READMe++: Benchmarking Multilingual Language Models for Multi-Domain Readability Assessment

Tarek Naous, Michael J. Ryan, Anton Lavrouk, Mohit Chandra, Wei Xu
College of Computing
Georgia Institute of Technology
{tareknaous, michaeljryan, antonlavrouk, mchandra9}@gatech.edu; wei.xu@cc.gatech.edu

Abstract

We present a comprehensive evaluation of large language models for multilingual readability assessment. Existing evaluation resources lack domain and language diversity, limiting the ability for cross-domain and cross-lingual analyses. This paper introduces READMe++, a multilingual multi-domain dataset with human annotations of 9757 sentences in Arabic, English, French, Hindi, and Russian, collected from 112 different data sources. This benchmark will encourage research on developing robust multilingual readability assessment methods. Using READMe++, we benchmark multilingual and monolingual language models in the supervised, unsupervised, and few-shot prompting settings. The domain and language diversity in READMe++ enable us to test more effective few-shot prompting, and identify shortcomings in state-of-the-art unsupervised methods. Our experiments also reveal exciting results of superior domain generalization and enhanced cross-lingual transfer capabilities by models trained on READMe++. We will make our data publicly available and release a python package tool for multilingual sentence readability prediction using our trained models at: https://github.com/tareknaous/readme

1 Introduction

Readability assessment is the task of determining how difficult it is for a specific audience to read and comprehend a piece of text (Vajjala, 2022). Developing methods for automatically predicting the readability of a sentence is beneficial for many applications such as controllable text simplification (Chi et al., 2023; Agrawal and Carpuat, 2019), ranking search engine results by their level of difficulty (Fourney et al., 2018), and selecting appropriate reading material for language learners (Xia et al., 2019). Making such technologies robust to textual variations and accessible to a global community with diverse languages requires readability prediction methods that generalize across different text domains and language families.

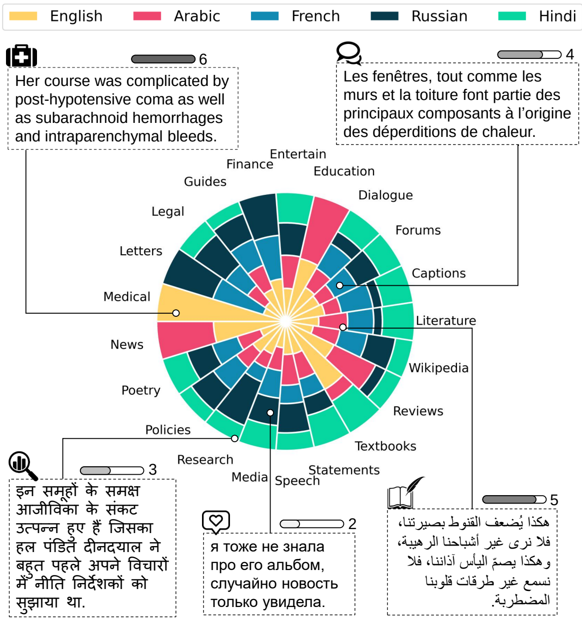

Figure 1: Language distribution per each domain in READMe++. Example sentences from each language are shown along with their human-annotated readability levels on a 6-point scale (1: easiest, 6: hardest).

Recent advancements in Language Models (LMs) (Xue et al., 2021; Conneau et al., 2020) have enabled the development of neural-based readability assessment methods (Martinc et al., 2021). Despite the progress made, the absence of a diverse benchmark limits the ability to effectively evaluate how well LM-based methods, whether supervised, unsupervised, or prompting-based, perform across domains and languages. Current evaluation resources for sentence readability assessment suffer from a few crucial shortcomings. First, existing datasets are primarily composed of sentences collected from Wikipedia (Naderi et al., 2019; Arase et al., 2022; Štajner et al., 2017) or news articles (Brunato et al., 2018). However, LMs have been shown to struggle when handling data from a differ-

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 12230–12266
November 12-16, 2024 ©2024 Association for Computational Linguistics

Dataset	Languages	Scripts	#Data Sources
MTDE (De Clercq and Hoste, 2016)	en, nl	Latin	4 (Wikipedia, BNC, Dutch Parallel Corpus, SoNaR)
S1131 (Štajner et al., 2017)	en	Latin	2 (Wikipedia, Newsela)
CompDS (Brunato et al., 2018)	en, it	Latin	2 (Italian UD Treebank, WSJ from Penn Treebank)
TextComplexityDE (Naderi et al., 2019)	de	Latin	1 (Wikipedia, Leichte Sprache)
CEFR-SP (Arase et al., 2022)	en	Latin	3 (Wikipedia, Newsela, SCoRE)
README++ **(Ours)**	ar, en, fr, hi, ru	Arabic, Brahmic, Cyrillic, Latin	**112** (examples in Table 2; full list in Appendix A)

Table 1: Summary of readability datasets with *sentence-level annotations*. Our README++ corpus provides more domain and typological diversity. There also exist more datasets with document-level readability ratings (§2).

ent domain outside of their training corpus (Plank, 2016; Farahani et al., 2021; Arora et al., 2021). For reliable readability assessment, it's critical for methods to perform well across various textual domains. Hence, a domain-diverse benchmark is essential in assessing model domain generalization. Past work also often utilized document-based readability data as an approximation for sentence-based readability (more in §2), due to a lack of human readability ratings on individual sentences (Martinc et al., 2021; Lee and Vajjala, 2022). Additionally, there is no existing benchmark for sentence readability assessment that covers a diverse set of language families, limiting the ability to perform cross-lingual evaluation and analysis.

To address these gaps in the field, we introduce README++, a diverse multi-domain dataset for multilingual sentence readability assessment. README++ consists of 9757 human-annotated sentences drawn from 112 distinct data sources and covers 5 different languages: Arabic, English, French, Hindi, and Russian (see examples in Figure 1). We focus on readability assessment for second language learners (Xia et al., 2019) and thus annotate sentences for their readability level based on the Common European Framework of Reference for Languages (CEFR) scale (§ 3.2).

Using README++, we benchmark a variety of monolingual and multilingual LMs for multi-domain readability assessment in the supervised, unsupervised, and few-shot prompting settings. The domain and language diversity in README++ enable us to analyze more effective few-shot prompting (§ 4.1) and identify shortcomings in existing unsupervised readability prediction methods, such as the effect of transliterations on their performance in languages with non-Latin script (§ 4.2). Finally, we show that LMs fine-tuned using README++ perform better on unseen domains and exhibit superior cross-lingual transfer capabilities from English to six target languages: Arabic, French, Hindi, Russian, Italian, and German, compared with LMs trained on previous datasets (§ 5).

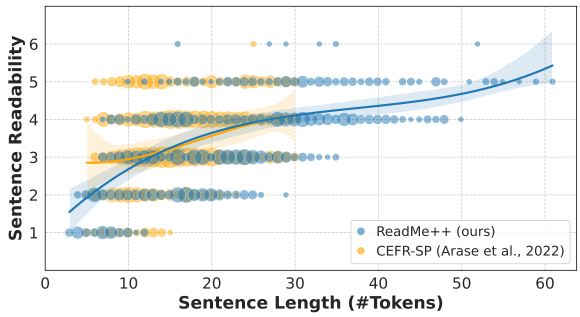

Figure 2: Distribution of sentence lengths across readability levels in the English portion of README++, compared with CEFR-SP (Arase et al., 2022). README++ offers a wider coverage of lengths and readability levels.

2 Related Work

Document-based Readability. Many datasets used in readability research have only document-level labels, as they were collected from sources (e.g., textbooks) that provide parallel or non-parallel text at varied levels of writing. These include WeeBit (Vajjala and Meurers, 2012), Newsela (Xu et al., 2015), Cambridge (Xia et al., 2016), OneStopEnglish (Vajjala and Lučić, 2018), VikiWiki (Azpiazu and Pera, 2019), Slovenian SB (Martinc et al., 2021), English-Chinese LR (Rao et al., 2021), ALC (Khallaf and Sharoff, 2021), Gloss (Khallaf and Sharoff, 2021), ZAEBUC (Habash and Palfreyman, 2022), SAMER (Alhafni et al., 2024), and Philippines Corpus (Imperial and Kochmar, 2023). While appropriate for assessing document readability, such datasets are suboptimal for sentence-level readability compared to resources with ground-truth readability labels for individual sentences (Cripwell et al., 2023).

