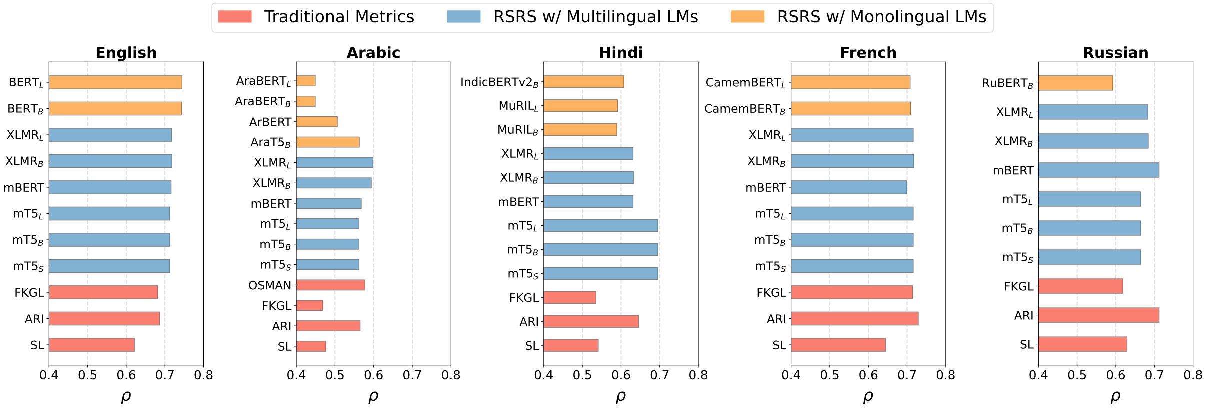

Figure 6: Pearson correlation (ρ) of **unsupervised** readability measurements on the test set of README++, including RSRS (Martinc et al., 2021) which leverages conditional word probabilities estimated by LMs. RSRS which uses multilingual LLMs performs better than RSRS which uses monolingual models in languages with non-Latin scripts.

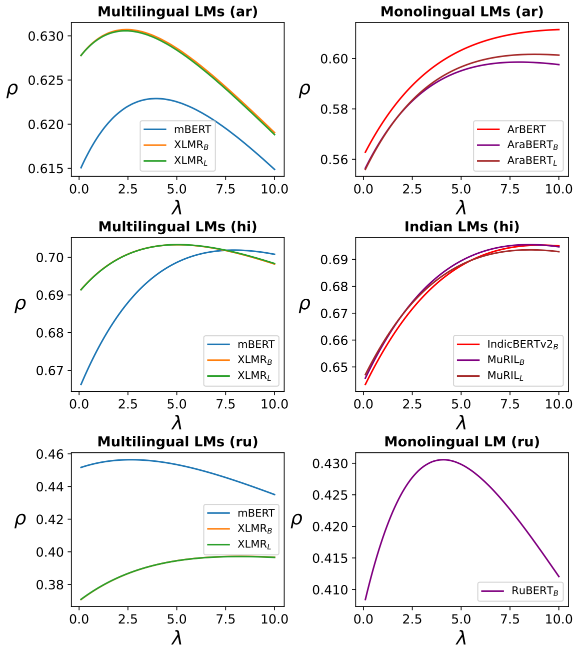

Figure 7: Effect of increasing the penalty factor (λ) on the Pearson correlation (ρ) between RSRS scores and human ratings for Arabic, Hindi and Russian sentences that contains transliterations. The plot shows a clear improvement in correlation as λ increases, which is more significant for monolingual than multilingual models.

diseases, historical figures, etc.) that the LM never saw during pre-training. We hypothesize that this design choice in RSRS degrades its performance on languages with non-Latin script since many of these transliterated words do not add to the difficulty level of the sentence for humans.

Unsupervised LM-based RSRS struggle with transliterations. To test the impact of transliterated words on RSRS scores, we asked Arabic, Hindi, and Russian annotators to indicate if a sentence contains transliterated words when annotating. This resulted in 320 sentences with transliterations in Arabic (16.45% of Arabic data), 561 sentences in Hindi (36.81% of Hindi data), and 120 sentences in Russian (6.82% of Russian data). We penalized the RSRS scores of those sentences by a factor $\frac{\lambda}{S}$, where λ is a penalty factor and S is the length of the sentence. We compute the correlation with human labels for an increasing penalty λ to analyze whether decreasing those scores results in a higher correlation since we assume transliterations cause RSRS scores to be unreasonably high. The results are shown in Figure 7 for 0.1 increments of λ. The trends corroborate with our hypothesis, where correlation increases as the penalty becomes higher up to a certain level. The improvement reaches up to 6-7% for monolingual LMs. Multilingual LMs (improvements of 1-3%) were less affected, indicating their greater robustness to transliterations. This underscores the need for careful consideration of transliterations in future research.

