On Faithfulness Disparity between Multilingual and Monolingual Models

Anonymous EMNLP submission

Abstract

In many application scenarios, practitioners not only aim to maximize predictive performance but also seek faithful explanations for the predictions. Rationales selected by faithful feature attribution methods provide insights into how different parts of the input contribute to the model prediction. Previous studies have explored how different factors affect faithfulness, however, these studies are mainly in the context of monolingual English models. On the other hand, the differences in explanation faithfulness between multilingual and monolingual models have yet to be explored. In this paper, we provide a comprehensive study on comparing the faithfulness between these two types of models. Our extensive experiments covering five languages and five popular feature attribution methods, showing that faithfulness varies between multilingual and monolingual models. For example, multilingual mBERT is more faithful than monolingual BERT, while multilingual RoBERTa is less faithful than monolingual RoBERTa. We show that the larger the multilingual model, the less faithful its rationales are compared to its counterpart monolingual model. Finally, we find that the faithfulness disparity is related to differences between multilingual and monolingual tokenizers, that when the tokenizers of multilingual models split words more aggressively, their faithfulness is closer to their monolingual counterparts.[1]

1 Introduction

Feature attribution methods (FAs) are commonly used to rank input features (i.e. tokens) according to their importance to a model's prediction (Kindermans et al., 2016; Sundararajan et al., 2017; DeYoung et al., 2020). Subsequently, the top-k ranked tokens are selected to form a rationale. The faithfulness of a FA method refers to what extent its selected rationales actually reflect the model's inner reasoning mechanism (Jacovi and Goldberg, 2020).

[1]Our code will be publicly released for reproducibility.

	Model	Rationales (highlighted)
FR	XLM-R	" bonjour **je** n ' **ai pas** recu l ' article commande, **car jai** commande **couleur** bois et jai recu noir! je n" **ai pas** du tout recu celui desire!!!"
	RoBERTa	" **bonjour** je n ' ai pas **recu** l ' article commande, car **jai** commande couleur bois et **jai recu** noir **!** je n ' ai pas du tout **recu** celui **desire** !!!"

Table 1: Rationales extracted for multilingual (XLM-R) and monolingual (French RoBERTa) models using the same FA and input for the same task (sentiment analysis in FR; prediction 'negative').

Previous work has mainly studied faithfulness in the context of monolingual models, i.e. especially English (Atanasova et al., 2020; Bastings and Filippova, 2020; Chan et al., 2022b). Furthermore, monolingual studies have investigated the impact of out-of-domain data (Chrysostomou and Aletras, 2022a), adversarial attacks (Sinha et al., 2021; Zhao et al., 2022a) and temporal shifts (Zhao et al., 2022b) on the faithfulness of FAs. Moreover, existing studies on interpreting multilingual models' behavior and their representations (Rama et al., 2020; Serikov et al., 2022; Gonen et al., 2022) have not focused on the faithfulness of FAs.

As shown in Table 1, even for the same input, prediction, and FA, the rationales selected are different between multi- and monolingual models. This indicates that they follow different inner processes for making predictions. It is unclear whether this difference is generally shared among input examples or even across other languages and models. Given that the performance of multilingual models might be on par with monolingual counterparts in various languages (Rust et al., 2021; Su et al., 2022), this leaves practitioners in a dilemma between choosing multilingual or monolingual models when the application scenario requires extracting faithful explanations for the model predictions. Therefore, we seek to answer *if there is a faithfulness disparity between multi- and monolingual models.*

Our main contributions are as follows:

- We perform a large empirical study across tasks in five languages, five popular FAs, and two types of monolingual and multilingual models;
- Our results reveal that the degree of faithfulness disparity can be attributed to the size of the models, i.e. larger multilingual models tend to have less faithful rationales compared to their monolingual counterparts;
- Our analysis shows that multilingual tokenizers split words into subwords more aggressively than monolingual models do. The more aggressively the multilingual models split words, their faithfulness is closer to their monolingual counterparts.

2 Related Work

2.1 Faithfulness of monolingual models

Feature attribution methods are commonly used to extract the importance degree of each token to the model prediction (Kindermans et al., 2016; Sundararajan et al., 2017; Belinkov et al., 2020; Kersten et al., 2021). The top-ranked tokens are considered as the rationales and their quality can be assessed in terms of plausibility and faithfulness (DeYoung et al., 2020; Jacovi and Goldberg, 2020). Faithfulness measures to what extent the rationales accurately reflect the model's internal reasoning process (Ribeiro et al., 2016; Zaidan et al., 2007; DeYoung et al., 2020; Jacovi and Goldberg, 2020; Pezeshkpour et al., 2021).[2]

Existing faithfulness studies on monolingual models mainly focus on English. Sinha et al. (2021) and Zhao et al. (2022a) explored how adversarial attacks affect the faithfulness of FAs by swapping tokens to create new inputs with the same semantics. Bastings et al. (2022) introduced ground truth, i.e. fully faithful rationales, with specific but meaningless tokens, to evaluate faithfulness. Chrysostomou and Aletras (2022a) investigated the impact of out-of-domain data on the model faithfulness, while Zhao et al. (2022b) studied the faithfulness on data from different time periods.

On the other hand, an increasing number of pretrained language models are made available for different languages (Antoun et al., 2020; Chan et al., 2020; Cañete et al., 2020; Le et al., 2020), there is no empirical evidence that non-English monolingual models are as faithful as English models.

2.2 Interpretability of multilingual models

Previous studies on the behavior of multilingual models focus on probing or analyzing the hidden representations, which are not directly related to the faithfulness of model explanations.

Santy et al. (2021) monitored the changes of attention heads in multilingual models when the model is further fine-tuned on monolingual and bilingual corpora. Rama et al. (2020) probed the representations of mBERT (multilingual BERT) between languages and they found that their distances correlate most with phylogenetic and geographical distances between languages. Gonen et al. (2022) analyzed the gender representations of multilingual models. Rust et al. (2021) studied the difference of multilingual models in processing different languages. They found that languages adequately represented in the multilingual model's vocabulary exhibit negligible performance decreases over their monolingual counterparts. Morger et al. (2022) examined the correlation between the human focus (eye-tracking) and model relative word importance on monolingual and multilingual language models.

Rather than studying the faithfulness of multilingual models, Zaman and Belinkov (2022) proposed a faithfulness evaluation method which they validated on multilingual models. They assume that an interpretation system is unfaithful if it provides different interpretations for similar inputs and outputs where the similar inputs have the same meaning in different languages.[3] While this work is relevant, it does not provide a comparison between monolingual and multilingual models.

2.3 Performance comparison of monolingual and multilingual models

Previous work has been conducted to compare the performance of monolingual and multilingual language models across languages. Nozza et al.

[2] Plausibility evaluates the extent to which the rationale aligns with human understanding (Jacovi and Goldberg, 2020; Chan et al., 2022a) and it is out of the scope of our study.

[3] The assumption that sentences in different languages are taken as "similar inputs" by the model has not been validated. It is unknown if models process similar yet different inputs in a similar manner (Jacovi and Goldberg, 2020; Ju et al., 2022).

Language	Model	Pre-training Corpus	#Tokens	Vocab	Params
Multi	mBERT	Wiki-100	3.3B	106K	167M
	XLM-R	CC-100	167B	250K	278M
English (EN)	BERT	Wikipedia, BookCorpus	3.3B	30K	109M
	RoBERTa	BookCorpus, cc news, Openwebtext, STORIES	40B	50K	125M
Chinese (ZH)	BERT	Wikipedia	0.4B	21K	103M
	RoBERTa	Wikipedia	0.4B	21K	102M
Spanish (ES)	BERT	Wikipedia, OPUS	3B	31K	110M
	RoBERTa	Web crawl	135B	50K	125M
French (FR)	BERT	Europeana	11B	32K	111M
	RoBERTa	Wikipedia, CC-100	59B	50K	124M
Hindi (HI)	BERT	L3Cube	0.3B	52K	126M
	RoBERTa	mc4, oscar, indic-nlp	1.5B	52K	83M

Table 2: Models' summary.

(2020) compared the performance between monolingual BERT variants and mBERT. Rönnqvist et al. (2019), Vulić et al. (2020) and Rust et al. (2021) conducted experiments with mBERT and monolingual BERT models with different selections of languages and testing tasks. Vulić et al. (2020) and Rust et al. (2021) further investigated the impact of lexical semantics and tokenizers on the performance differences respectively. A general observation drawn from these studies is that when the mono- and multilingual models have similar architectures and training objectives, their predictive performance is comparable regardless of the difficulty of the task.

Multilingual models' performance is often considered to suffer from the *"curse of multilinguality"* (Conneau et al., 2020; Pfeiffer et al., 2022), i.e. the inadequate capacity to represent all languages. To the best of our knowledge, no empirical study has validated this claim, let alone investigated how the curse of multilinguality impacts the faithfulness of multilingual models.

3 Experiments

Our aim is to compare the faithfulness between mono- and multilingual models across tasks and languages. For this purpose, we experiment with models of similar architectures and pre-training objectives following Rust et al. (2021). The main difference between them is the supported vocabularies. We evaluate models in various downstream tasks across a spectrum of typologically diverse and widely spoken languages.

