Comparing Explanation Faithfulness between Multilingual and Monolingual Fine-tuned Language Models

Anonymous ACL submission

Abstract

In many real natural language processing application scenarios, practitioners not only aim to maximize predictive performance but also seek faithful explanations for the model predictions. Rationales and importance distribution given by feature attribution methods (FAs) provide insights into how different parts of the input contribute to a prediction. Previous studies have explored how different factors affect faithfulness, mainly in the context of monolingual English models. On the other hand, the differences in FA faithfulness between multilingual and monolingual models have yet to be explored. Our extensive experiments, covering five languages and five popular FAs, show that FA faithfulness varies between multilingual and monolingual models. We find that the larger the multilingual model, the less faithful the FAs are compared to its counterpart monolingual models. Our further analysis shows that the faithfulness disparity is potentially driven by the differences between model tokenizers.[1]

1 Introduction

Feature attribution methods (FAs) are commonly used for ranking input 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 token importance scores and selected rationales actually reflect the model's inner reasoning mechanism (Jacovi and Goldberg, 2020).

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., 2022). Specifically, monolingual studies have investigated the impact of out-of-domain data (Chrysostomou and Aletras, 2022), adversarial attacks (Sinha et al., 2021; Zhao et al., 2022a) and temporal shifts (Zhao et al., 2022b) on the faithfulness of FAs. On the other hand, existing studies on interpreting multilingual models' behavior and their representations (Rama et al., 2020; Serikov et al., 2022; Gonen et al., 2022) have not investigated the faithfulness of FAs.

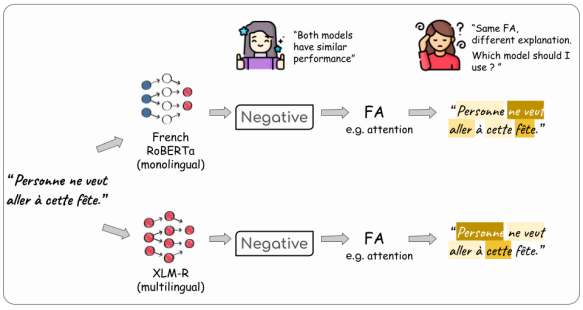

Figure 1: Model explanations given by the same feature attribution method, e.g. attention, for multilingual (XLM-R) and monolingual (French RoBERTa) models for the same task (sentiment analysis in FR).

As shown in Figure 1, even for the same input ("Personne ne veut aller à cette fête.", i.e. "Nobody wants to go to this party." in English), model prediction and FA, the token importance scores can be substantially different between multi- and monolingual models. This indicates that the models 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 of FAs when applied to multi- and monolingual models.*

Our main contributions are as follows:

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

- We perform a large empirical study across tasks in five languages, five popular FAs and two groups 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. FAs tend to give less faithful rationales for larger multilingual models, compared to their monolingual counterparts;
- Our analysis further shows that the discrepancies in faithfulness are potentially driven by differences in tokenization rather than how these models semantically process the input. For example, the more aggressive tokenization results in a larger faithfulness discrepancy between mono- and multilingual models.

2 Related Work

Faithfulness of monolingual models Faithfulness measures if a rationale extracted with a given FA, accurately reflects the model's internal reasoning process (Ribeiro et al., 2016; DeYoung et al., 2020; Jacovi and Goldberg, 2020; Pezeshkpour et al., 2021).[2]

On the one hand, 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 (2022) investigated the impact of out-of-domain data on faithfulness, while Zhao et al. (2022b) studied the effect of temporal concept drift on faithfulness. On the other hand, an increasing number of multilingual language models are made available for different languages (Antoun et al., 2020; Chan et al., 2020; Cañete et al., 2020; Le et al., 2020), but there is no empirical evidence that a FA is equally faithful between monolingual models and their counterpart multilingual models.

Interpretability of multilingual models Previous studies on multilingual models focus on probing or analyzing their 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 human focus 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 for multilingual models. It assumes 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 without comparing mono- and multilingual models.

Performance comparison of monolingual and multilingual models Previous work has compared the performance of monolingual and multilingual language models focusing on mBERT and BERT variants (Rönnqvist et al., 2019; Nozza et al., 2020; Vulić et al., 2020; Rust et al., 2021). Vulić et al. (2020) specifically investigated how lexical knowledge extraction strategies impact performance between mono- and multilingual models, while Rust et al. (2021) further investigated the impact of tokenizers. 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 phenomenon of overall performance decrease on monolingual as well as cross-lingual tasks beyond a certain number of languages. To the best of our knowledge, no study has investigated how the curse of multilinguality impacts FAs' faithfulness.

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

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, IndicNLP	1.5B	52K	83M

Table 1: Overview of models across languages.

3 Experiments

Our aim is to compare FA 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 differences between mono- and multilingual models are the tokenizers, supported vocabularies and the pre-training corpora. Using this setting allows for a realistic comparison between models, given the fact that in a real world scenario, a practitioner would choose between off-the-shelf, already pre-trained mono- or multilingual models without considering any specific implementation details (e.g. pre-training data). We compare models in various downstream tasks across a spectrum of typologically diverse and widely spoken languages, i.e. English, Chinese, Spanish, French and Hindi.

3.1 Models

Multilingual models. We use two popular multilingual models: (1) **mBERT**, a multilingual version of BERT (Devlin et al., 2019) trained on text from 104 languages in Wikipedia; and (2) **XLM-R** (Conneau et al., 2020), a multilingual version of RoBERTa (Liu et al., 2019) trained on text from 100 languages in the Common Crawl corpus.

Monolingual models. For each language, we include its corresponding **monolingual BERT** and **RoBERTa** models respectively. Table 1 provides an overview of all models across languages.

3.2 Datasets

Due to the lack of identical datasets in multiple languages, we include a variety of tasks that are similar. We experiment with (1) sentiment analysis; (2) topic classification; (3) reading comprehension; (4) paraphrase identification; and natural language inference. Further details of datasets and their preprocessing are presented in Appendix C.