Sentence-based Readability. Only a few existing datasets (De Clercq and Hoste, 2016; Štajner et al., 2017; Brunato et al., 2018; Naderi et al., 2019) were created by manually annotating indi-

Domain (Abrv)	#	Examples of Data Sources — Full list for all languages in Appendix A		
		Arabic (ar)	English (en)	Hindi (hi)
Captions (Cap)	9	**Images** (ElJundi et al., 2020)	**Videos** (Wang et al., 2019)	**Movies** (Lison and Tiedemann, 2016)
Dialogue (Dia)	7	**Open-domain** (Naous et al., 2020)	**Negotiation** (He et al., 2018)	**Task-oriented** (Malviya et al., 2021)
Dictionaries (Dic)	2	**Dictionaries** (almaany.com)	**Dictionaries** (dictionary.com)	—
Entertainment (Ent)	4	**Jokes** (almrsal.com)	**Jokes** (Weller and Seppi, 2019)	**Jokes** (123hindijokes.com)
Finance (Fin)	3	—	**Finance** (Malo et al., 2014)	—
Forums (For)	7	**QA Websites** (Nakov et al., 2016)	**StackOverflow** (Tabassum et al., 2020)	**Reddit** (reddit.com)
Guides (Gui)	6	**Online Tutorials** (ar.wikihow.com)	**Code Documentation** (mathworks.com)	**Cooking Recipes** (narendramodi.in)
Legal (Leg)	9	**UN Parliament** (Ziemski et al., 2016)	**Constitutions** (constitutioncenter.org)	**Judicial Rulings** (Kapoor et al., 2022)
Letters (Let)	3	—	**Letters** (oflosttime.com)	—
Literature (Lit)	3	**Novels** (hindawi.org/books/)	**History** (gutenberg.org)	**Biographies** (Public Domain Books)
Medical Text (Med)	1	—	**Clinical Reports** (Uzuner et al., 2011)	—
News Articles (New)	2	**Sports** (Alfonse and Gawich, 2022)	**Economy** (Misra, 2022)	—
Poetry (Poe)	5	**Poetry** (aldiwan.net)	**Poetry** (poetryfoundation.org)	**Poetry** (hindionlinejankari.com)
Policies (Pol)	7	**Olympic Rules** (specialolympics.org)	**Contracts** (honeybook.com)	**Code of Conduct** (lonza.com)
Research (Res)	15	**Politics** (jcopolicy.uobaghdad.edu.iq)	**Science & Engineering** (arxiv.org)	**Economics** (journal.ijarms.org)
Social Media (Soc)	3	**Twitter** (Zheng et al., 2022)	**Twitter** (Zheng et al., 2022)	**Twitter** (Zheng et al., 2022)
Speech (Spe)	4	**Public Speech** (state.gov/translations)	**Public Speech** (whitehouse.gov)	**Ted Talks** (ted.com/talks)
Statements (Sta)	6	**Quotes** (arabic-quotes.com)	**Rumours** (Zheng et al., 2022)	**Quotes** (wahh.in)
Textbooks (Tex)	3	**Business** (hindawi.org/books/)	**Agriculture** (open.umn.edu)	**Psychology** (ncert.nic.in)
User Reviews (Rev)	12	**Products** (ElSahar and El-Beltagy, 2015)	**Books** (goodreads.com)	**Movies** (hindi.webdunia.com)
Wikipedia (Wik)	1	**Wikipedia** (wikipedia.com)	**Wikipedia** (wikipedia.com)	**Wikipedia** (wikipedia.com)
Total	112			

Table 2: List of domains and example data sources in ReadMe++ (see full list for all 5 languages in Appendix A).

vidual sentences for their level of readability (see Table 1). However, these sentence-level annotated datasets are largely limited to high-resource English and European languages that use the Latin script. They are also collected from one or a few data sources and are thus insufficient for studying the robustness of readability assessment methods across text domains. Further, these past datasets are annotated with various rating scales that do no have a clear readability grounding. The recent CEFR-SP dataset (Arase et al., 2022) adopts the 6-level CEFR scale for annotation, which grounds sentence readability in the language capability of a second language learner. However, CEFR-SP only contains English sentences from Wikipedia, Newsela (Xu et al., 2015, leveled news articles), and SCoRE (Chujo et al., 2015, textbooks for learning English). In comparison, our work highlights the importance of both domain and language coverage, resulting in more data diversity (see Figure 2). ReadMe++ covers 112 different data sources and is annotated at the sentence level in 5 languages.

Multilingual Readability Assessment. Several works have leveraged neural approaches for multilingual readability assessment. Many adopt fine-tuning strategies of transformer LMs (Azpiazu and Pera, 2019; Le et al., 2018; Imperial et al., 2022; Chakraborty et al., 2021; Mesgar and Strube, 2018; Blaneck et al., 2022). However, training data is often unavailable except in a few high-resource languages. Other works explored cross-lingual transfer strategies (Imperial and Kochmar, 2023), demonstrating effective transfer from English to French/Spanish (Lee and Vajjala, 2022) and Chinese (Rao et al., 2021). The work of Martinc et al. (2021) proposed an unsupervised approach that leverages an LM's distribution to compute a likelihood-based sentence readability score. The majority of these past studies have used document-based readability datasets. Using our dataset, we benchmark various LMs in the supervised, unsupervised, and few-shot prompting settings in diverse language scripts (i.e., Arabic, Latin, Brahmic, and Cyrillic). We show that LMs trained using the English portion of ReadMe++ perform better cross-lingual transfer to 6 target languages compared to models trained on previous datasets.

3 Constructing ReadMe++ Corpus

We present the detailed procedure for constructing the ReadMe++ corpus. To maximize the diversity of domains, we identified 112 data sources that are either with open licenses or shareable for non-commercial purposes (see Table 2). A total of 9757 sentences (1945 Arabic, 1669 French, 2861 English, 1524 Hindi, 1758 Russian) were sampled from these sources and then manually annotated. ReadMe++ supports multilingual, cross-lingual, and cross-domain experiments (§4).

3.1 Data Collection

Selecting Diverse Data Sources. Our data collection process varies per source and can be categorized into four approaches: **(1)** obtaining content directly from a website (e.g., Wikipedia), **(2)** extracting text from sources in PDF format (e.g., contract templates, reports, etc.), **(3)** sampling text from existing datasets (e.g., dialogue, user reviews, etc.), or **(4)** manually collecting sentences (e.g., dictionary examples, etc.). Collection details per domain are provided in Appendix A. For each domain, we collected the available texts from one or more data sources and then sampled 50 paragraphs per domain. We increased the sampling rate to 100 for unstructured sources such as PDFs since they are likely to return text not useful for annotation (e.g., headers, titles, references, etc.) that needs to be filtered out. From each paragraph, we sample one sentence that we use for readability annotation. Lastly, we perform manual quality checking to filter out any low-quality sentences and sentences that contain toxic, hateful, or offensive language.

Considering the Influence of Contexts. In addition to the sampled sentences, we collect up to three preceding sentences as context if available. Many of the sampled sentences could be placed in the body of a paragraph. We provided annotators with optional access to context in case they needed to know the context in which a sentence appears. Such cases have not been adequately considered in previous work; for example, Arase et al. (2022) collected only the first sentence in a paragraph. We provide additional results in Appendix E.4 where context was provided to LMs during fine-tuning.

3.2 Readability Annotation

Using the CEFR Standards. Previous works on sentence-level readability have used various rating scales such as 0-100 (De Clercq and Hoste, 2016), 3-point (Štajner et al., 2017), or 7-point (Naderi et al., 2019; Brunato et al., 2018) scales. However, these scales are prone to annotator subjectivity due to the lack of a clear readability grounding. Instead, following Arase et al. (2022), we adopt the Common European Framework of Reference for Languages (CEFR), which defines the language ability of a person on a 6-point scale ($1_{(A1)}$, $2_{(A2)}$, $3_{(B1)}$, $4_{(B2)}$, $5_{(C1)}$, $6_{(C2)}$), where A is for basic, B for independent, and C for proficient. Each level of the scale is grounded by can-do descriptors of a language learner, which act as a guide for annotators (see CEFR level descriptors in Appendix B).

Rank-and-Rate Annotation. Rating each sentence independently on a scale of readability comes with the drawback of annotators eventually not differentiating between different sentences. This results in most samples being labeled within one or two levels, limiting their usefulness for statistical analyses (McCarty and Shrum, 2000). Instead of rating alone as in prior works, we utilize a Rank-and-Rate approach (Maddela et al., 2023) for readability annotation, which mitigates independent sentence rating issues by providing comparative texts. We randomly group sentences into batches of 5 and ask annotators to first rank sentences of a batch from most to least readable and then rate each sentence individually on the 6-point CEFR scale. By comparing and contrasting sentences within a batch, annotators can better differentiate between the readability of different sentences and produce less subjective ratings. In our initial pilot studies, we found that annotators express a better experience when using the rank-and-rate framework and achieve higher agreements compared with rating alone. Our interface is shown in Appendix F.

Dataset		α	ρ
README++	**Arabic**	0.67	0.78
	English	0.78	0.81
	French	0.76	0.78
	Hindi	0.67	0.71
	Russian	0.68	0.72
CEFR-SP	**WikiAuto**	0.66	0.73
(Arase et al., 2022)	**SCoRe**	0.44	0.66

Table 3: Annotator agreements measured by Krippendorff's alpha (α) and Pearson Correlation (ρ). The agreements reached in CEFR-SP (Arase et al., 2022) are provided for comparison.

Annotator Selection. We take several steps to ensure the quality of our annotations. First, four of our authors who can speak each language provided the first set of annotations. We then hired two additional annotators for each language, who were university students who can speak the language and had linguistic annotation experience, or annotators we hired through Prolific. Annotators were paid at rates of $16-18/hour. When recruiting annotators, we first conducted training sessions to familiarize them with the CEFR scale and the annotation framework. We then gave each candidate a batch of 250 sentences and only proceeded with candidates who achieved a sufficient enough correlation (> 0.7) with the first set of annotations.