3.1 Multilingual models

mBERT: A multilingual version of BERT (Devlin et al., 2019) following the same architecture and training objective of BERT. The primary difference is the training set that mBERT is trained on up to 104 languages from Wikipedia.

XLM-R: A multilingual version of RoBERTa (Conneau et al., 2020). The main difference is that XLM-R uses monolingual data from different languages and sample streams of text from each language. The training data includes 100 languages from Common Crawl.

3.2 Monolingual models

For each language, we include its monolingual BERT and RoBERTa as counterparts to mBERT and XLM-R respectively. We exclude monolingual models that are fine-tuned on bilingual or multilingual data. Table 2 in Appendix gives an overview of all models we experiment with across languages.

We fine-tune each model following the hyperparameter settings reported in the original papers describing the corresponding models and tasks. If not applicable, we use a batch size of 16, a learning rate of 1e-5 (1e-4 for the linear output layer), and an early stopping on 5 epochs. In all cases, our results are higher or comparable to the reported ones in previous studies. Further implementation details are given in the Appendix B. The predictive performance for each model on each task is reported in accuracy and F1 (Appendix D).

3.3 Datasets

Due to the lack of available data, it is impossible to use the exact same datasets in multiple languages. Therefore, we include a variety of tasks that are similar across languages. For example, we include sentiment analysis and language understanding tasks for each language. Details of datasets and their pre-processing are presented in Appendix C.

3.4 Feature attribution methods

We experiment with five popular FAs since there is no single best FA across models and tasks (Atanasova et al., 2020). Our aim is not to exhaustively benchmark various FAs but to explore faithfulness between mono- and multilingual models across different languages and tasks.

- **Attention (α):** Importance is computed using the corresponding normalized attention score (Jain et al., 2020).

- **Scaled attention ($\alpha\nabla\alpha$):** Attention scores scaled by their corresponding gradients (Serrano and Smith, 2019).

- **InputXGrad ($x\nabla x$):** It attributes importance by multiplying the input with its gradient computed with respect to the predicted class (Kindermans et al., 2016; Atanasova et al., 2020).

- **Integrated Gradients (IG):** This FA ranks input tokens by computing the integral of the gradients taken along a straight path from a baseline input (i.e. zero embedding vector) to the original input (Sundararajan et al., 2017).

- **DeepLift (DL):** It computes token importance according to the difference between the activation of each neuron and a reference zero embedding vector (Shrikumar et al., 2017).

Additionally, we include a random baseline that randomly assigns importance scores to each token.

3.5 Faithfulness evaluation

Sufficiency and comprehensiveness are two commonly-used metrics for evaluation faithfulness (DeYoung et al., 2020). Their normalized versions allow for a fairer comparison across models and tasks (Carton et al., 2020).

Normalized Sufficiency (Suff): Sufficiency captures the difference in predictive likelihood between retaining only the rationale $p(\hat{y}|\mathcal{R})$ and the full-text $p(\hat{y}|\mathbf{X})$:

$$\mathrm{S}(\mathbf{X}, \hat{y}, \mathcal{R}) = 1 - max(0, p(\hat{y}|\mathbf{X}) - p(\hat{y}|\mathcal{R}))$$
$$\text{Normalized } \mathrm{S}(\mathbf{X}, \hat{y}, \mathcal{R}) = \frac{\mathrm{S}(\mathbf{X}, \hat{y}, \mathcal{R}) - \mathrm{S}(\mathbf{X}, \hat{y}, 0)}{1 - \mathrm{S}(\mathbf{X}, \hat{y}, 0)} \quad (1)$$

where $\mathrm{S}(\mathbf{x}, \hat{y}, 0)$ is the sufficiency of a baseline input (zeroed out sequence) and $\hat{y}$ is the model predicted class using the full text $\mathbf{x}$ as input.

Normalized Comprehensiveness (Comp): It assesses how much information the rationale holds by measuring changes in predictive likelihoods when removing the rationale $p(\hat{y}|\mathbf{X}_{\backslash\mathcal{R}})$:

$$\mathrm{C}(\mathbf{X}, \hat{y}, \mathcal{R}) = max(0, p(\hat{y}|\mathbf{X}) - p(\hat{y}|\mathbf{X}_{\backslash\mathcal{R}}))$$
$$\text{Normalized } \mathrm{C}(\mathbf{X}, \hat{y}, \mathcal{R}) = \frac{\mathrm{C}(\mathbf{X}, \hat{y}, \mathcal{R})}{1 - \mathrm{S}(\mathbf{X}, \hat{y}, 0)} \quad (2)$$

Following DeYoung et al. (2020), we use the Area Over the Perturbation Curve (AOPC) for normalized sufficiency and comprehensiveness across different rationale lengths. We evaluate three different rationale ratios (10%, 20%, and 50%) and take the average value, similar to DeYoung et al. (2020) and Chan et al. (2022b).[4] The final sufficiency and comprehensiveness scores are computed after being divided by their corresponding random baseline (positive values of these ratios denote higher than random faithfulness, the higher the more faithful).

4 Results

Our experiments include two multilingual and ten monolingual models, five FAs, and 15 tasks. Specifically, we test four models (two multilingual and two monolingual), three tasks, and five FAs for each language, measuring sufficiency and comprehensiveness. This results in 120 faithfulness evaluation cases for each language, 600 cases in total. All sufficiency, comprehensiveness, and predictive performance (accuracy and F1) for each model and task can be found in Appendix D.

4.1 Faithfulness between monolingual and multilingual models

Language	Model	BERT			RoBERTa		
		Accuracy	Suff	Comp	Accuracy	Suff	Comp
English	Mono	**0.847**	1.146	1.525	**0.852**	**1.306**	**1.588**
	Multi	0.837	**1.224**	**1.604**	0.841	1.163	1.210
Chinese	Mono	**0.833**	1.101	1.142	0.816	**1.093**	**1.156**
	Multi	0.819	**1.137**	**1.271**	**0.825**	1.088	1.000
French	Mono	0.825	1.047	1.057	0.822	**1.242**	**1.510**
	Multi	**0.844**	**1.130**	**1.259**	**0.851**	1.049	1.055
Spanish	Mono	0.849	1.024	1.046	**0.857**	**1.235**	**1.176**
	Multi	**0.852**	**1.146**	**1.214**	0.849	1.082	1.055
Hindi	Mono	**0.716**	1.162	**1.177**	0.693	**1.094**	**1.097**
	Multi	0.685	**1.202**	1.157	**0.718**	1.086	1.084

Table 3: Predictive performance ("Accuracy") and faithfulness ("Suff" and "Comp") of mono- and multilingual models. For all values, the higher the better (F1 for prediction performance is available in Appendix D).

Table 3 overviews the predictive performance and faithfulness (sufficiency and comprehensiveness) of models, averaged on the three tasks and FAs for each language.

We first observe that the performance of mono- and multilingual models is comparable to each other. For instance, the difference between Spanish BERT and mBERT is merely 0.003. The largest gap is found between Hindi BERT (0.716) and mBERT (0.685), exhibiting a difference of 0.031. Our results are also in line with results reported by Rust et al. (2021), which tested BERT with a different language set, including Arabic, Finnish,

[4] For tasks of average length over 200, we evaluate rationale ratios of 1%, 5%, and 10% instead, to extract rationales in reasonable lengths.

Indonesian, Japanese, Korean, Russian, and Turkish (presented in Table 9 in Appendix).[5]

Second, we note that the faithfulness disparity of mono- and multilingual models is consistent and follows different directions between BERT and RoBERTa. Specifically, XML-R consistently obtains lower faithfulness (both sufficiency and comprehensiveness) than monolingual RoBERTas, whereas mBERT gains higher faithfulness than its monolingual BERTs (except for sufficiency in Hindi). Additionally, the faithfulness disparity of RoBERTa is more noticeable as half of the cases have a faithfulness difference greater than 0.1. For example, the comprehensiveness in French is 1.510 for French RoBERTa but only 1.055 for XLM-R, differing by 0.475. We further explore this differences between BERT and RoBERTa in Section 5.

4.2 Faithfulness disparity across FAs

	α	$\alpha\nabla\alpha$	$x\nabla x$	IG	DL	Avg Diff	P value
Sufficiency							
English	-0.082	-0.086	-0.097	-0.131	-0.319	-0.143	0.258
Chinese	0.065	0.056	-0.085	-0.040	-0.018	-0.005	0.946
Spanish	-0.070	-0.138	-0.336	-0.107	-0.111	-0.153	0.053
French	-0.206	-0.218	-0.133	-0.217	-0.188	-0.193	0.007
Hindi	-0.054	-0.047	0.045	-0.068	0.081	-0.009	0.888
Avg Diff	-0.070	-0.086	-0.121	-0.113	-0.111	-0.100	-
P value	0.535	0.462	0.041	0.033	0.076	-	0.006
Comprehensiveness							
English	-0.465	-0.436	-0.327	-0.333	-0.330	-0.378	0.000
Chinese	-0.230	-0.224	-0.111	-0.156	-0.062	-0.157	0.010
Spanish	-0.197	-0.116	-0.105	0.032	-0.218	-0.121	0.076
French	-0.486	-0.482	-0.232	-0.598	-0.475	-0.455	0.004
Hindi	0.071	0.062	-0.036	-0.268	0.082	-0.018	0.831
Avg Diff	-0.261	-0.239	-0.162	-0.265	-0.201	-0.226	-
P value	0.027	0.034	0.004	0.015	0.070	-	0.000

Table 4: Faithfulness difference between multilingual RoBERTa (XLM-R) and counterpart monolingual RoBERTa (plum indicates monolingual models are more faithful than multilingual models.)