3.3 Implementation details

We fine-tune each model using the hyperparameters from the original papers describing the corresponding models and tasks. If these are not available, we use a batch size of 16 and a learning rate of 1e-5 with an early stopping over five epochs. Full implementation details are given in Appendix B.

3.4 Feature attribution methods

We experiment with five popular FAs since it has been shown that 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 their faithfulness between mono- and multilingual models across different languages and tasks: (1) **Attention** (α) computes the token importance using the corresponding normalized attention score (Jain et al., 2020); (2) **Scaled attention** ($\alpha\nabla\alpha$) is similar to attention scores but scaled by their corresponding gradients (Serrano and Smith, 2019); (3) **InputXGrad** ($x\nabla x$) attributes importance by multiplying the input with its gradient computed with respect to the predicted class (Kindermans et al., 2016; Atanasova et al., 2020); (4) **Integrated Gradients (IG)** 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); and (5) **DeepLift (DL)** computes 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

We compare the faithfulness of a given FA by applying it to mono- and multilingual models.

Hard Sufficiency & Comprehensiveness. Sufficiency (Suff) and comprehensiveness (Comp) are two commonly-used metrics for evaluating faithfulness (DeYoung et al., 2020) using hard input perturbation. **Suff** captures the difference in predictive likelihood between retaining only the rationale. **Comp** assesses how much information the rationale holds by measuring changes in predictive likelihood when entirely removing the rationale. Their

		BERT & mBERT					RoBERTa & XLM-R				
Lang	Model	Acc	Suff	Comp	S-Suff	S-Comp	Acc	Suff	Comp	S-Suff	S-Comp
EN	Mono	**0.847**	1.146	1.525	1.172	1.201	**0.852**	**1.306**	**1.588**	1.207	**1.200**
	Multi	0.837	**1.224**	**1.604**	**1.180**	**1.204**	0.841	1.163	1.210	**1.220**	1.195
ZH	Mono	**0.833**	1.101	1.142	**1.012**	0.995	0.816	**1.093**	**1.156**	0.990	**1.004**
	Multi	0.819	**1.137**	**1.271**	0.990	**1.001**	**0.825**	1.088	1.000	**1.041**	0.999
ES	Mono	0.849	1.024	1.046	**1.148**	1.150	**0.857**	**1.235**	**1.176**	**1.141**	**1.182**
	Multi	**0.852**	**1.146**	**1.214**	1.130	**1.152**	0.849	1.082	1.055	1.129	1.148
FR	Mono	0.825	1.047	1.057	**1.099**	1.100	0.822	**1.242**	**1.510**	**1.087**	1.095
	Multi	**0.844**	**1.130**	**1.259**	1.096	**1.102**	**0.851**	1.049	1.055	1.083	**1.099**
HI	Mono	**0.716**	1.162	**1.177**	0.984	1.001	0.693	**1.094**	**1.097**	1.013	**1.012**
	Multi	0.685	**1.202**	1.157	**1.013**	1.001	**0.718**	1.086	1.084	**1.040**	0.998

Table 2: Predictive performance (Accuracy) and FA faithfulness (Suff, Comp, Soft-Suff, Soft-Comp) of mono- (BERT, RoBERTa) and multilingual models (mBERT, XLM-R). Full results per task and per FA are in Appendix D.

normalized versions allow for a fairer comparison across models (Carton et al., 2020). Following DeYoung et al. (2020), we use the Area Over the Perturbation Curve (AOPC) for normalized Suff and Comp across different rationale lengths (10%, 20%, and 50%) by taking the average, similar to DeYoung et al. (2020) and Chan et al. (2022).[3]

Soft Sufficiency & Comprehensiveness. Soft sufficiency (**Soft-Suff**) and comprehensiveness (**Soft-Comp**) use a soft input perturbation criterion to measure faithfulness (Zhao and Aletras, 2023). Each token is perturbed proportionally to its importance score assigned by a FA instead of being fully retained or removed. The 'soft' version of these metrics has been found to be more robust compared to their 'hard' counterparts.

The final scores for the four metrics are computed after being divided by their corresponding random baseline. Therefore, values greater than one 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) on three tasks, using five FAs in each language. This results in 480 faithfulness evaluation cases for each language, 2400 cases for five languages in total. We report accuracy and F1 of all models in Appendix D.

Predictive Performance Table 2 shows the predictive performance of models and faithfulness scores across FAs, averaged on three tasks for each language. Overall, we observe that the performance of mono- and multilingual models is consistently comparable to each other which demonstrates the importance of comparing FA faithfulness between these two types of models. For instance, the difference between Spanish BERT and mBERT is 0.003. The largest gap is found between Hindi BERT (0.716) and mBERT (0.685), exhibiting a difference of 0.031.

Faithfulness We note that FAs demonstrate inconsistent faithfulness discrepancies between mono- and multilingual models. In general, FAs obtain lower Suff and Comp with XML-R than monolingual RoBERTa models, and higher with mBERT compared to monolingual BERT models (except for Comp in Hindi). These discrepancies do not manifest though when we use Soft-Suff and Soft-Comp to measure faithfulness. In fact, the faithfulness of FAs between mono- and counterpart multilingual models are comparable to each other, i.e. the majority of differences are smaller than 0.01. For example, the greatest difference is only 0.051 (Soft-Comp between Chinese RoBERTa and XLM-R).

Moreover, the faithfulness disparity in Suff and Comp of RoBERTa-based models is more noticeable as half of the cases have a faithfulness difference greater than 0.1. For example, the Comp in French is 1.51 for French RoBERTa but only 1.055 for XLM-R. We further investigate these differences in faithfulness across metrics in Section 5.

Comparing FAs Figure 2 delves deeper into the faithfulness disparity of FAs by looking at each one separately. Disparity is computed as the faith-

[3]For tasks of average token length over 200, we evaluate rationale ratios of 1%, 5%, and 10% instead, to keep the rationales relatively short.