Inter-annotator Agreement. We report the Krippendorff's alpha (α) and average Pearson Correlation (ρ) between the three annotators for each language in Table 3. High agreements are achieved by our annotators (Artstein and Poesio, 2008), on par with the past work of Arase et al. (2022). We perform majority voting on the three annotations to obtain a final rating that we use in our experiments.

4 Benchmarking Experiments

As shown in Figures 2 and 3, the README++ corpus offers a diverse coverage of domains, readability levels, and sentence lengths, making it an ideal testbed for evaluating readability assessment methods. We benchmark supervised, unsupervised, and few-shot approaches using recently developed LMs. We use the same random train/valid/test split (detailed statistics in Appendix D.2) based on a 60/10/30% ratio per domain for all experiments, except the domain generalization study in §5.

4.1 Supervised & Prompting Methods

Supervised. We fine-tune LMs to classify sentence readability. We compare multilingual models, **mBERT** (Devlin et al., 2019) and **XLM-R** (Conneau et al., 2020), to monolingual models that include **BERT** (Devlin et al., 2019) for English, **AraBERT** (Antoun et al., 2020) and **ArBERT** (Abdul-Mageed et al., 2021) for Arabic, **CamemBERT** for French (Martin et al., 2020), and **RuBERT** (Kuratov and Arkhipov, 2019) for Russian. For Hindi, we use **MuRIL** (Khanuja et al., 2021) and **IndicBERTv2** (Kakwani et al., 2020), both pretrained on 12 Indian languages. We also consider encoder-decoder LMs, **mT5** (Xue et al., 2021), **Aya101** (Üstün et al., 2024), and **AraT5** (Elmadany et al., 2022). We fine-tune for 20 epochs using the cross-entropy loss and the Adam optimizer and tune the learning rate in the set $\{1e^{-5}, 1e^{-6}, 1e^{-7}\}$. We select checkpoints based on the best performance on the validation set. We report the average of 5 runs with different random initialization seeds.

Prompting. We perform in-context learning using **GPT3.5**, **GPT4** (Apr 2024), **Llama2-7b** (Touvron et al., 2023), **Llama3.1-8b** (Dubey et al., 2024), and **Aya23-8b** (Aryabumi et al., 2024). We provide LMs with a definition of readability and the descriptors of the six CEFR levels. We show the model five randomly sampled in-context examples from the train set and their corresponding CEFR levels, then ask the model to assess the readability of a new sentence based on the CEFR scale. Prompt details can be found in Appendix D.3.

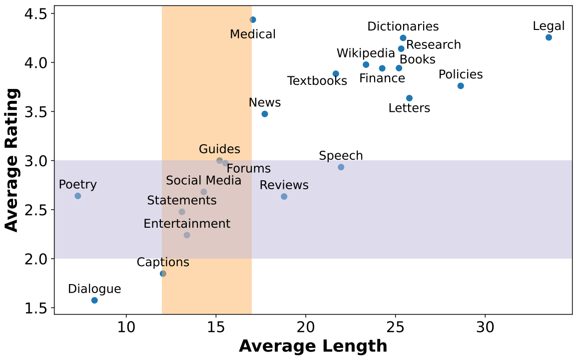

Figure 3: Average readability rating and sentence length per domain in the English portion of README++. Domain diversity presents additional challenges for readability assessment. Certain domains may be within the same readability range (e.g. [2, 3] that corresponds to A2 and B1 levels) but have varying lengths, while sentences within a length range (e.g. [12, 17] tokens) could be spread across the whole readability spectrum.

4.1.1 Results

The results are shown per language in Figure 4, where we report the Pearson Correlation (ρ) between the predictions and the ground-truth labels. Additional metrics are reported in Appendix E.1.

A gap exists between fine-tuning and few-shot performance. Fine-tuned models were able to achieve high correlation levels in the 0.7-0.9 range, with larger models showing improved performance. Overall, mT5$_L$ was among the best-performing fine-tuned models across all languages. However, the performance of prompted causal models with 5-shot examples was lower than that of fine-tuned models in all languages.

Domain diversity of in-context examples improves few-shot performance. We analyze the effect of the domain diversity of the few-shot examples on prompting performance. We prompt Llama2 by sampling examples from 1, 2, 4, and 8 domains. The domains from which the examples are sampled are also randomly sampled for each test sentence. The average correlation from 5 runs is shown in Figure 5, for an increasing number of shots. The performance gain from increasing domain diversity is clearly observed, with correlation improving all cases, reaching slightly above 0.7 in the best case. This improvement also outweighs the gains from increasing the number of shots, highlighting the importance of domain diversity.

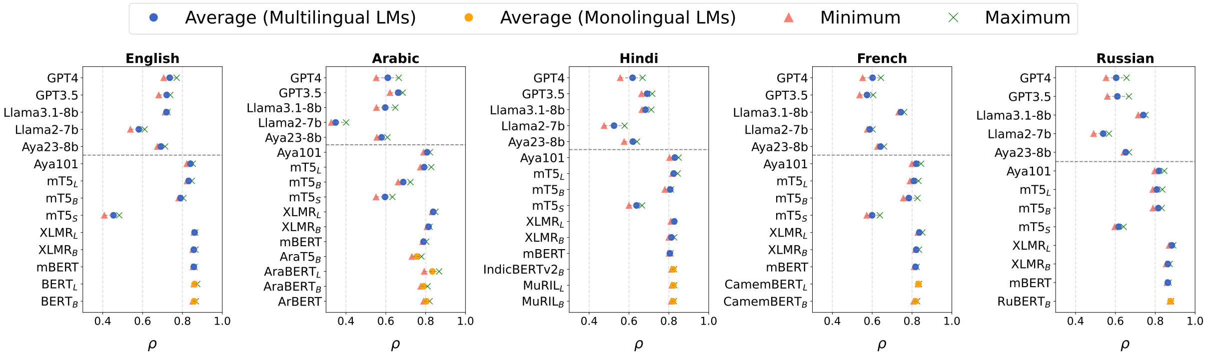

Figure 4: Pearson correlation (ρ) of **fine-tuned** multilingual and monolingual LMs, as well as **prompted** GPT3.5, GPT4, Aya23-8b, Llama2-7b, and Llama3.1-8b models with 5-shot examples, on the test set of README++. The small ($_S$), base ($_B$), and large ($_L$) sizes of the models are used. We report the min/max/average of performance across 5 runs using random seeds for fine-tuning initialization, or random sets of demonstrations in prompting.

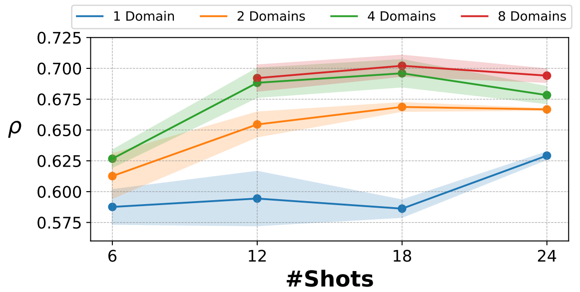

Figure 5: Effect of domain diversity of in-context examples on Llama2-7b performance on README++ (*en*). Correlation is greatly improved when examples are sampled from an increasing number of domains.

4.2 Unsupervised Methods

In the unsupervised setting, we leverage the LM distribution to compute a readability score without training. We also compare with several traditional length-based readability formulas.

LM-based Metrics. We use the Ranked Sentence Readability Score (**RSRS**) proposed by Martinc et al. (2021) which combines LM statistics with the sentence length. It computes a weighted sum of the individual word losses as follows:

$$\text{RSRS} = \frac{\sum_{i=1}^{S} [\sqrt{i}]^{\alpha}.\text{WNLL}(i)}{S}, \qquad (1)$$

where S is the sentence length, i is the rank of the word after sorting each Word's Negative Log Loss (WNLL) in ascending order. Words with higher losses are assigned higher weights, increasing the total score and reflecting less readability. α is equal to 2 when a word is an Out-Of-Vocabulary (OOV) token and 1 otherwise, assuming that OOV tokens represent rare, difficult words and thus are assigned higher weights by eliminating the square root. The WNLL is computed as follows:

$$\text{WNLL} = -(y_t \log y_p + (1-y_t)\log(1-y_p)), \quad (2)$$

where y_p is the predicted distribution by the LM, and y_t is the true distribution where the word appearing in the sequence holds a value of 1 while all other words have a value of 0.

Traditional Readability Metrics. We compare to several common traditional readability metrics (Ehara, 2021), which are based on word and sentence lengths. Specifically, we use the Sentence Length **(SL)**, Automated Readability Index **(ARI)** (Smith and Senter, 1967), Flesch-Kincaid Grade Level **(FKGL)** (Kincaid et al., 1975), and Open Source Metric for Measuring Arabic Narratives **(OSMAN)** (El-Haj and Rayson, 2016). The formulas for these metrics are provided in Appendix C.