Tables 4 and 5 delve deeper into the faithfulness disparity between mono- and multilingual models, presenting the results for RoBERTa and BERT models per FA. Disparity is computed as the multilingual faithfulness (sufficiency or comprehensiveness) score minus its monolingual counterpart.

Looking into individual FAs, IG shows a greater faithfulness disparity than other FAs. For example, it obtains the greatest disparity in comprehensiveness averaged over languages for both RoBERTa and BERT; and the greatest and the second greatest in sufficiency over languages for BERT and

	α	$\alpha\nabla\alpha$	$x\nabla x$	IG	DL	Avg Diff	P value
Sufficiency							
English	0.086	0.093	-0.024	0.187	0.048	0.078	0.292
Chinese	-0.018	-0.037	0.043	0.176	0.016	0.036	0.454
Spanish	0.200	0.202	0.006	0.190	0.015	0.123	0.049
French	0.184	0.173	-0.028	0.063	0.025	0.083	0.066
Hindi	-0.041	-0.035	0.010	0.266	-0.003	0.039	0.510
Avg Diff	0.082	0.079	0.001	0.176	0.020	0.072	-
P value	0.264	0.298	0.966	0.003	0.527	-	0.005
Comprehensiveness							
English	0.122	0.106	0.075	0.078	0.015	0.079	0.323
Chinese	0.211	0.213	0.028	0.176	0.016	0.129	0.053
Spanish	0.268	0.268	0.040	0.160	0.105	0.168	0.048
French	0.294	0.299	0.046	0.217	0.156	0.202	0.049
Hindi	-0.232	-0.234	-0.128	0.138	0.057	-0.080	0.307
Avg Diff	0.133	0.130	0.012	0.154	0.070	0.100	-
P value	0.258	0.263	0.758	0.040	0.081	-	0.007

Table 5: Faithfulness difference between multilingual BERT (mBERT) and counterpart monolingual BERT.

RoBERTa. IG computes the integral of gradients of each input element, compared to a blank input, modeling the absence of the feature/token (Sundararajan et al., 2017). According to Sundararajan et al. (2017), compared to $x\nabla x$, IG is less sensitive to unimportant features. This is because $x\nabla x$ is over-sensitive to all features, e.g. blank input which is supposed to be the most unimportant one. This still leads to a gradient value that is closer to non-blank inputs (Shrikumar et al., 2016). The large faithfulness disparity of IG intuitively indicates that multilingual and monolingual models consider different tokens as unimportant during inference.

The disparities of Attention-based FAs, i.e. α and $\alpha\nabla\alpha$, are consistently on par with each other. This indicates the attention values, scaled or not scaled by gradients, are unlikely to introduce big changes to the attention values as the attention being with greater magnitudes compared to the corresponding gradients values (Serrano and Smith, 2019; Jain et al., 2020).

Overall, attention-based FAs demonstrate great disparities. These are larger than $x\nabla x$ and DL in most cases of comprehensiveness in RoBERTa and BERT, and sufficiency in BERT. *It is, therefore, likely that mono- and multilingual models reach similar predictions by attending to different tokens.*

5 Analysis

5.1 RoBERTa vs. BERT

Figure 1 compares the overall faithfulness between mBERT and XLM-R.[6] Each point represents a

[5]We do not include these languages as they do not have RoBERTa monolingual models.

[6]The figures for monolingual models are presented in Appendix D.

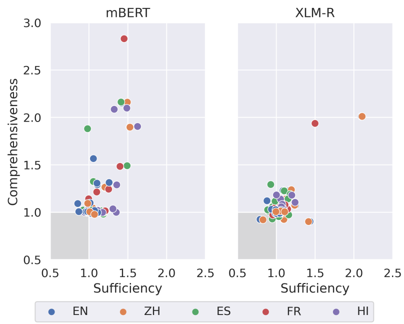

Figure 1: Faithfulness of the two multilingual models across languages. The dark grey area (bottom left) indicates unfaithfulness (low Suff and Comp).

	α	$\alpha\nabla\alpha$	$x\nabla x$	IG	DL	Avg Diff	P value
Sufficiency							
English	-0.360	-0.354	-0.124	-0.445	-0.214	-0.300	0.001
Chinese	-0.143	-0.133	-0.042	-0.220	-0.044	-0.116	0.157
Spanish	-0.172	-0.240	-0.352	-0.278	-0.160	-0.240	0.001
French	-0.309	-0.314	-0.120	-0.248	-0.188	-0.236	0.000
Hindi	0.010	0.012	0.039	-0.239	0.001	-0.035	0.711
Avg Diff	-0.195	-0.206	-0.120	-0.286	-0.121	-0.186	-
P value	0.057	0.050	0.045	0.000	0.035	-	0.000
Comprehensiveness							
English	-0.201	-0.314	-0.366	0.078	-0.448	-0.250	0.204
Chinese	-0.266	-0.254	-0.047	-0.303	-0.048	-0.183	0.055
Spanish	-0.184	-0.102	-0.003	-0.029	-0.177	-0.099	0.060
French	-0.484	-0.484	-0.124	-0.627	-0.449	-0.434	0.005
Hindi	0.103	0.091	-0.022	-0.364	0.101	-0.018	0.868
Avg Diff	-0.206	-0.212	-0.112	-0.249	-0.204	-0.197	-
P value	0.147	0.119	0.088	0.169	0.088	-	0.001

Table 6: Faithfulness difference between multilingual RoBERTa Large (XLM-R Large) and counterpart monolingual RoBERTa.

FA's sufficiency (x-axis) and comprehensiveness (y-axis), on a given task (not specify the task but its language by color). XLM-R shows lower variance among languages, indicated by a more dispersed distribution of data points than mBERT. One potential explanation for this is that English has the overwhelmingly largest portion in the pre-training corpus for mBERT, while XLM-R increases the portion of corpora in non-English languages. We include Figure 3 in Appendix E.1, which compares the pre-training corpus size of different languages for XLM-R and mBERT (Conneau et al., 2020). It shows the amount of data for languages such as French (FR) and Chinese (ZH) has increased by several orders of magnitude.

In Section 4.2, we observed contrasting directions of faithfulness disparities. XLM-R exhibited lower faithfulness compared to monolingual RoBERTa, whereas mBERT demonstrated higher faithfulness than monolingual BERT. We hypothesize that this phenomenon is linked to the gap in model size between mono- and multilingual models. Specifically, mBERT has at least 1.5 times more parameters than monolingual BERT models, while XLM-R has at least 2.2 times more parameters than monolingual RoBERTa models. The difference in model size may account for the opposite directions of faithfulness disparities between RoBERTa and BERT. If this holds true, we anticipate that when the model size gap increases, XLM-R will still provide less faithful rationales than monolingual RoBERTa while their disparity degree will increase.

5.2 Impact of model size

To further investigate the impact of the model size, we repeat all experiments with XLM-R large and compare its faithfulness with monolingual RoBERTa. In this case, the size difference between multi- and monolingual models is bigger than XLM-R base v.s. monolingual RoBERTa. XLM-R base and XLM-R large use the same pre-training corpus, pre-training objective, and similar model architectures, but differ in model parameter numbers[7] (Conneau et al., 2020). XLM-R large (550M parameters) is at least 4.7 times larger than monolingual RoBERTa models.

Table 6 shows the faithfulness disparity of multilingual RoBERTa large and monolingual RoBERTa. Full results of faithfulness are in Appendix E.2. First, we see that the faithfulness disparity direction remains the same as XLM-R base and monolingual RoBERTa. That is, monolingual RoBERTa is more faithful than XLM-R large.

Second, the overall sufficiency disparity increases from -0.100 to -0.186. It also increases for each individual FA and language, with IG being the only exception to remain almost the same (-0.120 and -0.121). For example, the average disparity in English increases from -0.143 to -0.300 and the average disparity for attention increases from -0.070 to -0.195. The overall comprehensiveness disparity of XLM-R large is on par with XLM-R base (-0.226 v.s. -0.197). Also, the changes of faithfulness disparity fluctuate on each FA and language that XLM-R large increases in some cases (e.g. Chinese and IG) and decreases in others (e.g. Spanish and attention).

Overall, the results confirm our assumption that the difference in model size is related to the faithful-

[7]Both are transformer-based, XLM-R base: L = 12, H = 768, A = 12; XLM-R large: L = 24, H = 1024, A = 16)

ness disparity. The larger the multilingual model, the less faithful its rationales are compared to its monolingual counterpart. One intuitive interpretation behind this is that when the model gets larger, it becomes intrinsically complex and therefore, it is harder to faithfully explain its predictions with FA methods. To summarize, *the more parameters the multilingual model has, the less faithful its rationales are compared to its monolingual counterparts. Therefore, we suggest using monolingual models for faithful rationales when the multilingual model is much larger than the monolingual counterpart.*

We acknowledge that our findings might not generalize to BERT because mBERT large (or different sizes) are not available to experiment with. To overcome this, we repeat all experiments on BERT-large and compare its faithfulness with BERT-base, to investigate the impact of model size from a different perspective. To keep the focus of the paper on the faithfulness disparity between mono- and multilingual models, we present the results and analysis in Table 14 in the Appendix.