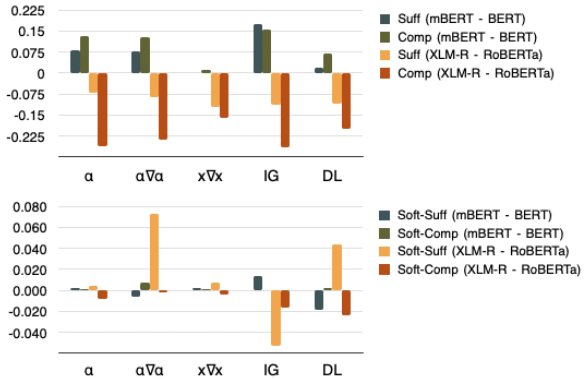

Figure 2: Faithfulness disparity of FAs averaged across languages. Values above zero indicate that the FAs are more faithful in the multilingual model.

fulness score on the multilingual model minus the faithfulness score on the monolingual counterpart.[4]

First, Figure 2 (top) shows contrasting directions of faithfulness disparities between RoBERTa-based and BERT-based models. That is, FAs exhibit lower faithfulness in Suff and Comp for XLM-R than monolingual RoBERTa (above zero), whereas FAs exhibit higher Suff and Comp for mBERT than monolingual BERT. This holds true across FAs and languages as shown in Tables 11 and 12 in Appendix E. Second, FAs do not show a consistent faithful disparity in terms of Soft-Suff and Soft-Comp (bottom) applied to mono- and multilingual models.

Moreover, we observe that IG has a larger faithfulness disparity than other FAs. For example, this is evident in Suff and Comp averaged over languages for both RoBERTa- and BERT-based models. IG is also the only FA showing significant differences in Suff and Comp disparity on both BERT- and RoBERTa-based models (see Table 11 and Table 12 in Appendix). This means, that when IG selects faithful rationales for a monolingual model, it might not be able to do so for the counterpart multilingual model. We also notice that the disparities of Attention-based FAs, i.e. α and $\alpha\nabla\alpha$, are consistently on par with each other. However, they demonstrate larger disparities compared to $x\nabla x$ and DL in most cases. Therefore, this indicates that mono- and multilingual models tend to employ attention differently for making predictions.

On the other hand, the results of Soft-Suff and Soft-Comp do not echo any of the above observations we made for Suff and Comp. Indeed, the Soft-Suff and Soft-Comp disparities of FAs are not significantly different between mono- and multilingual models (see p-values in Appendix E). That is, when evaluating the overall importance distribution given by a FA, there is no faithfulness discrepancy between using a multilingual or a monolingual model. One possible reason is that, the pre-defined rationale lengths introduce bias to the evaluation of Suff and Comp. To investigate this, we examine the faithfulness disparity on specific rationale lengths. We find that the faithfulness disparity varies across rationale lengths. Table 3 shows the Suff disparity at different rationale lengths, 10% and 50%, within BERT-based models. For example, FAs are more faithful with XLM-R than Spanish RoBERTa at a 10% rationale length, with a significant average difference of 0.236 (p-value of 0.03) across tasks. When looking at the rationale length of 50%, FAs are comparably faithful for XLM-R and Spanish RoBERTa (avg of 0.027, p-value of 0.245). This further inspires us to consider the impact of the tokenizers on faithfulness. Soft-Suff and Soft-Comp evaluate the importance scores over the whole input and this can be less sensitive to the tokenization. We examine this in Section 5.

5 Analysis

We further investigate if the tokenization and model size contribute to the contrasting directions of FA faithfulness disparity between BERT- and RoBERTa-based models.

Sufficiency at 10%							
	α	$\alpha\nabla\alpha$	$x\nabla x$	IG	DL	Avg Diff	P value
English	0.225	0.232	-0.021	0.168	0.035	0.128	0.202
Chinese	0.037	0.007	0.079	0.285	0.054	0.093	0.228
Spanish	0.41	0.407	0.084	0.346	-0.067	0.236	0.03
French	0.256	0.236	-0.011	0.21	-0.047	0.129	0.042
Hindi	-0.198	-0.185	-0.06	0.244	-0.075	-0.055	0.575
Avg Diff	0.146	0.139	0.014	0.251	-0.02	0.106	
P value	0.224	0.255	0.747	0.004	0.666		0.008
Sufficiency at 50%							
	α	$\alpha\nabla\alpha$	$x\nabla x$	IG	DL	Avg Diff	P value
English	-0.146	-0.15	-0.022	0.05	0.058	-0.042	0.3
Chinese	0.061	0.062	0.045	0.15	-0.042	0.055	0.271
Spanish	0.033	0.035	-0.026	0.075	0.019	0.027	0.245
French	0.096	0.09	-0.021	-0.046	0.059	0.036	0.313
Hindi	0.026	0.031	0.049	0.251	0.02	0.075	0.094
Avg Diff	0.014	0.014	0.005	0.096	0.023	0.03	
P value	0.752	0.765	0.831	0.039	0.51		0.082

Table 3: Sufficiency difference between mBERT and counterpart monolingual BERT on rationale ratio of 10% and 50%. Plum indicates that monolingual models are more faithful than multilingual models.

[4]Tables 12, 11, 14, and 13 in Appendix E show details per FA and language.

5.1 Impact of model size

We posit that the difference between RoBERTa and BERT-based models in disparity directions of FAs faithfulness is associated with the differences in model sizes of 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-based models. 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.

To further investigate the impact of the model size, we repeat all experiments using XLM-R large and compare its faithfulness with monolingual RoBERTa. Full results are in Table 16 and Table 15 in Appendix. In this case, we examine a RoBERTa-based model pair with a larger difference in model size than XLM-R base vs. 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[5] (Conneau et al., 2020). XLM-R large (550M parameters) is at least 4.7 times larger than the monolingual RoBERTa models.

The results first show that the faithfulness disparity direction remains the same as the one between XLM-R base and monolingual RoBERTa. This implies that FAs are more faithful with monolingual RoBERTa. 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 by remaining 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).