4.2.1 Results

The results achieved by unsupervised methods are shown in Figure 6. We find that **LM-based RSRS scores achieve better correlation than traditional readability metrics in English. This was not the case for other languages, where performance was model-dependent.** Interestingly, for languages with non-Latin script (Arabic, Hindi, Russian), we find that RSRS scores computed via monolingual LMs achieve noticeably lower correlations compared to multilingual LMs. The RSRS metric (§4.2 Eq. 1) assumes that all unseen words by the LM's tokenizer are rare, difficult words that should be assigned higher weights. However, these could also be transliterations from other languages (e.g., names of new politicians or artists, emerging

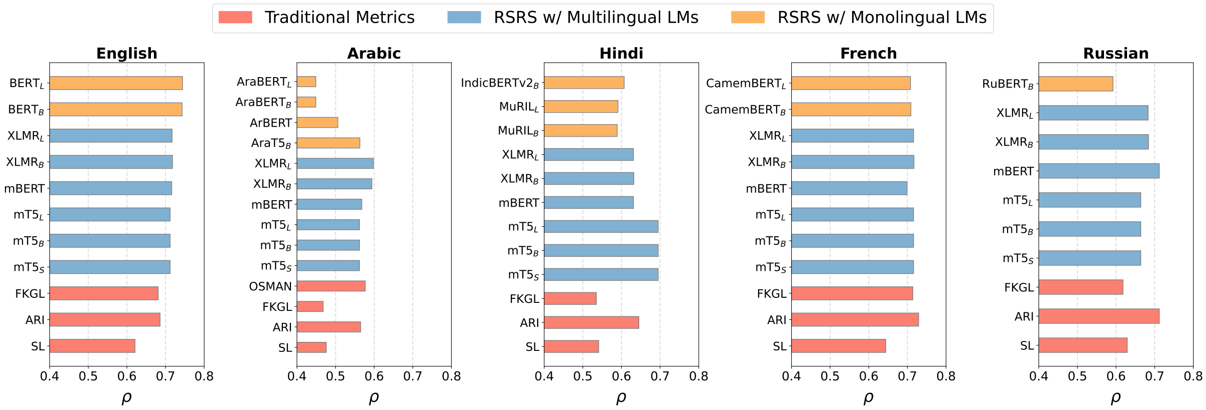

Figure 6: Pearson correlation (ρ) of **unsupervised** readability measurements on the test set of README++, including RSRS (Martinc et al., 2021) which leverages conditional word probabilities estimated by LMs. RSRS which uses multilingual LLMs performs better than RSRS which uses monolingual models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

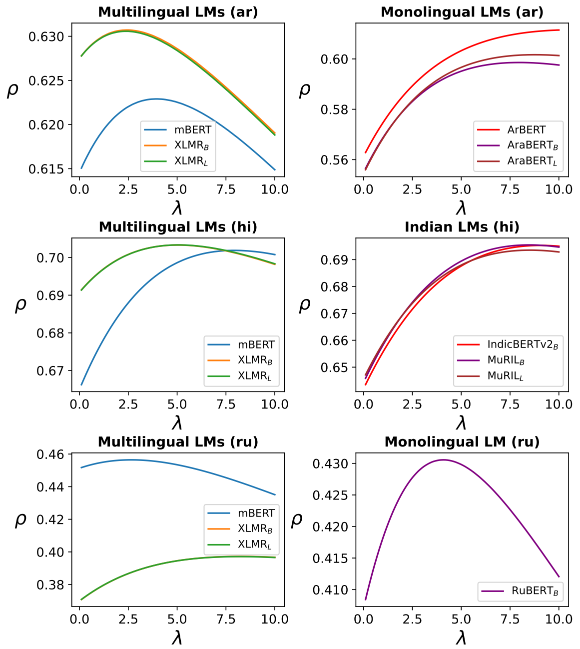

Figure 7: Effect of increasing the penalty factor (λ) on the Pearson correlation (ρ) between RSRS scores and human ratings for Arabic, Hindi and Russian sentences that contains transliterations. The plot shows a clear improvement in correlation as λ increases, which is more significant for monolingual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

International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1549–1564.

Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2016. Aspect based sentiment analysis in Hindi: resource creation and evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2703–2709.

Hend Al-Khalifa, Fetoun AlZahrani, Hala Qawara, Reema AlRowais, Sawsan Alowa, and Luluh AlDhubayi. 2022. A dataset for detecting humor in Arabic text. In *The 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022)*.

Marco Alfonse and Mariam Gawich. 2022. A novel methodology for Arabic news classification. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(2):e1440.

Bashar Alhafni, Reem Hazim, Juan David Pineros Liberato, Muhamed Al Khalil, and Nizar Habash. 2024. The SAMER arabic text simplification corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16079–16093.

Mohamed Aly and Amir Atiya. 2013. LABR: A large scale Arabic book reviews dataset. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 494–498.

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

Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. CEFR-based sentence difficulty annotation and assessment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6206–6219, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Abhinav Arora, Akshat Shrivastava, Mrinal Mohit, Lorena Sainz-Maza Lecanda, and Ahmed Aly. 2020. Cross-lingual transfer learning for intent detection of covid-19 utterances.

Udit Arora, William Huang, and He He. 2021. Types of out-of-distribution texts and how to detect them. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10687–10701, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, et al. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

Patrick Gustav Blaneck, Tobias Bornheim, Niklas Grieger, and Stephan Bialonski. 2022. Automatic readability assessment of German sentences with transformer ensembles. In *Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text*, pages 57–62.

Dominique Brunato, Lorenzo De Mattei, Felice Dell'Orletta, Benedetta Iavarone, and Giulia Venturi. 2018. Is this sentence difficult? do you agree? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2690–2699.

Susmoy Chakraborty, Mir Tafseer Nayeem, and Wasi Uddin Ahmad. 2021. Simple or complex? learning to predict readability of Bengali texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12621–12629.

Shuvamoy Chatterjee, Kushal Chakrabarti, Avishek Garain, Friedhelm Schwenker, and Ram Sarkar. 2021. JUMRv1: A sentiment analysis dataset for movie recommendation. *Applied Sciences*, 11(20):9381.

Alison Chi, Li-Kuang Chen, Yi-Chen Chang, Shu-Hui Lee, and Jason S Chang. 2023. Learning to paraphrase sentences to different complexity levels. *arXiv preprint arXiv:2308.02226*.

Kiyomi Chujo, Kathryn Oghigian, and Shiro Akasegawa. 2015. A corpus and grammatical browsing system for remedial EFL learners. *Multiple affordances of language corpora for data-driven learning*, pages 109–130.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard 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.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Simplicity level estimate (sle): A learned referenceless metric for sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Tobias Daudert and Sina Ahmadi. 2019. CoFiF: A corpus of financial reports in french language. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, pages 21–26.

Orphée De Clercq and Véronique Hoste. 2016. All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch. *Computational Linguistics*, 42(3):457–490.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Martin d'Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. 2020. FQuAD: French question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1193–1208.

Pavel Efimov, Andrey Chertok, Leonid Boytsov, and Pavel Braslavski. 2020. Sberquad–russian reading comprehension dataset: Description and analysis. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings 11*, pages 3–15. Springer.

Yo Ehara. 2021. Evaluation of unsupervised automatic readability assessors using rank correlations. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 62–72.

Mahmoud El-Haj and Paul Rayson. 2016. OSMAN — a novel Arabic readability metric. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 250–255.

Obeida ElJundi, Mohamad Dhaybi, Kotaiba Mokadam, Hazem M Hajj, and Daniel C Asmar. 2020. Resources and end-to-end neural network models for Arabic image captioning. In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP,*, pages 233–241. INSTICC, SciTePress.

AbdelRahim Elmadany, Muhammad Abdul-Mageed, et al. 2022. AraT5: Text-to-text transformers for arabic language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647.

Hady ElSahar and Samhaa R El-Beltagy. 2015. Building large Arabic multi-domain resources for sentiment analysis. In *International conference on intelligent text processing and computational linguistics*, pages 23–34. Springer.

Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R Arabnia. 2021. A brief review of domain adaptation. *Advances in Data Science and Information Engineering: Proceedings from ICDATA 2020 and IKE 2020*, pages 877–894.

Adam Fourney, Meredith Ringel Morris, Abdullah Ali, and Laura Vonessen. 2018. Assessing the readability of web search results for searchers with dyslexia. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1069–1072.

Nizar Habash and David Palfreyman. 2022. ZAEBUC: An annotated arabic-english bilingual writer corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 79–88.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343.

HindiMovieReviews. Hindi movie reviews dataset. https://www.kaggle.com/datasets/disisbig/hindi-movie-reviews-dataset. (Accessed on 05/03/2023).

Addison Howard, Deepak Nathani, Divy Thakkar, Julia Elliott, Partha Talukdar, and Phil Culliton. 2021. chaii - Hindi and Tamil question answering.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Joseph Marvin Imperial and Ekaterina Kochmar. 2023. Automatic readability assessment for closely related languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Joseph Marvin Imperial, Lloyd Lois Antonie Reyes, Michael Antonio Ibanez, Ranz Sapinit, and Mohammed Hussien. 2022. A baseline readability model for Cebuano. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 27–32.