5.3 Impact of tokenization

Previous research has shown the essential impact of the tokenizer on multilingual models (Ruan et al., 2021; Zhang et al., 2022). Intuitively, multilingual tokenizers are less specialized than their counterpart monolingual tokenizers for the specific language. For example, as shown in Table 2, the multilingual BERT tokenizer has a vocabulary size of 105K covering 104 languages, while the five monolingual BERT tokenizers cover a vocabulary of 167k tokens already. Therefore, we investigate the impact of tokenizers on the faithfulness disparity. BERT-based models use WordPiece as their tokenizers (Wu et al., 2016). Monolingual RoBERTa-based models use BytePair-Encoding (BPE) (Sennrich et al., 2016), and multilingual XLM-R uses SentencePiece (Kudo and Richardson, 2018). We do not compare their splitting mechanisms but their splitting results, especially how aggressively they split words into subwords. The superficial splitting of a tokenizer intuitively reflects how many unique tokens it knows for the language, i.e. how well the tokenizer knows the language. Following Rust et al. (2021), we examine two metrics across tokenizers, fertility and splitting ratio.

- **Fertility** measures the average number of subwords produced per tokenized word, a.k.a subword fertility (Rust et al., 2021). The minimum fertility value is 1 when the tokenizer's vocabulary contains every word in the text. The higher the fertility, the larger the number of subwords generated when splitting words.
- **Splitting ratio** calculates the proportion of words split during tokenization (Rust et al., 2021).[8] The maximum splitting ratio is 1 when the tokenizer splits each word into subwords. The higher the splitting ratio, the more words are split during tokenization.

Fertility indicates how many subwords a tokenizer splits a word into, the splitting ratio shows how often a tokenizer splits words. Intuitively, low scores are preferable for both metrics as they indicate that the tokenizer is well-suited to the language (Rust et al., 2021).

Table 7 shows the fertility and splitting ratio difference between monolingual and multilingual models (i.e. multilingual score minus its counterpart monolingual).[9] Faithfulness disparity values are taken from Tables 4 and 5.

First, for both RoBERTa and BERT, the positive values of fertility and splitting ratio difference indicates that multilingual models tend to be more aggressive in splitting words than monolingual ones. For example, as shown in Table 15 in Appendix, 26.1% English words (underlined in table) are split by multilingual RoBERTa tokenizer but only 7.6% (underlined in table) by monolingual RoBERTa tokenizer.

Second, RoBERTa has larger gaps in both fertility and splitting ratio than BERT for all languages. For all three languages, the fertility and the splitting ratio differences are greater than 0.1 for RoBERTa, but less than 0.1 for BERT. This is because SentencePiece (multilingual XLM-R's tokenizer) is generally more aggressive in splitting words. Taking English as an example, the fertility gap among monolingual RoBERTa (BPE), monolingual BERT (WordPiece) tokenizers, and multilingual BERT (WordPiece) is relatively smaller, 1.125, 1.115, and 1.179 respectively, while the fertility of XLM-R (SentencePiece) is 1.319. However, this is counterintuitive given the much larger vocabulary size

[8] Different tokenizers present subwords and non-subwords differently. Details can be found in Appendix G.

[9] Hindi and Chinese are excluded from this analysis because Hindi does not show a significant difference between mono- and multilingual in either sufficiency or comprehensiveness for RoBERTa and BERT; Chinese is a logographic language without white spaces.

	Multi Fertility	Mono Fertility	Fertility Diff	Multi Splitting	Mono Splitting	Splitting Diff	Suff Diff	Comp Diff
RoBERTa								
English	1.319	1.125	0.195	0.261	0.076	0.185	-0.300	-0.250
Spanish	1.409	1.290	0.119	0.299	0.195	0.104	-0.240	-0.099
French	1.531	1.345	0.186	0.325	0.211	0.114	-0.236	-0.434
Avg	1.420	1.253	0.167	0.312	0.203	0.134	-0.259	-0.261
BERT								
	Multi Fertility	Mono Fertility	Fertility Diff	Multi Splitting	Mono Splitting	Split ratio Diff	Suff Diff	Comp Diff
English	1.179	1.115	0.064	0.111	0.059	0.052	0.078	0.079
Spanish	1.369	1.283	0.086	0.152	0.090	0.062	0.123	0.168
French	1.461	1.456	0.005	0.139	0.134	0.005	0.083	0.202
Avg	1.336	1.285	0.052	0.134	0.094	0.040	0.095	0.150

Table 7: Fertility, splitting ratio, sufficiency, and comprehensiveness difference between multilingual and monolingual models (positive values indicate multilingual is more faithful). Full results of fertility and splitting ratio for each dataset can be found in Table 15 in Appendix H.

of multilingual RoBERTa (over two times bigger than multilingual BERT, see Figure 2). One potential explanation is that XLM-R saves capacity for representing the vocabulary for other low-resource languages. On the other hand, the greater aggressiveness in tokenization of multilingual RoBERTa potentially explains the different disparity direction to BERT models. That is, only when the fertility difference is greater than 0.1, do multilingual models gain higher faithfulness than their monolingual counterparts.[10] An intuitive reason might be that more fine-grained tokenization breaks the balance of keeping certain linguistic units together during faithfulness evaluation.

Last, the differences in sufficiency and comprehensiveness demonstrate a high negative relationship to the fertility difference (Table 8). That is, the larger the fertility difference between mono- and multilingual models, the smaller the faithfulness disparity. Particularly, the fertility and the comprehensiveness difference show a very high negative correlation (-0.91).

ρ	Suff Diff	Comp Diff
Splitting Diff	-0.86	-0.79
Fertility Diff	-0.86	-0.91

Table 8: Pearson correlation coefficient between fertility, splitting ratio, and faithfulness disparity.

To sum up, *multilingual tokenizers split words into subwords more aggressively than monolingual tokenizers. The degree of splitting difference is strongly correlated with the faithfulness disparity between models. The aggressive tokenization of multilingual models might result in lower faithfulness, particularly when the fertility and splitting differences are greater than 0.1, compared to their monolingual counterparts.*

[10]We demonstrate this pattern in Figure 4, Appendix I.

5.4 Qualitative analysis

For a qualitative evaluation, we examine the rationales extracted by the same faithful FAs for both types of models. We observe that rationales of multilingual models more often contain pronouns, prepositions, postpositions, conjunction, and article words, while monolingual models' prefer nouns and adjectives. We suspect the different preferences in parts of speech are due to monolingual models being more specialized for the language so that its rationales contain more specific nouns and adjectives rather than general functional words such as pronouns, prepositions, postpositions, and conjunctions. We also observe examples where multilingual tokenizers tokenize more aggressively, e.g. the word "defectos" in Spanish ("defects" in English) is not split into subwords by Spanish BERT, but split into "'def', '##ecto', '##s' by mBERT; "desagradable" in Spanish ("unpleasant") is not split by Spanish BERT but split into 'desa', '##grada', '##ble' by mBERT, echoing observations in Section 5.3.

6 Conclusion

To the best of our knowledge, our study is the first to investigate the faithfulness disparity between monolingual and multilingual models. We have conducted a comprehensive empirical study and found that faithfulness gaps exist across languages, models, and FAs. Our study further reveals that the larger the multilingual model, the less faithful its rationales are compared to its monolingual counterpart models. Finally, we found that the disparity is highly correlated to the gap between mono- and multilingual tokenizers on how aggressively they split words. Future work includes exploring models for low-resource languages and other language families, such as Austronesian and Afroasiatic.

Limitations

As outlined in the paper, one significant challenge we encountered during our research was the absence of monolingual models in various languages. First, monolingual models are only available in a few languages, such as monolingual BERT and RoBERTa models used in this paper. Second, more recent decoder-based models, such as T5, Llama, and GPT2, are multilingual by default.

Furthermore, it would be intriguing to explore the faithfulness disparity and behavior of feature attributions for low-resource languages, particularly given their limited corpus during pre-training.

An additional uncontrolled factor is the impact of the different pre-training corpora between monolingual and multilingual models (see Table 2). However, it is not feasible to disentangle this factor in our experiments since we would need to obtain comparable corpora in various languages and pretrain from scratch all models.

Last, it is important to acknowledge that multilingual studies focusing on Indo-European and Sino-Tibetan languages may not necessarily apply to languages outside these language families. We hope future work can contribute resources to facilitate the development of a more diverse range of monolingual language models.

References

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

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

Jasmijn Bastings, Sebastian Ebert, Polina Zablotskaia, Anders Sandholm, and Katja Filippova. 2022. “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*, pages 976–991, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick. 2020. Interpretability and analysis in neural NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–5, Online. Association for Computational Linguistics.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020(2020):1–10.