Overall, the results confirm our assumption that the difference in model size is related to the faithfulness 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. To summarize, *the more parameters the multilingual model has, the less faithful its rationales are compared to its monolingual counterparts.*

We acknowledge that our findings might not generalize to BERT because mBERT of different sizes are not available to experiment with. To overcome this, we repeat all experiments with BERT-large and compare its faithfulness with BERT-base, to investigate the impact of model size from a different perspective. The results show that FAs obtain lower Suff and Comp with the larger BERT model across FAs and tasks. This observation agrees with our assumption above that model sizes might impact faithfulness disparity. To keep the focus of the paper on the faithfulness disparity between mono- and multilingual models, we present the results and analysis in Table 17 in the Appendix.

5.2 Impact of tokenization

Previous research has shown the impact of tokenization 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, 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 (see Table 1). BERT-based models use WordPiece as their tokenizers (Wu et al., 2016). Monolingual RoBERTa models use BytePair-Encoding (BPE) (Sennrich et al., 2016), and the multilingual XLM-R uses SentencePiece (Kudo and Richardson, 2018).

Therefore, we investigate the impact of tokenization on faithfulness disparity. The effectiveness of a tokenizer in text splitting intuitively reflects how many unique tokens it knows in a particular language. Following Rust et al. (2021), we examine two metrics for assessing tokenization, fertility and splitting ratio. Fertility indicates how many subwords a tokenizer splits a word into, while the splitting ratio shows how often a tokenizer splits words. Intuitively, low scores are preferable for both metrics indicating that the tokenizer is well-suited to the language (Rust et al., 2021).

Table 4 shows the fertility and splitting ratio difference between monolingual and multilingual models (i.e. multilingual score minus its counter-

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

RoBERTa								
	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
BERT								
	Multi Fertility	Mono Fertility	Fertility Diff	Multi Splitting	Mono Splitting	Splitting Diff	Suff Diff	Comp Diff
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

Table 4: Fertility, splitting ratio, sufficiency, and comprehensiveness difference between multilingual ("Multi Fertility", "Multi Splitting") and monolingual models ("Mono Fertility", "Mono Splitting"). For "Suff Diff" and "Comp Diff", positive values indicate that the FA is more faithful to the multilingual model. Full results of fertility and splitting ratio for each dataset can be found in Table 18 in Appendix I.

part monolingual).[6] Faithfulness disparity values are taken from Tables 11 and 12.

First, for both RoBERTa and BERT-based models, the positive values of fertility and splitting ratio difference indicate that multilingual models tend to be more aggressive in splitting words than monolingual ones. For example, as shown in Table 18 in Appendix I, 26.1% English words (underlined in table) are split by the SentencePiece tokenizer of XLM-R but only 7.6% (underlined in table) by BPE which is used in monolingual RoBERTa models.

Second, RoBERTa-based models has larger gaps in both fertility and splitting ratio than BERT-based across all three languages. The fertility and the splitting ratio differences are greater than 0.1 for RoBERTa-based, but less than 0.1 for BERT-based. This is because SentencePiece (multilingual XLM-R's tokenizer) is generally more aggressive in splitting words. Taking English as an example, the fertility gaps among monolingual RoBERTa (BPE), monolingual BERT (WordPiece) tokenizers, and multilingual BERT (WordPiece) are 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 of XLM-R, over two times bigger than multilingual BERT (see Figure 1). 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 XLM-R potentially explains the different disparity direction in Suff and Comp to BERT models. That is, only when the fertility difference is greater than 0.1, FAs are more faithful with multilingual models than with monolingual counterparts.[7]

Last, the differences in Suff and Comp demonstrate a highly negative relationship to the fertility difference (Table 19 in Appendix I.2). 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).

To sum up, *multilingual tokenizers split words into subwords more aggressively than monolingual tokenizers, which potentially contribute to the faithfulness disparity between models. The aggressive tokenization of multilingual models might result in lower faithfulness.*

5.3 Disentangling the impact of the model

To further investigate how tokenization and model selection affects faithfulness, we experiment with (1) adapting a monolingual model to a different language (i.e. EN to FR); and (2) adapting a multilingual model to FR. This allows us to disentangle the impact of the model itself while observing the faithfulness changes of FAs across tokenization strategies. We experiment with RoBERTa-based models (RoBERTa and XLM-R) in French because this was the case where we observed the greatest faithfulness discrepancy between a mono- and multilingual model (see Table 12).

We use WECHSEL (Minixhofer et al., 2022) to adapt an English RoBERTa to French. We replace the tokenizer of the English RoBERTa with the tokenizer of the French RoBERTa. French token embeddings are initialized such that they are semantically similar to the English tokens by using multilingual static word embeddings covering English

[6] Hindi and Chinese are excluded from this analysis. For Hindi, we do not observe a substantial difference between mono- and multilingual models in Suff or Comp. Chinese is a logographic language without white spaces.

[7] We further demonstrate this pattern in Figure 4 in Appendix J.

	Sufficiency					Soft Sufficiency				
	α	$\alpha\nabla\alpha$	$x\nabla x$	IG	DL	α	$\alpha\nabla\alpha$	$x\nabla x$	IG	DL
RoBERTa (FR)	1.287	1.289	1.198	1.232	1.201	1.091	1.033	1.106	1.087	1.120
RoBERTa (EN → FR)	1.241	1.243	1.193	1.225	1.197	1.090	1.029	1.098	1.083	1.122
XLM-R (Multi → FR)	1.230	1.200	1.242	1.246	1.032	1.088	1.032	1.102	1.086	1.120
XLM-R	1.081	1.071	1.065	1.015	1.013	1.046	1.144	1.012	1.100	1.114
	Comprehensiveness					**Soft Comprehensiveness**				
	α	$\alpha\nabla\alpha$	$x\nabla x$	IG	DL	α	$\alpha\nabla\alpha$	$x\nabla x$	IG	DL
RoBERTa (FR)	1.573	1.567	1.267	1.667	1.476	1.097	1.093	1.090	1.102	1.092
RoBERTa (EN → FR)	1.305	1.317	1.321	1.474	1.309	1.099	1.09	1.087	1.099	1.091
XLM-R (Multi → FR)	1.394	1.401	1.266	1.435	1.353	1.097	1.090	1.089	1.100	1.091
XLM-R	1.087	1.085	1.035	1.069	1.001	1.100	1.097	1.099	1.101	1.097

