.8}$	–
	Beam	$22.5^{+0.7}_{-0.8}$	$187.0^{+10.4}_{-10.3}$	–