José Cañete, Sebastian Donoso, Felipe Bravo-Marquez, Andrés Carvallo, and Vladimir Araujo. 2022. ALBETO and DistilBETO: Lightweight Spanish language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4291–4298, Marseille, France. European Language Resources Association.

Samuel Carton, Anirudh Rathore, and Chenhao Tan. 2020. Evaluating and characterizing human rationales. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9294–9307, Online. Association for Computational Linguistics.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Aaron Chan, Maziar Sanjabi, Lambert Mathias, Liang Tan, Shaoliang Nie, Xiaochang Peng, Xiang Ren, and Hamed Firooz. 2022a. UNIREX: A unified learning framework for language model rationale extraction. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 51–67, virtual+Dublin. Association for Computational Linguistics.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German’s next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Chun Sik Chan, Huanqi Kong, and Liang Guanqing. 2022b. A comparative study of faithfulness metrics for model interpretability methods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5029–5038, Dublin, Ireland. Association for Computational Linguistics.

George Chrysostomou and Nikolaos Aletras. 2022a. An empirical study on explanations in out-of-domain settings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6920–6938, Dublin, Ireland. Association for Computational Linguistics.

George Chrysostomou and Nikolaos Aletras. 2022b. Flexible instance-specific rationalization of nlp models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10545–10553.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Asier Gutiérrez Fandiño, Jordi Armengol Estapé, Marc Pàmies, Joan Llop Palao, Joaquin Silveira Ocampo, Casimiro Pio Carrino, Carme Armentano Oller, Carlos Rodriguez Penagos, Aitor Gonzalez Agirre, and Marta Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68.

Hila Gonen, Shauli Ravfogel, and Yoav Goldberg. 2022. Analyzing gender representation in multilingual models. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 67–77, Dublin, Ireland. Association for Computational Linguistics.

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

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. 2020. Learning to faithfully rationalize by construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4459–4473, Online. Association for Computational Linguistics.

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

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

Tom Kersten, Hugh Mee Wong, Jaap Jumelet, and Dieuwke Hupkes. 2021. Attention vs non-attention for a shapley-based explanation method. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 129–139, Online. Association for Computational Linguistics.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

Pieter-Jan Kindermans, Kristof Schütt, Klaus-Robert Müller, and Sven Dähne. 2016. Investigating the influence of noise and distractors on the interpretation of neural networks. *arXiv preprint arXiv:1611.07270*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Seventh International Conference on Learning Representations ICLR 2019*.

Felix Morger, Stephanie Brandl, Lisa Beinborn, and Nora Hollenstein. 2022. A cross-lingual comparison of human and model relative word importance. In *Proceedings of the 2022 CLASP Conference on*

(Dis)embodiment, pages 11–23, Gothenburg, Sweden. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2020. What the [mask]? making sense of language-specific bert models. *arXiv preprint arXiv:2003.02912*.

Pouya Pezeshkpour, Sarthak Jain, Byron Wallace, and Sameer Singh. 2021. An empirical comparison of instance attribution methods for NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 967–975, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

Taraka Rama, Lisa Beinborn, and Steffen Eger. 2020. Probing multilingual BERT for genetic and typological signals. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1214–1228, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. 2019. Is multilingual BERT fluent in language generation? In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 29–36, Turku, Finland. Linköping University Electronic Press.

Xiaoyi Ruan, Meizhi Jin, Jian Ma, Haiqin Yang, Lianxin Jiang, Yang Mo, and Mengyuan Zhou. 2021. Sattiy at SemEval-2021 task 9: An ensemble solution for statement verification and evidence finding with tables. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1255–1261, Online. Association for Computational Linguistics.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Sebastin Santy, Anirudh Srinivasan, and Monojit Choudhury. 2021. BERTologiCoMix: How does code-mixing interact with multilingual BERT? In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 111–121, Kyiv, Ukraine. Association for Computational Linguistics.

Stefan Schweter. 2020. Europeana bert and electra models.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Oleg Serikov, Vitaly Protasov, Ekaterina Voloshina, Viktoria Knyazkova, and Tatiana Shavrina. 2022. Universal and independent: Multilingual probing framework for exhaustive model interpretation and evaluation. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 441–456, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3145–3153. PMLR.

Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. 2016. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*.

Sanchit Sinha, Hanjie Chen, Arshdeep Sekhon, Yangfeng Ji, and Yanjun Qi. 2021. Perturbing inputs for fragile interpretations in deep natural language processing. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 420–434, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hui Su, Weiwei Shi, Xiaoyu Shen, Zhou Xiao, Tuo Ji, Jiarui Fang, and Jie Zhou. 2022. RoCBert: Robust Chinese bert with multimodal contrastive pretraining. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 921–931, Dublin, Ireland. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Shagun Uppal, Vivek Gupta, Avinash Swaminathan, Haimin Zhang, Debanjan Mahata, Rakesh Gosangi, Rajiv Ratn Shah, and Amanda Stent. 2020. Two-step classification using recasted data for low resource settings. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 706–719, Suzhou, China. Association for Computational Linguistics.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267, Rochester, New York. Association for Computational Linguistics.

Kerem Zaman and Yonatan Belinkov. 2022. 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*, pages 1556–1576, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shiyue Zhang, Vishrav Chaudhary, Naman Goyal, James Cross, Guillaume Wenzek, Mohit Bansal, and Francisco Guzman. 2022. How robust is neural machine translation to language imbalance in multilingual tokenizer training? In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 97–116, Orlando, USA. Association for Machine Translation in the Americas.

Yang Zhao, Zhang Yuanzhe, Jiang Zhongtao, Ju Yiming, Zhao Jun, and Liu Kang. 2022a. Can we really trust explanations? evaluating the stability of feature attribution explanation methods via adversarial attack. In *Proceedings of the 21st Chinese National Conference on Computational Linguistics*, pages 932–944, Nanchang, China. Chinese Information Processing Society of China.

Zhixue Zhao, George Chrysostomou, Kalina Bontcheva, and Nikolaos Aletras. 2022b. On the impact of temporal concept drift on model explanations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4039–4054, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Comparison of predictive performance

Lg	Model	NER Test F1	SA Test Acc	QA Dev EM / F1	UDP Test UAS/LAS	POS Test Acc
Arabic AR	Monolingual mBERT	91.1 90	95.9 95.4	68.3/82.4 66.1/80.6	90.1/85.6 88.8/83.8	96.8 96.8
English	Monolingual mBERT	91.5 91.2	91.6 89.8	80.5/88.0 80.9/88.4	92.1/89.7 91.6/89.1	97 96.9
Finnish	Monolingual mBERT	92 88.2	- -	69.9/81.6 66.6/77.6	95.9/94.4 91.9/88.7	98.4 96.2
Indonesian	Monolingual mBERT	91 93.5	96 91.4	66.8/78.1 71.2/82.1	85.3/78.1 85.9/79.3	92.1 93.5
Japanese	Monolingual mBERT	72.4 73.4	88 87.8	- -	94.7/93.0 94.0/92.3	98.1 97.8
Korean	Monolingual mBERT	88.8 86.6	89.7 86.7	74.2/91.1 69.7/89.5	90.3/87.2 89.2/85.7	97 96
Russian	Monolingual mBERT	91 90	95.2 95	64.3/83.7 63.3/82.6	93.1/89.9 91.9/88.5	98.4 98.2
Turkish	Monolingual mBERT	92.8 93.8	88.8 86.4	60.6/78.1 57.9/76.4	79.8/73.2 74.5/67.4	96.9 95.7
Chinese	Monolingual mBERT	76.5 76.1	95.3 93.8	82.3/89.3 82.0/89.3	88.6/85.6 88.1/85.0	97.2 96.7
AVG	Monolingual mBERT	87.4 87	92.4 91	70.8/84.0 69.7/83.3	90.0/86.3 88.4/84.4	96.9 96.4

Table 9: Comparison of predictive performance between mBERT and monolingual BERT across languages and tasks. Results are drawn from Rust et al. (2021)

As shown in Table 9, the predictive performance of mBERT is comparable with monolingual BERT in most cases. Particularly, on Russian and Chinese, the difference between monolingual and multilingual models is not greater than 1.2 and 1.5 across each task, respectively.

B Model Implementation Details

Language	Models	Huggingface ID	
Multilingual	mBERT XLM-R XLM-R large	bert-base-multilingual-uncased xlm-roberta-base xlm-roberta-large	Devlin et al. (2019) Conneau et al. (2020) Conneau et al. (2020)
English	BERT RoBERTa	bert-base-uncased roberta-base	Devlin et al. (2019) Liu et al. (2019)
Chinese	BERT RoBERTa	bert-base-chinese hfl/chinese-roberta-wwm-ext	Devlin et al. (2019) Cui et al. (2021)
Spanish	BERT RoBERTa	dccuchile/bert-base-spanish-wwm-uncased PlanTL-GOB-ES/roberta-base-bne	Cañete et al. (2020) Fandiño et al. (2022)
French	BERT RoBERTa	dbmdz/bert-base-french-europeana-cased ClassCat/roberta-base-french	Schweter (2020) n/a
Hindi	BERT RoBERTa	l3cube-pune/hindi-bert-scratch flax-community/roberta-hindi	Joshi (2022) n/a

Table 10: Model references

We use pre-trained models from the Huggingface library (Wolf et al., 2020). We use the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of $1e^{-5}$ for fine-tuning ($1e^{-4}$ for the linear output layer). We fine-tune all models for 5 epochs using a linear scheduler, with 10% of the data in the first epoch as warming up. We also use a grad-norm of 1.0. The model with the lowest loss on the development set is selected. All models are trained across 3 random seeds, and we report the average prediction performance. The best model among the 3 runs is used to extract rationales. Experiments are run on a single NVIDIA A100 GPU.