Table 5: Faithfulness of French RoBERTa, XLM-R and two adapted models to French averaged over French tasks.

and French. We refer to this model as **RoBERTa (EN → FR)**. For monolingual specialization (i.e. French) of XLM-R, we use FOCUS (Dobler and de Melo, 2023) to replace the tokenizer of XLM-R with the tokenizer of French RoBERTa. FOCUS first finds the shared tokens between the French RoBERTa and XLM-R vocabularies which XLM-R can use directly. New tokens (French tokens not in XLM-R) are represented as combinations of overlapping tokens in the French RoBERTa and XLM-R vocabularies. We refer to this model as **XLM-R (Multi → FR)**.[8]

The first three models in Table 5, namely French RoBERTa (FR), RoBERTa (EN → FR) and XLM-R (Multi → FR), use the same tokenizer, i.e. the same tokenization aggressiveness, while the embedding initialization is either identical or semantically similar (Minixhofer et al., 2022; Dobler and de Melo, 2023). Notably, all FAs obtain more similar Suff and Comp between RoBERTa (FR) and the two hybrid RoBERTa, rather than between XLM-R and the two hybrid RoBERTa. For example, each FA on XLM-R (Multi → FR) almost mirrors the sufficiency of RoBERTa (FR) with the greatest difference of 0.232 and the smallest of 0, even though XLM-R (Multi → FR) shares the same model parameters (non-embedding weights) with XLM-R (their greatest and smallest differences are 0.366 and 0.001, respectively). Therefore, we summarize that *FAs tend to be of similar faithfulness on models with the same tokenizer or the similar tokenizers regarding the splitting aggressiveness level.* Further, this leads to promising future research questions. First, do FAs really reflect the inner reasoning process of models? Second, when token units and their embeddings are identical or similar, different models tend to converge to a point after fine-tuning where they process these inputs in a similar way?

[8]See Appendix J.1 for the implementation details of these two models.

5.4 Qualitative analysis

For a qualitative evaluation, we examine the rationales extracted by the same 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 5.2.

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. Further experiments support the assumption on the impact of tokenization: the discrepancies in faithfulness are primarily driven by differences in tokenization rather than underlying differences in how mono- and multilingual models semantically process the input.

Limitations

A 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 and Llama, 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 presence in the pre-training corpora.

Finally, 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.

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.

JosÃl' CaÃśete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge PÃl'rez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

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. 2022. 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. 2022. 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.

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.

Gianna M Del Corso, Antonio Gulli, and Francesco Romani. 2005. Ranking a stream of news. In *Proceedings of the 14th international conference on World Wide Web*, pages 97–106.

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.

Konstantin Dobler and Gerard de Melo. 2023. Focus: Effective embedding initialization for monolingual specialization of multilingual models.

Asier GutiÃľrrez FandiÃśo, Jordi Armengol EstapÃľ, Marc PÃămies, 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*.

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*.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

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.

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.

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

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.

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 and Nikolaos Aletras. 2023. Incorporating attribution importance for improving faithfulness metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4732–4745, Toronto, Canada. Association for Computational Linguistics.

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

Table 6: 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 6, 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	bert-base-multilingual-uncased	Devlin et al. (2019)
	XLM-R	xlm-roberta-base	Conneau et al. (2020)
	XLM-R large	xlm-roberta-large	Conneau et al. (2020)
English	BERT	bert-base-uncased	Devlin et al. (2019)
	RoBERTa	roberta-base	Liu et al. (2019)
Chinese	BERT	bert-base-chinese	Devlin et al. (2019)
	RoBERTa	hfl/chinese-roberta-wwm-ext	Cui et al. (2021)
Spanish	BERT	dccuchile/bert-base-spanish-wwm-uncased	Cañete et al. (2020)
	RoBERTa	PlanTL-GOB-ES/roberta-base-bne	Fandiño et al. (2022)
French	BERT	dbmdz/bert-base-french-europeana-cased	Schweter (2020)
	RoBERTa	ClassCat/roberta-base-french	n/a
Hindi	BERT	l3cube-pune/hindi-bert-scratch	Joshi (2022)
	RoBERTa	flax-community/roberta-hindi	n/a

Table 7: 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 8 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

Table 9 on page 14 shows the Suff and Comp of each feature attribution method on each dataset. Table 10 on page 15 shows the Soft-Suff and Soft-Comp of each feature attribution method on each dataset. 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.

E Faithfulness disparity on FAs and languages

F RoBERT v.s. BERT

F.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 mBERT and 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.

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

Table 15 on page 17 presents the original Suff and Comp results of each feature attribution method on each task for XLM-R large. It was used in Section

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	Socher et al. (2013)
		Agnews	Topic classification	102,000 / 18,000 / 7,600	36	F1	Del Corso et al. (2005)
		MultiRC	Reading Comprehension	24,029 / 3,214 / 4,848	290/17	F1	DeYoung et al. (2020); Jain et al. (2020)
Chinese	Sino-Tibetan	Ant	Reading Comprehension	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)
French	Indo-European	CSL	Sentiment analysis	2,000 / 1,200 / 1,200	28	Accuracy	Le et al. (2020); Keung et al. (2020)
		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 8: 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

Table 9: Full results of faithfulness in Suff and Comp, and prediction performance. All faithfulness results are presented by being divided by the random baseline.