5 Cross-Domain Cross-Lingual Analyses

We test the ability of LMs trained on README++ to generalize to unseen domains (5.1) and transfer to other languages (5.2) compared with models trained on previous datasets.

5.1 Performance on Unseen Domains

To test how well fine-tuned models perform on unseen domains, we create new train/val/test splits from README++ by removing an increasing num-

	#Unseen Domains (#Data Sources)	#train/val	#test	ReadMe++		CEFR-SP	
				F1	ρ	F1	ρ
English	2 (7): Wik, Res	1995 / 235	631	**37.57**	**0.611**	20.95	0.439
	4 (7): Let, Ent, Soc, Gui	2285 / 267	309	**40.16**	**0.761**	24.91	0.649
	6 (14): Res, Fin, Sta, Ent, Dia, New	1885 / 221	755	**34.61**	**0.780**	20.69	0.517
	8 (25): Pol, Cap, Sta, Res, Rev, Leg, Soc, Poe	1653 / 191	1017	**43.88**	**0.828**	23.80	0.690

	#Unseen Domains (#Data Sources)	#train/val	#test	ReadMe++		ALC Corpus	
				F1	ρ	F1	ρ
Arabic	2 (2): Tex, New	1540 / 180	225	**47.54**	**0.626**	6.80	-0.208
	4 (7): Poe, Gui, Ent, Dia	1457 / 173	315	**39.24**	**0.683**	7.27	-0.043
	6 (11): For, New, Spe, Cap, Wik, Res	910 / 106	929	**34.47**	**0.609**	10.25	0.083
	8 (13): Ent, For, Leg, Spe, Wik, Dia, Poe, Res	918 / 109	918	**29.56**	**0.523**	6.79	0.144

Table 4: Supervised mBERT-based readability model fine-tuned on our ReadMe++ corpus achieve much better performance on unseen domains than the same model trained on existing datasets, namely CEFR-SP (Arase et al., 2022) for English and the ALC Corpus (Khallaf and Sharoff, 2021) for Arabic.

src → tgt	ReadMe++		CEFR-SP		CompDS	
	F1	ρ	F1	ρ	F1	ρ
en → ar	**31.48**	**0.606**	8.81	0.071	5.99	0.322
en → hi	**23.87**	**0.702**	13.15	0.267	10.38	0.381
en → fr	**30.29**	**0.768**	11.06	-0.026	5.92	0.335
en → ru	**24.60**	**0.760**	15.69	0.173	10.33	0.412
en → it	**14.68**	**0.239**	9.88	-0.043	10.06	0.099
en → de	**22.19**	**0.701**	10.00	-0.092	11.84	0.408

Table 5: Zero-shot cross-lingual transfer results using XLMR$_L$. LMs fine-tuned on English data (en) of ReadMe++ significantly outperform LMs fine-tuned with CEFR-SP (Arase et al., 2022) or CompDS (Brunato et al., 2018) in transfer to Arabic (ar), Hindi (hi), French (fr), Russian (ru), Italian (it), and German (de).

ber of randomly sampled domains from the dataset (Table 4). We use the sentences from the removed domains as the test set and use the rest of the dataset for training and validation. For direct comparison, we randomly sample the same amount of train/val sentences in each experiment from the open-sourced Wikipedia-based portion of CEFR-SP (Arase et al., 2022) to fine-tune mBERT models. We evaluate on the unseen domains test set from ReadMe++. The results in Table 4 show that **models fine-tuned using ReadMe++ achieve good performance on unseen domains and outperform models trained using CEFR-SP**, demonstrating the advantage of domain diversity in ReadMe++.

We perform the same experiments in Arabic by comparing to the ALC Corpus (Khallaf and Sharoff, 2021), which is labeled on 5-scale CEFR levels (A1, A2, B1, B2, C). We convert the labels in ReadMe++ to the same scale of ALC Corpus by combining C1 and C2 into C and then perform a 5-way classification. We observe the same trend, where models trained using the Arabic portion of ReadMe++ achieve good performance on unseen domains and outperform models trained on ALC.