C Datasets

Table 11 on page 14 gives details of each task. Following Su et al. (2022), we use the small version of ChnSentiCorp data. Following (Le et al., 2020), we sample 2000 data from the original French CSL dataset as the training set and also 2000 for the testing and development set separately. We do the same for Hindi CSL and Spanish CSL. Further, for tasks without a published testing set and a published development set, we split the original set into an 8:1:1 training:testing:development split with the same label distribution.

D Full Results of Faithfulness

D.1 Faithfulness full results

Table 12 on page 14 shows the sufficiency and comprehensiveness of each feature attribution method on each dataset. "Suff" is short for sufficiency, "comp" for comprehensiveness. All faithfulness results are presented as the ratio after being divided by the random baseline (i.e. assigning a random importance distribution to the token sequence and then computing the sufficiency and the comprehensiveness). The predictive results, F1 and accuracy, are the average over three runs. The best model from the three runs is taken to extract and evaluate the rationales with each feature attribution method separately.

D.2 Faithfulness overview of monolingual models

Figure 2 on page 15 is the monolingual counterpart figure for Figure 1 on page 6. It overviews the faithfulness of monolingual BERT and RoBERTa, regardless of noticing the feature attribution used. The points in the grey area (left bottom) are unfaithful in both sufficiency and comprehensiveness. As shown in the figure, most cases are faithful on at least one of sufficiency or comprehensiveness. This validates our comparison of faithfulness and faithfulness disparity. Otherwise, it is not reasonable to say one is more faithful than the other if both are unfaithful.

Language	Language Family	Dataset	Task	Training set size	Avg length	Metrics	Papers
English	Indo-European	SST	Sentiment analysis	6,920 / 872 / 1,821	17	F1	Chrysostomou and Aletras (2022b)
		Agnews	Topic classification	102,000 / 18,000 / 7,600	36	F1	Chrysostomou and Aletras (2022b)
		MultiRC	Multi-Sentence Reading Comprehension	24,029 / 3,214 / 4,848	290/17	F1	Chrysostomou and Aletras (2022b)
Chinese	Sino-Tibetan	Ant	Financial Question Matching	30,018 / 4,316 / 4,316	13/13	Accuracy	Su et al. (2022)
		KR	Keyword Recognition	17,000 / 3,000 / 3,000	266/29	Accuracy	Su et al. (2022)
		ChnSentiCorp	Sentiment analysis	2,000 / 1,200 / 1,200	107	Accuracy	Su et al. (2022)
Spanish	Indo-European	CSL	Sentiment analysis	2,000 / 1,200 / 1,200	27	Accuracy	Keung et al. (2020)
		PAWS-X	Paraphrase Identification	49,400 / 2,000 / 2,000	20/20	Accuracy	Yang et al. (2019)
		XNLI	Natural Language Inference	393,000 / 5,010 / 2,490	19/9	Accuracy	Conneau et al. (2020), Conneau et al. (2020)
French	Indo-European	CSL	Sentiment analysis	2,000 / 1,200 / 1,200	28	Accuracy	Le et al. (2020),keung-etal-2020-multilingual
		PAWS-X	Paraphrase Identification	49,400 / 2,000 / 2,000	20/20	Accuracy	Yang et al. (2019),Le et al. (2020),Cañete et al. (2022)
		XNLI	Natural Language Inference	393,000 / 5,010 / 2,490	20/10	Accuracy	Le et al. (2020), Conneau et al. (2020),Cañete et al. (2022)
Hindi	Indo-Aryan	BBC NLI	Natural Language Inference	15,552 / 2,580 / 2,592	7/5	Accuracy	Uppal et al. (2020)
		News Topic	Topic classification	15,552 / 2,580 / 2,592	13	F1	Uppal et al. (2020)
		XNLI	Natural Language Inference	392,702 / 2,490 / 5,010	21/10	Accuracy	Conneau et al. (2020)

Table 11: Datasets summary. For tasks of two inputs, e.g. paraphrase identification tasks and inference tasks, their average text lengths are shown separately for the first input and the second input as *length 1 / length 2*