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	BERT	1.1174	1.1115	1.0811	1.3015	1.1817	1.2007	1.2005	1.2002	1.1995	1.1998
SST	mBERT	1.2650	1.3227	1.0719	1.1442	1.1330	1.2095	1.2131	1.2053	1.1965	1.2158
SST	RoBERTa	1.1197	1.0435	1.1384	1.2683	1.1531	1.1977	1.1985	1.2008	1.1992	1.1970
SST	XLM-R	1.2423	1.1081	1.1750	1.1196	1.2824	1.2054	1.2038	1.2006	1.2022	1.2069
Agnews	BERT	1.2185	1.2676	1.0728	1.0625	1.0420	1.2064	1.2095	1.2079	1.2023	1.2045
Agnews	mBERT	1.1117	1.3246	1.2289	1.0848	1.0826	1.2014	1.2007	1.2003	1.2000	1.1996
Agnews	RoBERTa	1.0688	1.0713	1.3344	1.3121	1.2398	1.2016	1.1992	1.2008	1.1983	1.2024
Agnews	XLM-R	1.2147	1.3088	1.3372	1.0413	1.1902	1.1984	1.1972	1.1978	1.1996	1.2001
MultiRC	BERT	1.2937	1.2236	1.2478	1.0459	1.3163	1.1974	1.1963	1.1986	1.1952	1.2012
MultiRC	mBERT	1.1990	1.0981	1.1447	1.2792	1.2093	1.2023	1.2028	1.2003	1.2017	1.2033
MultiRC	RoBERTa	1.2802	1.1747	1.3438	1.2485	1.3111	1.2059	1.1938	1.2087	1.2030	1.1969
MultiRC	XLM-R	1.2223	1.2089	1.3215	1.1894	1.3373	1.1825	1.1719	1.1878	1.1771	1.1932
ChnSentiCorp	BERT(zh)	1.1092	1.0514	1.1237	0.9149	0.8834	1.0038	0.9987	1.0025	0.9975	1.0013
ChnSentiCorp	mBERT	1.0982	1.1434	1.1168	0.9094	0.9811	0.9983	0.9997	0.9990	0.9977	1.0003
ChnSentiCorp	RoBERTa(zh)	1.1400	0.8794	0.9780	1.1358	0.9518	1.0042	1.0052	1.0011	1.0032	1.0022
ChnSentiCorp	XLM-R	0.8697	1.1414	1.0343	1.1033	1.0253	0.9964	0.9999	0.9981	0.9973	1.0007
KR	BERT(zh)	0.8564	0.9301	1.0576	0.9246	1.0910	1.0001	1.0002	0.9999	0.9998	1.0003
KR	mBERT	0.9053	0.9865	1.1194	0.9745	0.9071	0.9990	1.0001	0.9997	1.0006	0.9993
KR	RoBERTa(zh)	0.9594	0.9000	1.0735	1.1351	0.9105	0.9991	1.0013	0.9996	1.0009	1.0004
KR	XLM-R	0.9091	1.1106	0.9407	1.0584	1.1169	0.9911	0.9938	1.0038	1.0016	0.9964
ANT	BERT(zh)	0.8968	1.0331	1.1381	1.0937	1.0724	1.0001	0.9014	1.0001	1.0213	1.0002
ANT	mBERT	0.9319	0.9484	0.8653	0.9424	1.0193	1.0048	1.0034	1.0071	1.0022	1.0059
ANT	RoBERTa(zh)	0.8758	1.0080	1.0925	0.9138	0.8952	1.0154	1.0031	1.0062	1.0125	1.0093
ANT	XLM-R	1.1315	0.8814	1.1122	1.1218	1.0655	0.9949	1.0105	0.9905	1.0069	1.0030
Hindi XNLI	BERT(hi)	0.9956	0.9982	0.9964	0.9973	0.9991	1.0038	1.0017	1.0030	1.0024	1.0009
Hindi XNLI	mBERT	0.9927	1.0037	0.9899	0.9955	0.9983	1.0039	1.0000	1.0049	1.0030	1.0020
Hindi XNLI	RoBERTa(hi)	0.9973	0.9991	1.0009	0.9982	0.9964	1.0051	1.0017	0.9984	1.0034	1.0069
Hindi XNLI	XLM-R	1.0002	0.9985	0.9996	1.0007	0.9979	0.9973	0.9999	0.9981	0.9963	1.0008
Hindi BBC Nli	BERT(hi)	0.8755	1.1450	1.1292	0.9658	0.9224	1.0033	1.0065	0.9986	1.0015	1.0050
Hindi BBC Nli	mBERT	0.9492	0.8837	1.1009	1.0972	1.0853	1.0001	0.9993	1.0018	1.0024	1.0030
Hindi BBC Nli	RoBERTa(hi)	0.8649	1.1123	0.9938	1.1217	0.9366	0.9789	0.9914	1.0917	1.0068	1.0853
Hindi BBC Nli	XLM-R	1.0760	0.9141	1.1469	1.0084	1.1297	0.9916	0.9960	1.0112	0.9975	0.9911
Hindi BBC Topic	BERT(hi)	0.9914	0.8890	0.9382	0.8912	1.0277	0.9976	0.9992	0.9969	1.0008	0.9984
Hindi BBC Topic	mBERT	1.0715	0.8891	1.0780	0.9357	1.1289	0.9984	0.9977	0.9991	1.0004	0.9998
Hindi BBC Topic	RoBERTa(hi)	1.0792	0.9567	1.0483	1.0080	1.0797	0.9978	1.0011	0.9989	1.0032	1.0022
Hindi BBC Topic	XLM-R	0.9972	1.0010	1.1432	1.1174	1.0763	0.9999	0.9981	0.9961	1.0007	0.9972
French CSL	BERT(fr)	0.9700	1.1767	1.1052	1.2127	1.1523	1.1056	1.0937	1.1078	1.0968	1.1029
French CSL	mBERT	1.1124	1.0003	1.0063	1.1548	1.1742	1.1014	1.1023	1.1019	1.1006	1.1028
French CSL	RoBERTa(fr)	1.0482	0.9482	1.1602	1.1290	1.0932	1.1006	1.0994	1.0987	1.0975	1.0981
French CSL	XLM-R	0.9456	1.1939	0.9480	1.0073	1.1293	1.1014	1.0941	1.1039	1.1064	1.0966
French Paws	BERT(fr)	1.0163	1.0772	1.0279	1.0206	1.2284	1.0972	1.1007	1.0986	1.0979	1.0993
French Paws	mBERT	1.0518	1.1542	1.1038	1.2155	1.0163	1.1018	1.1006	1.1022	1.1026	1.1014
French Paws	RoBERTa(fr)	1.1262	1.0503	1.0594	1.0303	1.1667	1.0894	1.0819	1.0678	1.1084	1.0764
French Paws	XLM-R	1.0932	1.1402	0.9883	1.1941	1.1115	1.1021	1.0968	1.0991	1.1000	1.0975
French XNLI	BERT(fr)	1.0956	1.0985	1.1030	1.0971	1.1015	1.1026	1.1009	1.0982	1.1017	1.0991
French XNLI	mBERT	1.0972	1.0822	1.0934	1.0897	1.0859	1.1019	1.1056	1.1028	1.1038	1.1047
French XNLI	RoBERTa(fr)	1.0994	1.1012	1.0981	1.1006	1.0987	1.1012	1.0975	1.1035	1.0988	1.1024
French XNLI	XLM-R	1.1003	1.0986	1.1007	1.0994	1.0999	1.0959	1.1000	1.0949	1.0980	1.0970
Spanish CSL	BERT(es)	1.2813	0.9915	1.2320	1.1154	1.1478	1.1456	1.1568	1.1545	1.1522	1.1478
Spanish CSL	mBERT	1.0206	1.1762	1.1969	1.0003	0.9930	1.1537	1.1550	1.1524	1.1564	1.1576
Spanish CSL	RoBERTa(es)	0.9958	1.0382	1.0512	1.2614	1.2268	1.1541	1.1562	1.1602	1.1520	1.1583
Spanish CSL	XLM-R	1.0035	1.2053	1.2477	1.0237	1.2107	1.1447	1.1465	1.1483	1.1474	1.1456
Spanish Paws	BERT(es)	1.2447	1.2738	1.0227	0.9928	1.1452	1.1539	1.1487	1.1526	1.1513	1.1474
Spanish Paws	mBERT	1.2031	1.1840	1.2014	1.0241	1.2006	1.1504	1.1516	1.1522	1.1498	1.1492
Spanish Paws	RoBERTa(es)	1.1039	1.0871	1.1247	1.2679	1.2201	1.2160	1.1559	1.0993	1.3304	1.3864
Spanish Paws	XLM-R	1.0206	1.1617	1.1097	1.0455	1.1732	1.1466	1.1431	1.1478	1.1458	1.1448
Spanish XNLI	BERT(es)	1.1556	1.1593	1.1575	1.1537	1.1519	1.1479	1.1465	1.1472	1.1486	1.1493
Spanish XNLI	mBERT	1.1435	1.1459	1.1485	1.1534	1.1558	1.1524	1.1519	1.1515	1.1505	1.1501
Spanish XNLI	RoBERTa(es)	1.1491	1.1497	1.1484	1.1487	1.1494	1.1534	1.1511	1.1556	1.1545	1.1523
Spanish XNLI	XLM-R	1.1475	1.1478	1.1484	1.1481	1.1487	1.1526	1.1519	1.1505	1.1512	1.1497