Russian Jokes. Russian jokes dataset - Kaggle. https://www.kaggle.com/datasets/konstantinalbul/russian-jokes.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

Arnav Kapoor, Mudit Dhawan, Anmol Goel, TH Arjun, Akshala Bhatnagar, Vibhu Agrawal, Amul Agrawal, Arnab Bhattacharya, Ponnurangam Kumaraguru, and Ashutosh Modi. 2022. HLDC: Hindi legal documents corpus. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3521–3536.

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.

Nouran Khallaf and Serge Sharoff. 2021. Automatic difficulty classification of Arabic sentences. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 105–114.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. MuRIL: Multilingual representations for Indian languages. *arXiv preprint arXiv:2103.10730*.

J Peter Kincaid, Robert P. Fishburne Jr., Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for navy enlisted personnel. *Naval Technical Training Command Millington TN Research Branch*.

Diego Kozlowski, Elisa Lannelongue, Frédéric Saudemont, Farah Benamara, Alda Mari, Véronique Moriceau, and Abdelmoumene Boumadane. 2020. A three-level classification of french tweets in ecological crises. *Information Processing & Management*, 57(5):102284.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*.

Dieu-Thu Le, Cam-Tu Nguyen, and Xiaoliang Wang. 2018. Joint learning of frequency and word embeddings for multilingual readability assessment. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 103–107.

Justin Lee and Sowmya Vajjala. 2022. A neural pairwise ranking model for readability assessment. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3802–3813.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A learnable evaluation metric for text simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408, Toronto, Canada. Association for Computational Linguistics.

P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65.

Shrikant Malviya, Rohit Mishra, Santosh Kumar Barnwal, and Uma Shanker Tiwary. 2021. HDRS: Hindi dialogue restaurant search corpus for dialogue state tracking in task-oriented environment. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2517–2528.

Louis Martin, Benjamin Muller, Pedro Ortiz Suarez, Yoann Dupont, Laurent Romary, Éric Villemonte De La Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

John A McCarty and Larry J Shrum. 2000. The measurement of personal values in survey research: A test of alternative rating procedures. *Public Opinion Quarterly*, 64(3):271–298.

Mohsen Mesgar and Michael Strube. 2018. A neural local coherence model for text quality assessment. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4328–4339.

Rishabh Misra. 2022. News category dataset. *arXiv preprint arXiv:2209.11429*.

Babak Naderi, Salar Mohtaj, Kaspar Ensikat, and Sebastian Möller. 2019. Subjective assessment of text complexity: A dataset for German language. *arXiv preprint arXiv:1904.07733*.

Preslav Nakov, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, Abed Alhakim Freihat, Jim Glass, and Bilal Randeree. 2016. SemEval-2016 task 3: Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 525–545.

Tarek Naous, Wissam Antoun, Reem Mahmoud, and Hazem Hajj. 2021. Empathetic BERT2BERT conversational model: Learning Arabic language generation with little data. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 164–172, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Tarek Naous, Christian Hokayem, and Hazem Hajj. 2020. Empathy-driven Arabic conversational chatbot. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 58–68.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in NLP. *arXiv preprint arXiv:1608.07836*.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Quora.com. 2017. Quora question pairs. `https://www.kaggle.com/competitions/quora-question-pairs`.

Simin Rao, Hua Zheng, and Sujian Li. 2021. Cross-lingual leveled reading based on language-invariant features. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2677–2682.

Ankit Rathi. 2020. Deep learning apporach for image captioning in Hindi language. In *2020 International Conference on Computer, Electrical & Communication Engineering (ICCECE)*, pages 1–8. IEEE.

Biswarup Ray, Avishek Garain, and Ram Sarkar. 2021. An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews. *Applied Soft Computing*, 98:106935.

Shigehiko Schamoni, Julian Hitschler, and Stefan Riezler. 2018. A dataset and reranking method for multimodal mt of user-generated image captions. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 140–153.

Alok Singh, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2022. Attention based video captioning framework for Hindi. *Multimedia Systems*, 28(1):195–207.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, et al. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619*.

Sergey Smetanin. 2022. Rusentitweet: A sentiment analysis dataset of general domain tweets in russian. *PeerJ Computer Science*, 8:e1039.

Sergey Smetanin and Michail Komarov. 2019. Sentiment analysis of product reviews in russian using convolutional neural networks. In *2019 IEEE 21st Conference on Business Informatics (CBI)*, volume 01, pages 482–486.

Edgar A Smith and RJ Senter. 1967. *Automated readability index*, volume 66. Aerospace Medical Research Laboratories.

Sanja Štajner, Simone Paolo Ponzetto, and Heiner Stuckenschmidt. 2017. Automatic assessment of absolute sentence complexity. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI*, volume 17, pages 4096–4102.

Jeniya Tabassum, Mounica Maddela, Wei Xu, and Alan Ritter. 2020. Code and named entity recognition in StackOverflow. In *The Annual Meeting of the Association for Computational Linguistics (ACL)*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

TripAdvisor. Topic modelling on Trip Advisor dataset - Kaggle. https://www.kaggle.com/code/imnoob/topic-modelling-lda-on-trip-advisor-dataset/notebook.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Sowmya Vajjala. 2022. Trends, limitations and open challenges in automatic readability assessment research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5366–5377.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 297–304.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the seventh workshop on building educational applications using NLP*, pages 163–173.

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. 2021. From masked language modeling to translation: Non-English auxiliary tasks improve zero-shot spoken language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2479–2497.

Mengting Wan, Rishabh Misra, Ndapandula Nakashole, and Julian McAuley. 2019. Fine-grained spoiler detection from large-scale review corpora. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2605–2610.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591.

Orion Weller and Kevin Seppi. 2019. Humor detection: A transformer gets the last laugh. 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 3621–3625.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2019. Text readability assessment for second language learners. *arXiv preprint arXiv:1906.07580*.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Bang Yang, Fenglin Liu, Xian Wu, Yaowei Wang, Xu Sun, and Yuexian Zou. 2023. MultiCapCLIP: Auto-encoding prompts for zero-shot multilingual visual captioning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11908–11922. Association for Computational Linguistics.

Qingyu Zhang, Xiaoyu Shen, Ernie Chang, Jidong Ge, and Pengke Chen. 2022. MDIA: A benchmark for multilingual dialogue generation in 46 languages. *arXiv preprint arXiv:2208.13078*.

Jonathan Zheng, Ashutosh Baheti, Tarek Naous, Wei Xu, and Alan Ritter. 2022. Stanceosaurus: Classifying stance towards multilingual misinformation. *arXiv preprint arXiv:2210.15954*.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations Parallel Corpus v1. 0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534.

A More details about ReadMe++

A.1 Textual Domains

This section provides a description of how sentences were collected from each domain of ReadMe++. Table 15 shows statistics of the corpus and Table 16 summarizes the sources from which data was collected for each domain in each language, including publicly available web resources or open-source datasets.

- Wikipedia: Wikipedia is an attractive source of multilingual text since most articles are available in a large number of languages. Further, articles belong to a variety of topics where writing style and technicality differ significantly. We select 9 Wikipedia topics and, from each, randomly sample 5 different articles that discuss a certain sub-topic within that topic. For example, an article on *"Information Theory"* belongs to the *"Technology"* topic. We scrape the Arabic, English, French Hindi, and Russian versions of each article.

- News Articles: We leverage resources used for news category classification research, which we find publicly available datasets for in Arabic (Alfonse and Gawich, 2022) and English (Misra, 2022). No similar public resource was found for the other languages.

- Research: We collect text from medical, law, politics, and economics research papers in each language if available. We use open-access research archives such as arxiv[1] or HAL[2]. We also search for open-access research articles published under a Creative Commons license on Google Scholar using the same keyword in each language. We notice that research papers from natural sciences or technology are much less frequent in non-English languages as most researchers in those areas publish their work in English.

- Literature: We collect sentences from different types of literature *(Novels, History, Biographies, Children's Stories)* using books that are in the public domain. For English, French, and Russian, we use Project Gutenberg[3] that archives old books for which U.S. copyright has expired. For Arabic, we use Hindawi Books[4] which provide free Arabic books in many genres and topics. For Hindi, the law in India states that the copyright terms of books end 60 years after the death of an author and comes under the public domain[5]. Similar laws for most countries of the world are present with varying number of years[6]. We thus manually search for books in Hindi whose copyrights have expired according to these lengths. For example, we used Hindi novels by Premchand, Sarat Chandra Chattopadhyay, Rabindranath Tagore and Devaki Nandan Khatri.

- Textbooks: Textbooks are obtained from the Open Textbook Library[7] for English and Hindawi Books for Arabic which provide openly licensed textbooks. For Hindi textbooks, we use publicly available school textbooks from the National Council of Educational Research and Training in India [8] which provides books at various high-school levels and in different subjects. No similar openly available resource was found for French and Russian.

- Legal: We identify multiple governmental type of documents that we group under the "legal" domain, which include:

  Constitutions: We sample sentences from the U.S. constitution for English, the Lebanese constitution for Arabic, the Indian constitution for Hindi, the French constitution for French, and the Russian constitution for Russian.