Dataset	Model	α Suff	$\alpha\nabla\alpha$ Suff	$x\nabla x$ Suff	IG Suff	DL Suff	α Comp	$\alpha\nabla\alpha$ Comp	$x\nabla x$ Comp	IG Comp	DL Comp	F1	Accuracy
SST	mBERT	1.2063	1.205	0.9991	1.3995	1.2594	1.2576	1.2643	1.0433	1.4835	1.3135	0.8627	0.8627
SST	XLM-R	1.0914	1.0976	1.0329	1.1125	1.0558	0.9242	0.9244	0.9537	1.0787	0.9878	0.8718	0.8719
SST	BERT	1.174	1.1771	1.0207	1.1636	1.0726	1.5571	1.5597	1.1582	1.6837	1.1955	0.9156	0.9156
SST	RoBERTa	1.2623	1.2693	1.3215	1.4922	1.1866	1.6021	1.6144	1.2723	1.438	1.3409	0.8893	0.8898
Agnews	mBERT	1.7087	1.712	0.9817	1.4523	1.0573	3.2063	3.203	1.8811	2.8304	1.5659	0.9303	0.9304
Agnews	XLM-R	2.0947	2.105	0.9287	1.4987	0.8806	2.0106	2.0107	1.2924	1.9369	1.1211	0.9261	0.9264
Agnews	BERT	1.1553	1.1266	0.9105	1.0425	1.0719	2.5436	2.5968	1.5426	2.4037	1.6445	0.9357	0.9357
Agnews	RoBERTa	1.3137	1.3242	0.8989	1.452	1.4351	2.1323	2.1408	1.66	1.9998	1.0854	0.9347	0.9346
MultiRC	mBERT	1.1821	1.177	0.9611	1.0904	0.9612	1.0	1.0011	1.0004	1.0031	1.0065	0.7081	0.7186
MultiRC	XLM-R	0.7907	0.829	0.9001	0.9677	1.0648	0.9247	0.9204	1.0124	1.0424	1.0109	0.718	0.7245
MultiRC	BERT	1.5089	1.512	1.0829	1.1752	0.9888	0.9959	0.9948	0.9978	0.9942	1.0022	0.6815	0.6896
MultiRC	RoBERTa	1.648	1.6946	0.9313	1.0268	1.3368	1.5195	1.4091	1.3068	1.6189	1.6841	0.7295	0.7317
KR	mBERT	1.1229	1.0541	1.1878	1.3514	1.1128	1.0077	1.0082	0.9979	0.9989	0.9966	0.842	0.8424
KR	XLM-R	1.4342	1.4154	0.8885	1.0773	0.938	0.9022	0.9014	1.0259	1.0089	1.0307	0.8401	0.8403
KR	BERT (zh)	1.239	1.2241	1.0296	1.0242	0.9226	1.0105	1.0157	0.996	0.9907	1.0165	0.8399	0.84
KR	RoBERTa (zh)	0.8657	0.8376	1.0082	0.9963	0.9782	0.9912	0.9932	0.9882	0.9901	0.9989	0.8443	0.8446
ANT	mBERT	1.0425	1.0471	0.9258	0.9767	0.8555	1.049	1.0455	1.0228	1.0208	1.0915	0.6282	0.703
ANT	XLM-R	1.0033	0.991	0.948	1.0205	1.0631	0.953	0.9601	0.9287	0.9879	1.0229	0.6588	0.7139
ANT	BERT (zh)	1.2248	1.2319	0.9675	1.0107	0.9884	1.0216	1.0212	1.0032	1.0105	1.0051	0.6738	0.7237
ANT	RoBERTa (zh)	1.0773	1.0945	1.0446	1.1371	1.1157	1.0063	1.0033	1.0057	1.0261	1.0252	0.5241	0.6601
ChnSentiCorp	mBERT	1.4906	1.4942	1.0566	1.325	1.0146	2.1555	2.1608	1.324	2.0856	1.0983	0.9119	0.9119
ChnSentiCorp	XLM-R	1.2483	1.2368	1.0077	1.055	0.9944	1.0723	1.0738	0.9931	1.1389	0.9942	0.9217	0.9217
ChnSentiCorp	BERT (zh)	1.2466	1.2516	1.0455	1.09	1.0243	1.548	1.5388	1.2609	1.5762	1.1181	0.9355	0.9356
ChnSentiCorp	RoBERTa (zh)	1.5482	1.5435	1.0476	1.1406	0.9543	1.6196	1.6116	1.2854	1.5884	1.2097	0.9428	0.9428
Spanish CSL	mBERT	1.5244	1.5274	1.0999	1.6256	1.1076	1.898	1.8972	1.2135	1.9047	1.2905	0.886	0.8862
Spanish CSL	XLM-R	1.1065	1.0896	0.9543	1.1994	1.0514	0.986	0.9887	0.9715	1.1801	0.9913	0.878	0.8782
Spanish CSL	BERT (es)	0.9975	0.976	0.9957	1.1277	1.0645	1.0698	1.0788	1.0955	1.4271	1.0004	0.9062	0.9063
Spanish CSL	RoBERTa (es)	1.2901	1.4932	1.5522	1.5633	1.5125	1.5761	1.3826	1.3995	1.0484	1.5366	0.8914	0.8917
Spanish XNLI	mBERT	1.0031	1.0043	1.0258	1.0382	1.0331	1.0165	1.0164	0.9964	1.0028	0.9872	0.7877	0.7875
Spanish XNLI	XLM-R	1.0314	1.0457	1.0887	1.0738	1.0521	1.0485	1.0479	1.0285	1.0469	0.9918	0.7958	0.7956
Spanish XNLI	BERT (es)	1.0791	1.0922	1.037	1.0228	1.0331	1.0327	1.03	0.9938	1.0017	0.9721	0.7847	0.7842
Spanish XNLI	RoBERTa (es)	1.3083	1.3127	1.5799	1.1294	0.9508	1.102	1.1	1.0525	1.0146	1.0096	0.7958	0.7956
Spanish Paws	mBERT	1.1325	1.1348	0.9959	0.9616	0.9826	0.994	0.9952	0.9968	0.9999	1.0062	0.8811	0.8823
Spanish Paws	XLM-R	1.1797	1.1944	1.0948	1.0857	0.9884	1.2369	1.2376	1.0415	1.0452	0.987	0.8703	0.872
Spanish Paws	BERT (es)	0.9825	0.9919	1.0713	0.9047	0.9792	1.0016	0.997	0.9985	0.9988	0.9965	0.8555	0.8565
Spanish Paws	RoBERTa (es)	0.9294	0.9379	1.0151	0.9883	0.9621	1.1832	1.1391	0.9047	1.1132	1.0781	0.8823	0.883
French CSL	mBERT	1.4165	1.413	0.9956	1.4875	1.1035	2.1526	2.1624	1.1415	2.0983	1.3063	0.8772	0.8773
French CSL	XLM-R	1.1488	1.16	0.9952	1.0022	1.0042	0.9769	0.9721	1.0087	1.1822	0.9862	0.8863	0.8865
French CSL	BERT (fr)	1.0753	1.0857	0.9524	1.2311	0.8271	1.2186	1.211	0.9881	1.4274	0.852	0.8824	0.8825
French CSL	RoBERTa (fr)	1.3471	1.3482	1.1526	1.4631	1.4639	2.0347	2.0311	1.4313	2.5163	2.3467	0.8663	0.8668
French XNLI	mBERT	1.0997	1.0732	1.0201	1.1127	1.0719	1.0147	1.0175	0.9985	1.0194	1.0179	0.7748	0.7746
French XNLI	XLM-R	1.0058	0.9517	1.1456	1.0234	1.0441	1.0544	1.0577	1.0324	1.027	0.9889	0.789	0.7885
French XNLI	BERT (fr)	0.9795	0.9862	1.0337	1.0762	1.0819	1.0503	1.0484	1.0077	1.0389	0.9974	0.7643	0.7638
French XNLI	RoBERTa (fr)	1.5508	1.5543	1.4092	1.183	1.1098	1.527	1.5246	1.2518	1.0796	0.9975	0.7326	0.7323
French Paws	mBERT	1.1789	1.1849	0.9801	0.9469	0.8695	0.9808	0.9798	0.9963	0.998	1.0062	0.8781	0.8788
French Paws	XLM-R	1.087	1.1021	1.0529	1.0192	0.9929	1.2295	1.2255	1.0622	0.997	1.0263	0.8774	0.8778
French Paws	BERT (fr)	1.0875	1.0796	1.0948	1.0518	1.0596	0.9987	1.0022	1.0028	0.9994	1.0144	0.8274	0.8297
French Paws	RoBERTa (fr)	0.9629	0.9655	1.0318	1.0507	1.0304	1.1575	1.1452	1.1168	1.4052	1.0831	0.7729	0.8678
Hindi BBC Nli	mBERT	1.1255	1.1278	1.1362	1.175	1.0102	1.0044	1.0039	1.003	0.998	1.005	0.7862	0.7864
Hindi BBC Nli	XLM-R	1.1809	1.1789	1.0289	1.0762	1.0578	1.18	1.19	1.0317	1.0842	1.0125	0.7887	0.7888
Hindi BBC Nli	BERT (hi)	0.9799	0.9779	1.0385	1.0574	1.0385	1.0122	1.016	0.9989	1.0046	1.0045	0.8124	0.8128
Hindi BBC Nli	RoBERTa (hi)	1.0349	1.0225	0.9337	0.9863	0.9436	0.6561	0.6876	1.1159	1.0714	0.9546	0.7953	0.8094
Hindi BBC Topic	mBERT	1.4883	1.4913	1.2533	1.3573	0.984	1.4896	1.4907	1.2431	1.2887	1.0935	0.5123	0.5918
Hindi BBC Topic	XLM-R	1.1243	1.1513	1.0942	1.2419	1.1083	1.1409	1.1413	1.0351	1.1042	1.0049	0.5606	0.6425
Hindi BBC Topic	BERT (hi)	1.8746	1.8729	1.498	0.9446	1.0336	2.1692	2.1703	1.6329	0.8877	0.8943	0.617	0.6753
Hindi BBC Topic	RoBERTa (hi)	0.9569	0.9527	0.9921	1.2189	0.9464	0.9823	0.9841	1.04	1.4999	0.9481	0.5268	0.6395
Hindi XNLI	mBERT	1.1363	1.1501	1.2088	1.3084	1.071	1.0187	1.0159	1.0147	1.0359	0.9775	0.6754	0.676
Hindi XNLI	XLM-R	1.0099	0.9844	0.985	1.0652	0.9954	1.1142	1.1161	1.0214	1.0578	1.0056	0.7235	0.7237
Hindi XNLI	BERT (hi)	1.0199	1.0234	1.0304	1.0419	1.0032	1.0266	1.0248	1.0126	1.0165	1.0048	0.6607	0.6607
Hindi XNLI	RoBERTa (hi)	1.4853	1.4795	1.0466	1.3833	1.0287	1.5834	1.589	1.0399	1.4781	0.8741	0.6316	0.6314
Average faithfulness across	datasets and models	1.210	1.213	1.062	1.155	1.049	1.299	1.295	1.115	1.284	1.097		

Table 12: Full results of faithfulness and prediction performance. All faithfulness results are presented by being divided by the random baseline.

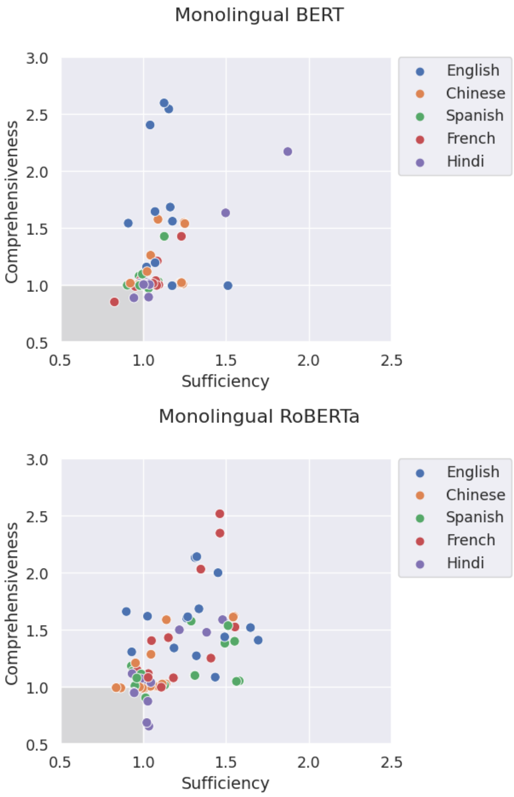

Figure 2: Faithfulness results for different languages on monolingual models.

E RoBERT v.s. BERT

E.1 The language distribution comparison of the pre-training corpus between mBERT and XLM-R

Figure 3 on page 16 compares the data amount and distribution in different languages between multilingual BERT (mBERT) and multilingual RoBERTa (XLM-R). As shown in the figure, XLM-R has significantly increased the pre-training data amount by several orders of magnitude in all languages. It has also increased the percentages of non-English data.

E.2 Full results of faithfulness for XLM-R large

Table 13 on page 16 presents the original sufficiency and comprehensiveness results of each feature attribution method on each task for XLM-R large. It was used in Section 5 to investigate the impact of the model size gap on the faithfulness disparity.

F Exploring the impact of model size on BERT

The results indicate a lower faithfulness on the larger BERT model across FAs and tasks. Specifically, the sufficiency and comprehensiveness of the monolingual English BERT-large are higher than its counterpart BERT-base (13 out of 16 comparison pairs as shown in Table 1), except for cases of sufficiency and comprehensiveness on IG and the comprehensiveness on MultiRC (where both base and large BERTs' faithfulness are on par with the random baseline, values close to one). This observation agrees with our assumption above that model sizes might impact faithfulness disparity. Given that our focus is on faithfulness disparity, we leave a more in-depth and comprehensive study with specifically designed methods in the future for the impact of model size on faithfulness.