5.2 Performance on Cross-lingual Transfer

We perform zero-shot cross-lingual transfer from English to 6 different languages by fine-tuning multilingual models using the English subset of ReadMe++. For comparison, we also fine-tune on the same number of train/valid sentences that we randomly sample from the open-sourced Wikipedia-based portion of CEFR-SP (Arase et al., 2022) and the full English CompDS (Brunato et al., 2018) corpora. We evaluate on the Arabic, Hindi, French, and Russian test sets from ReadMe++, as well as Italian CompDS (Brunato et al., 2018) and German TextComplexityDE (Naderi et al., 2019). Since CompDS and TextComplexityDE rate on scales from 1-7 instead of 1-6 but have only a few level-7 sentences, we merged their level 6 and 7 together. The results are shown in Table 5 for XLMR$_L$, where we find that **the model fine-tuned using ReadMe++ performs much better cross-lingual transfer across all tested languages** compared to models fine-tuned using CEFR-SP or CompDS, reaching high correlation values of 0.7 in most languages. In several cases, training on ReadMe++ leads to a 50% increase in performance. This trend is also observed across several models which we report in Appendix E.3.

6 Conclusion

We introduced ReadMe++, a multi-domain dataset for multilingual sentence readability assessment. ReadMe++ provides 9757 sentences in Arabic, English, French, Hindi, and Russian that are collected from 112 different data sources and annotated by humans based on the CEFR scale.

We showed that LMs trained using ReadMe++ achieve strong performance across different textual domains and perform well in cross-lingual transfer from English to 6 target languages, outperforming models trained on previous datasets. By releasing ReadMe++, we hope to encourage and enable the development and evaluation of more effective and robust methods for multilingual sentence readability assessment.

Limitations

ReadMe++ offers a diversity of text domains in multiple languages. Most domains in our dataset include texts in all the languages we considered, with a few exceptions where openly accessible data was not available in every language. The medical text domain, which consists of clinical reports, is only available in English. However, medical-related texts in other languages are covered within other domains, such as Research and Wikipedia.

In our experiments on cross-lingual transfer, we showed that models fine-tuned on ReadMe++ transfer well to other languages and outperform models trained on previous datasets. However, our dataset does not cover low-resource languages, which limits the ability to perform evaluation in such scenarios. Future work can extend ReadMe++ to include such languages. We will be releasing our rank-and-rate annotation interface that will enable easy extensions of our resource to additional languages by the research community.

We analyzed how transliterations can negatively impact the performance of the LM-based RSRS unsupervised metric due to its approach to handling rare words. However, certain rare words such as jargon and complex terminology could well add to the difficulty of a sentence. The language and domain diversity of our resource will encourage future studies to make a more in-depth exploration of this particular point and enable the development and evaluation of better unsupervised metrics.

Ethical Considerations

We are committed to upholding ethical standards in constructing and disseminating the ReadMe++ corpus. To ensure the integrity of our data collection process, we have made our best effort to obtain data from sources that are available in the public domain, released under Creative Commons or similar licenses, or can be used freely for personal and non-commercial purposes according to the resource's Terms and Conditions of Use. These sources include public domain books, publicly available documents/reports, and publicly available datasets. We use a small number of randomly sampled sentences for academic research purposes, specifically for labeling sentence readability. We have included a full list of licenses and terms of use for each source in Appendix G. We would like to note that two of the sources we used require access permission from the original authors, specifically the i2b2/VA (Uzuner et al., 2011) and Hindi Product Reviews (Akhtar et al., 2016) datasets. Therefore, sentences and annotations from these sources will not be shared with the community unless access permission has been obtained from the original authors.

Every annotator was informed that their annotations were being used to create a dataset for readability assessment. When collecting sentences from social media and forums, we have excluded any sampled sentences containing offensive/hateful speech, stereotypes, or private user information.

Acknowledgments

The authors would like to thank Nour Allah El Senary, Govind Ramesh, Suraj Mehrotra, and Ryan Punamiya for their help in data annotation. This research is supported in part by the NSF awards IIS-2144493 and IIS-2112633, NIH award R01LM014600, ODNI and IARPA via the HIATUS program (contract 2022-22072200004). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of NSF, NIH, ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, et al. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. 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 7088–7105.

Sweta Agrawal and Marine Carpuat. 2019. Controlling text complexity in neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th*