Table 10: Full results of faithfulness in Soft-Suff and Soft-Comp. All faithfulness results are presented by being divided by the random baseline.

5 to investigate the impact of the model size gap on the faithfulness disparity.

G Impact of model size

The results indicate a lower faithfulness on the larger BERT model across FAs and tasks. Specifically, the Suff and Comp of the monolingual English BERT-large are higher than its counterpart

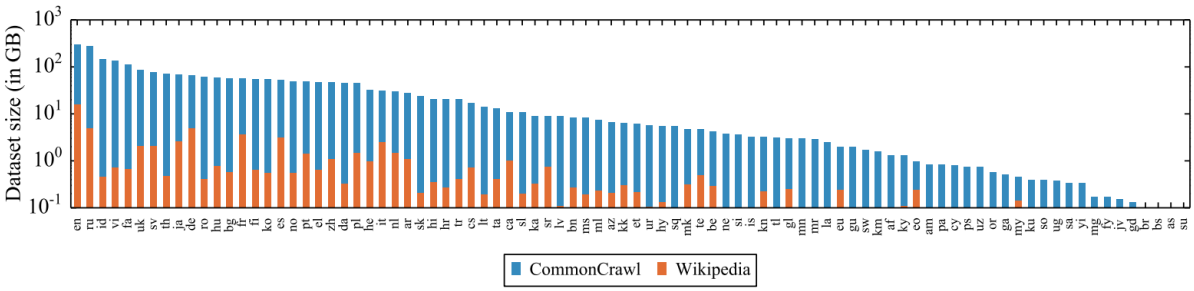

Figure 3: Amount of data in GiB (log-scale) for the 88 languages that appear in both the Wiki-100 corpus (used for mBERT) and the CC-100 (XLM-R). CC-100 increases the amount of data by several orders of magnitude, in particular for low-resource languages (Conneau et al., 2020).

Sufficiency							
	α	$\alpha\nabla\alpha$	$x\nabla x$	IG	DL	Avg Diff	P value
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							
	α	$\alpha\nabla\alpha$	$x\nabla x$	IG	DL	Avg Diff	P value
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 11: Suff and Comp difference between multilingual BERT (mBERT) and counterpart monolingual BERT. (plum indicates monolingual models are more faithful than multilingual models.

Sufficiency							
	α	$\alpha\nabla\alpha$	$x\nabla x$	IG	DL	Avg Diff	P value
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							
	α	$\alpha\nabla\alpha$	$x\nabla x$	IG	DL	Avg Diff	P value
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 12: Suff and Comp difference between XLM-R (multilingual RoBERTa) and counterpart monolingual RoBERTa (plum indicates monolingual models are more faithful than multilingual models.)