  Judicial Rulings: We used recent public decisions by law courts, such as the Supreme Court in the US [9], to collect sentences from judicial rulings, in addition to using legal datasets with such content (Kapoor et al., 2022).

  United Nations Parliament: We collect samples from the United Nations (UN) Parallel Corpus (Ziemski et al., 2016) which contains official records and parliamentary documents of the UN.

[1] arxiv.org
[2] hal.science
[3] gutenberg.org
[4] hindawi.org
[5] https://copyright.gov.in/Documents/handbook.html
[6] en.wikipedia.org/wiki/List_of_countries%27_copyright_lengths
[7] open.umn.edu/opentextbooks/books
[8] ncert.nic.in/
[9] law.cornell.edu/supremecourt/text

The corpus is available all languages we consider except for Hindi since it is not considered one of the official languages of the UN.

- USER REVIEWS: User text reviews for products, movies, books, hotels, and restaurants, are sampled from open-source datasets in each language when available. Most these datasets are used in sentiment analysis research.

- DIALOGUE: Conversational text data is collected from three different types of open-source dialogue datasets: **Open-domain** dialogue datasets which focus on open-ended general conversation (Naous et al., 2021; Li et al., 2017; Zhang et al., 2022), **Task-oriented** datasets that are design to train human-assistance or customer support dialogue models (van der Goot et al., 2021; Malviya et al., 2021), and **Negotiation** dialogues that are used in developing automated sales dialogue agents with negotiation capabilities (He et al., 2018).

- FINANCE: We leverage the Financial Phrasebank dataset (Malo et al., 2014) which provides English sentences with financial references and content collected from finance-focused news, and the CoFiF corpus (Daudert and Ahmadi, 2019) which provides financial reports in French.

- FORUMS: We collect text from several online forums. These include:
Reddit: Reddit is a popular platform where online communities discuss common interests and passions. We used the latest version of the Reddit dump available at the time of this study to sample user posts. We filtered posts for language using the fasttext language identification model with a confidence > 0.9. NSFW and Over 18 content were automatically filtered before sampling. Further, any sampled sentence that still contained sexual or offensive content was manually removed.

  QA Websites: We collected questions and answers from QA websites using publicly available datasets for Question Answering research (Nakov et al., 2016; Quora.com, 2017; Howard et al., 2021; d'Hoffschmidt et al., 2020; Efimov et al., 2020).

  StackOverflow: Sentences were collected from the StackOverflow NER dataset (Tabassum et al., 2020) which contains user posts that describe what the user is trying to accomplish, a problem they are facing, or questions to seek advice from the community.

- SOCIAL MEDIA: We sample tweets from the the Stanceosaurus dataset (Zheng et al., 2022) which provides thousands of tweets in English, Arabic, and Hindi that discuss recent region-specific rumors. French tweets were sampled from the dataset of Kozlowski et al. (2020) built to detect crisis messages in French tweets, while Russian tweets were sampled from the RuSentiTweet dataset (Smetanin, 2022) for sentiment analysis in Russian. Tweets that include offensive or hate speech were manually omitted.

- POLICIES: We group under "Policies" several type of documents that delineate plans of what to do in a particular situation. This includes text extracted from: freely available **contract** templates for apartment/house leasing and job employment, **Special Olympics rules** which are available in multiple languages among which are but not in Hindi, and online **codes of conduct** of different organizations that we identify.

- GUIDES: Several domains that aim at providing instructions to the reader are grouped under "Guides". We extract data from Samsung Smartphones **User Manuals** which are available in a variety of languages. Another source is **Online Tutorials** which we collect from WikiHow that provides how-to articles in multiple languages. We also manually collect **Recipe Instructions** from multiple online cooking resources for each language. Additionally, we collect **Code Documentation** sentences from documentation of different functions of the Matlab software[10].

- CAPTIONS: We collect four different types of captions: image and video captions from various public datasets used in automatic captioning research, movie subtitles from the OpenSubtitles (Lison and Tiedemann, 2016) dataset used in machine translation research, and YouTube captions that we manually collect from video released under a Creative Commons license. While high-quality YouTube captions are easy to find for English, we could not find any high-

[10] mathworks.com

quality YouTube captions for non-English languages.

- MEDICAL TEXT: We use clinical reports written by medical professionals from the i2b2/VA dataset (Uzuner et al., 2011). We could not find similar high-quality medical resources for non-English languages.

- DICTIONARIES: We manually collect sentence examples from Arabic and English dictionaries using words that have appeared in the Word of the Day. No similar resource under a Creative Commons license was found for Hindi, French, and Russian.

- ENTERTAINMENT: We use Humour detection datasets to collect jokes (Al-Khalifa et al., 2022; Weller and Seppi, 2019; Jokes). Hindi jokes were manually collected.

- SPEECH: Two types of sources for speech data are used: **publicly available presidential speeches** that are usually posted on governmental websites. We used speeches by the United States President that are posted on the department of state's website. These speeches are also professionally translated to Arabic. We also collect sentences from **TED Talk transcriptions**, which are professionally translated from English to multiple languages.

- STATEMENTS: Two different types of standalone sentences that we group under "statements" were identified which are: Rumours, and quotes. We collect rumours in Arabic, English, and Hindi from the Stanceosaurus dataset (Zheng et al., 2022) used in misinformation detection. The rumours/claims are collected from various fact-checking websites in the Arab World, India, and the U.S. We also manually collected quotes in the three languages from various online resources. We did not collect mere translations of famous English quotes to other languages but focused on quotes by old scholars and thinkers of the Arab World, France, Russia and India for more cultural representation.

- POETRY: Poetry lines are extracted from English, Arabic, and Hindi poems, some of which date back several centuries ago. To have culture specific samples, we focus on non-English poems from original Arab, French, Indian, and Russian authors, and not poems translated from English.

- LETTERS: English letters were collected from online archives of historic letters. No high-quality authentic letters were found in Arabic or Hindi.

A.2 Domain Distribution

Table 6 shows the distribution of the domains in each readability level for each language. Basic readability levels (A1, A2) mostly contains sentences from domains that have text that is straightforward to read and contains day-to-day vocabulary such as Captions, Dialogue, User Reviews, User Guides. Intermediate readability levels (B1, B2) largely contain sentences from domains that present factual content such as books, Wikipedia articles, policy documents, news articles, etc. Proficient levels (C1, C2) contain domains that are scientific and technical such as finance, medical, legal documents, or highly literary text such as Arabic Poetry. We show the distribution of readability levels per domain in Figure 8.

A.3 Sentence Examples

Example sentences from various domains are shown in Table 13 for English, Table 14 for Arabic, Figure 13 for Hindi, Figure 14 for French, and Figure 15 for Russian.

B CEFR Levels Descriptors

The CEFR levels descriptors are provided in Table 7. Each level is described by specific capabilities of a language learner which we used to familiarize annotators with the intuition behind the scale being used prior to labeling.

Lang	Readability Level	Distribution (>5%)
ar	A1	Captions (50.62%) Dialogue (28.4%) Reviews (7.41%)
	A2	Reviews (19.44%) Dialogue (18.65%) Guides (17.46%) Captions (12.7%) Social Media (5.45%) Literature (5.95%)
	B1	Wikipedia (22.37%) Reviews (15.76%) Guides (13.23%) News (10.12%) Speech (6.03%) Legal (5.84%)
	B2	News (21.59%) Wikipedia (21.06%) Reviews (6.9%) Entertainment (6.73%) Legal (6.55%) Policies (6.37%) Speech (5.31%)
	C1	Wikipedia (40.29%) Research (14.53%) Literature (13.43%) Textbooks (5.71%)
	C2	Poetry (24.04%) Wikipedia (26.23%) Novels (18.58%) Dictionaries (9.84%) Quotes (6.01%)
fr	A1	Captions (44.29%) Dialogue (9.29%) Twitter (8.57%) Poetry (7.86%) Quotes (5%)
	A2	Recipes (9.02%) Dialogue (12.02%) Twitter (7.1%) Quotes (7.1%) QA Websites (6.28%) Children Stories (5.46%)
	B1	Wikipedia (21.85%) Guides (15.32%) Books (10.36%) Legal (6.98%) Reddit (5.41%)
	B2	Wikipedia (43.47%) Legal (10.51%) Policies (9.66%) Books (7.39%) Guides (6.25%)
	C1	Wikipedia (46.47%) Policies (12.03%) Research (9.96%) Finance (7.74%)
	C2	Research (21.43%) Policies (7.14%) Finance (6.39%)
en	A1	Dialogue (38.25%) Captions (27.87%) Reviews (10.38%) Guides (5.46%)
	A2	Captions (16.74%) Reviews (13.33%) Statements (8.15%) Guides (10.03%) Dialogue (8.74%) Forums (7.41%) Entertainment (5.63%)
	B1	Wikipedia (16.72%) Reviews (13.85%) News (11.74%) Forums (7.8%) Guides (8.12%) Textbooks (7.17%)
	B2	Wikipedia (21.94%) News (11.8%) Research (10.8%) Textbooks (11.03%) Policies (7.83%) Literature (7.39%)
	C1	Wikipedia (24.23%) Research (13.14%) Literature (12.82%) Legal (9.54%) Textbooks (9.28%) Policies (5.67%) News (5.65%)
	C2	Wiki-Natural Sciences (16.25%) Literature (18.75%) Clinical Reports (11.25%) Research (8.7%) Textbooks (7.5%)
hi	A1	Captions (33.09%) Literature (16.91%) Dialogue (12.82%) Jokes (9.56%) Reviews (5.15%)
	A2	Captions (12.88%) Dialogue (12.88%) Forums (7.46%) Statements (7.46%) Children Stories (6.78%) (5.37%) Guides (5.76%)
	B1	Wikipedia (15.02%) Literature (13.31%) Guides (11.26%) Reviews (9.56%) Statements (8.53%) Forums (8.53%)
	B2	Wikipedia (21.27%) Textbooks (9.7%) Literature (9.33%) Poetry (8.96%) Research (7.46%) Policies (7.46%) Quotes (5.6%)
	C1	Wikipedia (31.08%) Textbooks (12.16%) Legal (10.36%) Research (10.36%) Literature (8.53%) Forums (7.21%) Poetry (5.41%)
	C2	Wikipedia (44.25%) Textbooks (10.92%) Legal (10.9%) Research (8.05%)
ru	A1	Reviews (10.7%) Recipes (9.2%) Twitter (9.45%) Dialogue (8.21%) Jokes (7.96%) Captions (5.97%)
	A2	Wikipedia (23.80%) Guides (15.36%) Research (8.19%) Speech (7.14%)
	B1	Wikipedia (32.76%) Guides (6.11%) Policies (5.62%) Legal (5.62%)
	B2	Wikipedia (34.05%) Research (20.86%) Legal (12.88%) Policies (9.51%) Community Websites (6.13%)
	C1	Wikipedia (31.65%) Research (26.16%) Legal (19.38%) Policies (8.81%)
	C2	Legal (28.42%) Research (17.58%) Policies (6.59%)