G The tokenization for different languages

All monolingual and multilingual BERT tokenizers in this paper use "##" to indicate the second and the rest subwords of a split word, i.e. non-first subword of a split word. For example, "sdfnsksi cklx" will be tokenize to 'sd', '##fn', '##sk', '##si', 'ck', '##l', '##x'.

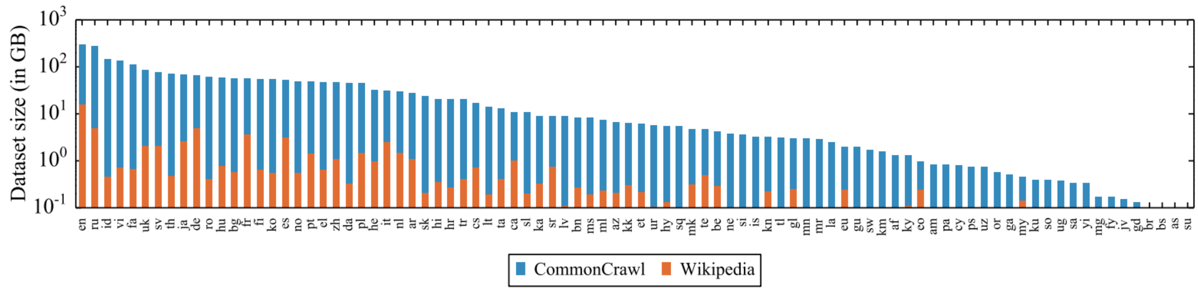

Figure 3: Amount of data in GiB (log-scale) for the 88 languages that appear in both the Wiki-100 corpus (used for multilingual BERT) and the CC-100 (multilingual RoBERTa). CC-100 increases the amount of data by several orders of magnitude, in particular for low-resource languages (Conneau et al., 2020).

Dataset	Model	α Suff	$\alpha\nabla\alpha$ Suff	$x\nabla x$ Suff	IG Suff	DL Suff	α Comp	$\alpha\nabla\alpha$ Comp	$x\nabla x$ Comp	IG Comp	DL Comp
SST	XLM-R large	0.9555	0.9547	1.0189	0.7746	1.0062	0.9437	0.9382	1.1265	0.6697	1.0576
Agnews	XLM-R large	1.1866	1.2698	0.7601	0.8642	0.9089	2.8766	2.6539	1.3965	1.3442	1.0955
MultiRC	XLM-R large	1.0007	1.0004	1.0007	0.9967	1.4006	0.8311	0.6314	0.6188	3.2761	0.6126
KR	XLM-R large	1.1857	1.1985	1.0159	0.9569	0.9741	1.0487	1.0408	1.0543	1.1403	1.0179
ANT	XLM-R large	1.0355	1.0395	0.9159	0.7393	1.0027	1.0278	1.0178	0.887	0.6333	1.0025
ChnSentiCorp	XLM-R large	0.8405	0.8372	1.044	0.918	0.9405	0.7424	0.7871	1.1985	0.9229	1.0699
Spanish CSL	XLM-R large	1.2667	1.2688	0.9961	0.9862	1.0137	1.2989	1.304	1.0417	1.0722	1.0519
Spanish XNLI	XLM-R large	0.8986	0.8959	1.0609	0.9614	0.9873	0.8655	0.8668	1.1609	1.0007	1.0213
Spanish Paws	XLM-R large	0.8478	0.8579	1.0342	0.9004	0.9432	1.1443	1.1448	1.1444	1.0152	1.0204
French CSL	XLM-R large	1.0388	1.0278	1.1031	1.0849	1.0313	1.0364	1.0361	1.0631	1.1244	1.0435
French XNLI	XLM-R large	1.0388	1.0403	1.079	0.9644	0.9943	1.0899	1.085	1.1397	1.0307	1.0227
French Paws	XLM-R large	0.8575	0.8583	1.051	0.9031	1.0132	1.1394	1.1289	1.2237	0.9642	1.0129
Hindi BBC Nli	XLM-R large	0.8731	0.8478	1.0379	1.0646	0.9734	0.7646	0.7796	1.0062	1.0786	1.0222
Hindi BBC Topic	XLM-R large	1.6458	1.6491	0.9722	0.8833	1.0009	1.7309	1.7246	0.9697	0.9469	1.0661
Hindi XNLI	XLM-R large	0.9875	0.995	1.0806	0.9227	0.947	1.0358	1.0309	1.1539	0.9326	0.9913

Table 13: Full results of faithfulness for XLM-R large. All faithfulness results are presented by being divided by the random baseline.

Sufficiency								
	α	$\alpha\nabla\alpha$	$x\nabla x$	IG	DL	SST	Agnews	MultiRC
BERT base (109M)	1.279	1.272	1.005	1.127	1.044	1.122	1.061	1.253
BERT large (340M)	1.045	1.037	1.005	1.158	1.025	1.017	1.041	1.105
Comprehensiveness								
	α	$\alpha\nabla\alpha$	$x\nabla x$	IG	DL	SST	Agnews	MultiRC
BERT base (109M)	1.699	1.717	1.233	1.694	1.281	1.431	2.146	0.997
BERT large (340M)	1.564	1.581	1.134	1.731	1.053	1.270	1.963	1.005

Table 14: Sufficiency and comprehensiveness of BERT-base and BERT-large models averaged on each FA (the first two to seven columns from left) and each task (the last three columns from right).

Monolingual RoBERTa indicates a space and its following word with 'ă'. Therefore, except for the first token, tokens without 'ă' are subwords. Multilingual RoBERTa uses "_" to indicate the start of a whole word.

H Full results for fertility and splitting ratio

Table 15 includes the full results of fertility and splitting ratio for each model. The results here are used for calculating the average values demonstrated in Table 7.

I Disparity in tokenization aggressiveness

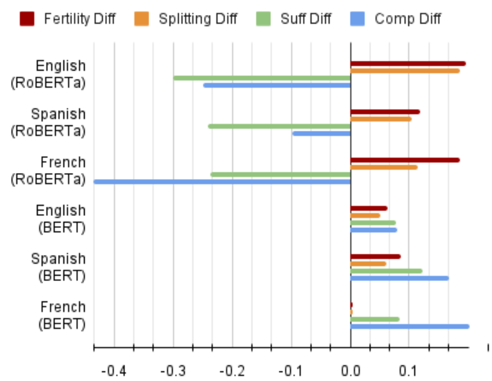

Figure 4: The impact of tokenization aggressiveness ("Fertility Diff" and "Splitting Diff") on faithfulness disparity ("Suff Diff" and "Comp Diff").

Figure 4 demonstrates the difference between multi- and monolingual models in terms of tokenization aggressiveness and faithfulness. Both are calculated as: the score of the multilingual model minus the corresponding score of the monolingual counterpart model. We observe that multilingual models consistently tokenize more aggressively

	RoBERTa			BERT			RoBERTa			BERT		
Dataset	Multi Fertility	Mono Fertility	Fertility Diff	Multi Fertility	Mono Fertility	Fertility Diff	Multi Splitting Ratio	Mono Splitting Ratio	Splitting Diff	Multi Splitting Ratio	Mono Splitting Ratio	Splitting Diff
SST	1.2941	1.1327	0.1615	1.2229	1.1237	0.0992	0.2358	0.0893	0.1466	0.1674	0.0863	0.0811
Agnews	1.3392	1.1519	0.1873	1.1780	1.1325	0.0455	0.2724	0.0765	0.1959	0.0884	0.0504	0.0380
MultiRC	1.3250	1.0901	0.2350	1.1365	1.0890	0.0475	0.2734	0.0618	0.2116	0.0768	0.0397	0.0371
Spanish CSL	1.3418	1.2018	0.1399	1.3796	1.2138	0.1658	0.2587	0.1596	0.0991	0.1716	0.0618	0.1098
Spanish PAWS-X	1.4706	1.4286	0.0419	1.3605	1.4034	-0.0429	0.3203	0.2441	0.0762	0.1303	0.1406	-0.0103
Spanish XNLI	1.4134	1.2387	0.1747	1.3679	1.2317	0.1362	0.3173	0.1819	0.1355	0.1543	0.0675	0.0868
French CSL	1.4511	1.3134	0.1377	1.4668	1.3768	0.0900	0.2921	0.1904	0.1016	0.1553	0.1091	0.0462
French PAWS-X	1.5818	1.3652	0.2166	1.4257	1.5555	-0.1298	0.3511	0.2195	0.1316	0.1257	0.1921	-0.0664
French XNLI	1.5598	1.3557	0.2041	1.4912	1.4353	0.0558	0.3307	0.2233	0.1074	0.1358	0.1011	0.0347

Table 15: Fertility and splitting ratio of multilingual and monolingual RoBERTa and BERT on tasks.

than their monolingual counterparts. When the fertility of the multilingual model is higher than its monolingual by more than 0.1, the multilingual model gains lower faithfulness than its monolingual counterpart model.