BERT-base (13 out of 16 comparison pairs as shown in Table 1), except for cases of Suff and

Soft-Sufficiency							
	α	$\alpha\nabla\alpha$	$x\nabla x$	IG	DL	Avg Diff	P value
English	-0.018	0.048	0.015	0.033	-0.038	0.008	0.821
Chinese	0.024	0.021	-0.073	-0.036	-0.046	-0.022	0.45
Spanish	-0.105	0.027	0.045	-0.028	-0.032	-0.019	0.539
French	0.06	-0.039	-0.011	0.043	-0.069	-0.003	0.915
Hindi	0.05	-0.085	0.035	0.058	0.088	0.029	0.284
Avg Diff	0.002	-0.006	0.002	0.014	-0.019	-0.001	-
P value	0.931	0.869	0.942	0.635	0.499	-	0.92
Soft-Comprehensiveness							
	α	$\alpha\nabla\alpha$	$x\nabla x$	IG	DL	Avg Diff	P value
English	0.003	0.003	0	0	0.004	0.002	0.273
Chinese	-0.001	0.034	0.001	-0.006	0.001	0.006	0.408
Spanish	0.003	0.002	0.001	0.002	0.004	0.002	0.036
French	0	0.004	0.001	0.004	0.003	0.002	0.041
Hindi	-0.001	-0.003	0.002	0	0	0	0.711
Avg Diff	0.001	0.008	0.001	0	0.002	0.002	-
P value	0.482	0.254	0.406	0.979	0.081	-	0.094

Table 13: Soft-Suff and Soft-Comp difference between mBERT and counterpart monolingual BERT (plum indicates monolingual models are more faithful than multilingual models.)

Soft-Sufficiency							
	α	$\alpha\nabla\alpha$	$x\nabla x$	IG	DL	Avg Diff	P value
English	0.07	0.112	0.006	-0.16	0.035	0.013	0.699
Chinese	-0.022	0.115	-0.019	0.033	0.15	0.052	0.245
Spanish	-0.026	0.08	0.061	-0.154	-0.021	-0.012	0.692
French	-0.045	0.111	-0.094	0.014	-0.006	-0.004	0.893
Hindi	0.044	-0.051	0.082	0	0.064	0.028	0.356
Avg Diff	0.004	0.073	0.007	-0.053	0.044	0.015	-
P value	0.901	0.053	0.787	0.165	0.075	-	0.3
Soft-Comprehensiveness							
	α	$\alpha\nabla\alpha$	$x\nabla x$	IG	DL	Avg Diff	P value
English	-0.006	-0.006	-0.008	-0.007	0.001	-0.005	0.105
Chinese	-0.012	-0.002	-0.005	-0.004	-0.004	-0.005	0.011
Spanish	-0.027	-0.007	0.011	-0.064	-0.086	-0.035	0.102
French	0.003	0.004	0.009	0	0.005	0.004	0.19
Hindi	0.002	0	-0.028	-0.006	-0.035	-0.013	0.112
Avg Diff	-0.008	-0.002	-0.004	-0.016	-0.024	-0.011	-
P value	0.143	0.346	0.564	0.203	0.181	-	0.018

Table 14: Soft-Suff and Soft-Comp difference between XLM-R and counterpart monolingual RoBERTa.)

Comp 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

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 15: 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	Avg Diff	P value
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							
	α	$\alpha\nabla\alpha$	$x\nabla x$	IG	DL	Avg Diff	P value
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 16: Suff and Comp difference between XLM-R Large and counterpart monolingual RoBERTa.

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 17: Suff and Comp 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).

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.

H 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'.

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

I Fertility and splitting ratio

I.1 Full results

Table 18 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 4.

I.2 Correlation tokenization aggressiveness and faithfulness disparity

J Disparity in tokenization aggressiveness

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

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

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

Table 19: Pearson correlation coefficient between fertility, splitting ratio, and faithfulness disparity.

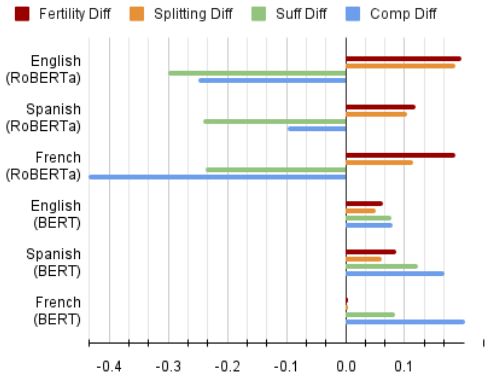

Figure 4: The impact of tokenization aggressiveness ("Fertility Diff" and "Splitting Diff") on faithfulness disparity ("Suff Diff" and "Comp Diff").

the corresponding score of the monolingual counterpart model. We observe that multilingual models consistently tokenize more aggressively 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.

J.1 Implementation of WECHSEL and FOCUS

WECHSEL first copies all non-embedding inner parameters from the English RoBERTa, and exchanges the tokenizer with the tokenizer of French RoBERTa. In this paper, we use the WECHSEL model in French published by the author (Minixhofer et al., 2022).[9].

FOCUS transfer part of the pre-trained embedding matrix from XLM-R to the French RoBERTa tokenizer, and for the new tokens to XLM-R, FOCUS represents them as a weighted mean of overlapping tokens' embeddings. The overlapping tokens between both tokenizers are the anchor points to find the similar tokens for calculating the weighted mean. For FOCUS, following (Dobler and de Melo, 2023), we use FastText embeddings in French to obtain the static token embeddings and find similar tokens. We follow the default hyperparameter settings in FOCUS.[10] The overlapping tokens are selected based on semantic similarity in an auxiliary static token embedding space.

[9] https://github.com/CPJKU/wechsel

[10] https://github.com/konstantinjdobler/focus