Table 6: Distribution of domains for each readability level in each language. Only domains that compose more than 5% of the distribution are show.

C Traditional Metrics

ARI and FKGL are statistical formulas based on the number of words, characters, and syllables.

Automated Readability Index (ARI). ARI aims at approximating the grade level needed by an individual to understand a text. It is computed by:

$$\mathrm{ARI} = 4.71\left(\frac{\#\mathrm{Chars}}{\#\mathrm{Words}}\right) + 0.5\left(\frac{\#\mathrm{Words}}{\#\mathrm{Sents}}\right) - 21.43 \tag{3}$$

Flesch-Kincaid Grade Level (FKGL). FKGL also aims at predicting the grade level, but unlike ARI, considers the total number of syllables in the text. It is computed as follows:

$$\mathrm{FKGL} = 0.39\left(\frac{\#\mathrm{Words}}{\#\mathrm{Sents}}\right) + 11.8\left(\frac{\#\mathrm{Sylla}}{\#\mathrm{Words}}\right) - 15.59 \tag{4}$$

Open Source Metric for Measuring Arabic Narratives (OSMAN). OSMAN is computed according to the following formula:

$$\mathrm{OSMAN} = 200.791 - 1.015\left(\frac{A}{B}\right) + 24.181\left(\frac{C}{A} + \frac{D}{A} + \frac{G}{A} + \frac{H}{A}\right) \tag{5}$$

where A is the number of words, B is the number of sentences, C is the number of words with more than 5 letters, D is the number of syllables, G is the number of words with more than four syllabus, and H is the number of "Faseeh" words, which contain any of the letters (ظ، ذ، ؤ، ئ، ء) or end with (ون، وا).

D Experimental Details

D.1 Language Models

The details of the pre-trained LMs used in our experiments are provided in Table 8, including the number of parameters and pre-training data sources. The majority of models have been pre-trained using CommonCrawl data. Aya is based on mT5$_{XXL}$ and further instruction-tuned using the Aya dataset (Singh et al., 2024). Training was performed using

CEFR Level	Description
A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.
A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. basic personal information, employment, etc.). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.
B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.
B2	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
C1	Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.
C2	Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.

Table 7: Level descriptions of the CEFR scale used for readability annotation.

Model	#Params	Pre-training Sources			
		Wiki	News	Books	CC
Multilingual LMs					
mBERT	177M	✓			
XLMR_B	278M				✓
XLMR_L	559M				✓
mT5_S	60M				✓
mT5_B	220M				✓
mT5_L	770M				✓
Aya101	13B				✓
Monolingual Arabic LMs					
AraBERT_B	135M	✓	✓		
AraBERT_L	369M	✓	✓		✓
ArBERT	163M	✓	✓	✓	✓
AraT5_B	220M	✓	✓	✓	✓
Monolingual French LMs					
CamemBERT_B	110M				✓
CamemBERT_L	335M				✓
Monolingual English LMs					
BERT_B	110M	✓		✓	
BERT_L	350M	✓		✓	
Indian LMs					
MuRIL_B	237M	✓			✓
MuRIL_L	506M	✓			✓
IndicBERTv2_B	278M		✓		✓
Monolingual Russian LMs					
RuBERT_B	180M	✓			

Table 8: Summary of LMs used in experiments. **CC** stands for Common Crawl.

four NVIDIA A40 GPUs. We fine-tuned Aya using LoRA (Hu et al., 2021) and 4-bit quantization. We set LoRa hyperparameters as follows: rank=8, alpha=16, dropout=0.05.

D.2 Corpus Split

The train/validation/test split statistics of README++ are shown in Table 9 for each language. Those splits are obtained based on taking a 60%/10%/30% split for train/validation/test per domain, ensuring all domains are covered in each split.

Lang	Split	Readability Class						
		$1_{(A1)}$	$2_{(A2)}$	$3_{(B1)}$	$4_{(B2)}$	$5_{(C1)}$	$6_{(C2)}$	Total
ar	#train	49	151	307	324	207	114	1152
	#val	6	25	53	62	35	17	198
	#test	26	76	154	179	108	52	595
fr	#train	78	226	270	200	144	72	990
	#val	13	35	34	44	22	15	163
	#test	49	105	140	108	75	39	516
en	#train	105	414	354	536	245	49	1703
	#val	20	61	64	99	30	8	282
	#test	58	200	210	272	113	23	876
hi	#train	158	182	170	148	121	118	897
	#val	29	27	27	28	29	12	152
	#test	85	86	96	92	72	44	475
ru	#train	235	174	252	191	151	49	1052
	#val	42	23	42	35	20	13	175
	#test	125	96	115	100	66	29	531

Table 9: Number of sentences per readability level for each data split of README++.

D.3 Few-shot Prompt

The prompt used for GPT3.5, GPT4, and Llama-7B is provided in Table 10. The prompt contains 5 primary parts: The task description, definition of readability, example CEFR levels, example sentences with readability scores, and finally the new sentence for evaluation. When investigating the importance of the few-shot demonstrations we modified how we sampled the few-shot examples from the training set, however the prompt scaffolding remained the same.

```
Rate the following sentence on it's readability level.  The readabilty is defined
as the cognitive load required to understand the meaning of the sentence.  Rate
the readabilty on a scale from very easy to very hard.  Base your scores off the
CEFR scale for L2 Learners.  You should use the following key:

1 = Can understand very short, simple texts a single phrase at a time, picking up
familiar names, words and basic phrases and rereading as required.
2 = Can understand short, simple texts on familiar matters of a concrete type
3 = Can read straightforward factual texts on subjects related to his/her field
and interest with a satisfactory level of comprehension.
4 = Can read with a large degree of independence, adapting style and speed of
reading to different texts and purpose
5 = Can understand in detail lengthy, complex texts, whether or not they relate
to his/her own area of speciality, provided he/she can reread difficult sections.
6 = Can understand and interpret critically virtually all forms of the written
language including abstract, structurally complex, or highly colloquial literary
and non-literary writings.


EXAMPLES:
Sentence:  "[EX 1]"
Given the above key, the readability of the sentence is (scale=1-6):  [EX RATING 1]

Sentence:  "[EX 2]"
Given the above key, the readability of the sentence is (scale=1-6):  [EX RATING 2]

...

Sentence:  "[EX N]"
Given the above key, the readability of the sentence is (scale=1-6):  [EX RATING N]

Sentence:  "[SENTENCE]"
Given the above key, the readability of the sentence is (scale=1-6):
```

Table 10: Prompt provided to GPT4, GPT3.5, Aya23-8b, Llama2-7b, and Llama3.1-8b models to assess in-context learning readability assessment capabilities.

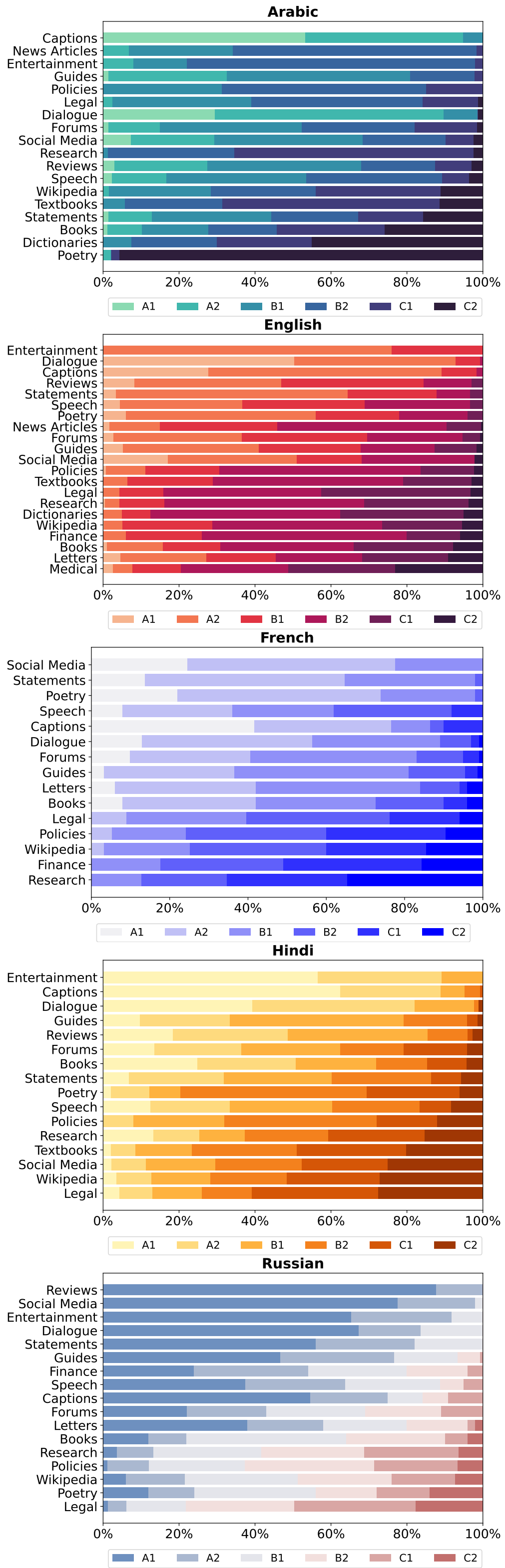

Figure 8: The readability levels vary greatly across domains and languages in README++, highlighting the importance to consider diversity of data sources.

E Additional Results

E.1 Main Results: Additional Metrics

The F1 scores obtained by the fine-tuned models are shown in Figure 9. We also report the Spearman Correlation (ρ_S) as an additional correlation measure in Figure 10. The same trends for models observed in §4.1 hold for other metrics.

E.2 Domain Correlation

To explore the utility of the large data diversity in README++, we investigate the performance of models trained on both README++ and CEFR-SP across several specific domains. We train XLMR$_L$ using the publicly available Wikipedia splits of CEFR-SP (1 data source) compared to the public data from README++ (112 data sources) The correlation of model predictions with human annotated labels are shown for 21 different textual domains in Figure 11. In 18 out of the 21 domains, the model trained on README++ clearly outperforms the model trained on CEFR-SP underscoring the importance of data diversity in fine-tuning LMs for readability assessment.

E.3 Zero-shot Cross Lingual Transfer

The zero-shot cross lingual results for several multilingual models are shown in Table 11. Similar to what is observed in §5, fine-tuning on README++ leads to significantly better cross-lingual transfer to 6 different target languages compared to fine-tuning on previous datasets. The improvement and trend is consistent across various models. We provide in Table 12 per-domain correlation results of XLMR$_L$ when transferring to Arabic, French, Hindi, and Russian, where we see superiority across domains by the model fine-tuned on README++ compared with fine-tuning on the single-domain Wikipedia-based CEFR-SP.

E.4 Effect of Context

We study the effect of providing models with context during training, which consists of up to three sentences that precede a sentence lying within a paragraph, on performance in the supervised setting. We prepend the context to the input sentence when available and separate them with a `[SEP]` token. Figure 12 shows the results with and without the addition of context when available. Overall, we find that pre-pending context information during fine-tuning decreased model performance in the majority of cases, or had little to no effect.

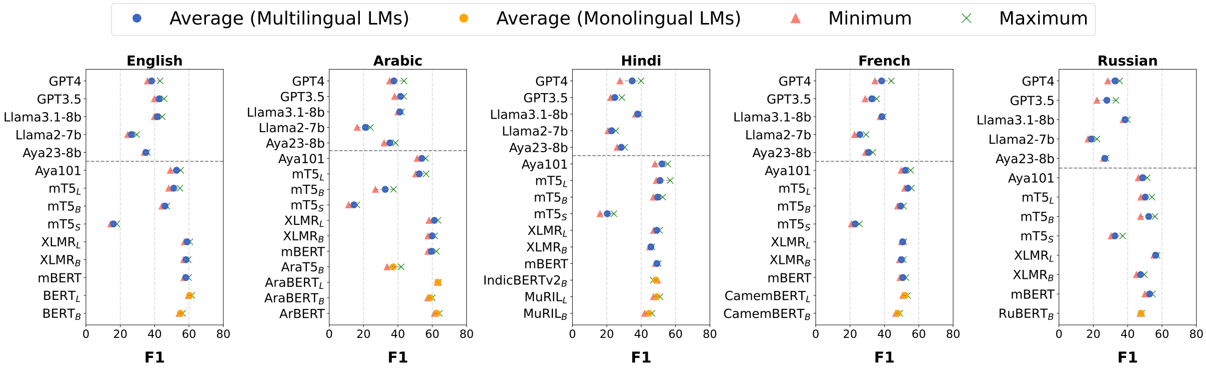

Figure 9: F1 score results of supervised fine-tuning and few-shot prompting on the test set of README++.

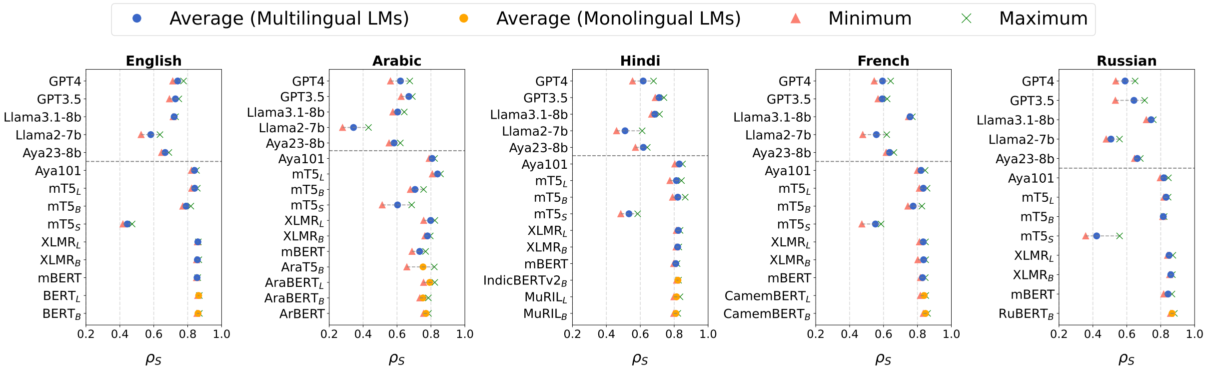

Figure 10: Spearman Correlation (ρ_S) of supervised fine-tuning and few-shot prompting on the test set of README++.

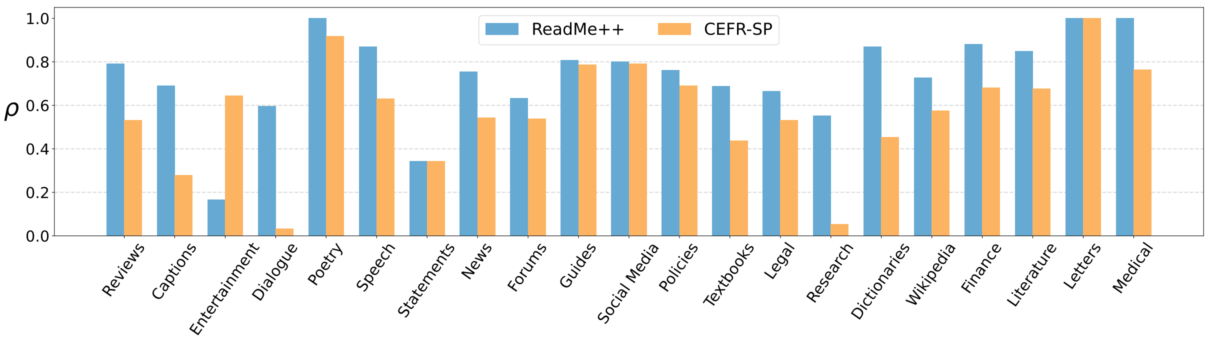

Figure 11: Pearson Correlation per domain for XLMR$_L$ trained using README++ and CEFR-SP. The model trained with README++ achieves better domain generalization, shown by higher correlation in all but one domain (Entertainment).

F Annotation Interface

Figures 16 and 17 show screenshots of our developed annotation interface for English sentences, where annotators perform a rank-and-rate approach to assign readability scores to 5 sentences in each batch. Annotators are asked to first rank sentences which they can do by simply dragging them. They are then asked to choose a rating for each sentence from a drop-down list. For each sentence, we provide the option to show its context, which shows the sentence in the paragraph to which it belongs. Figures 18 and 19 show screenshots of the interface for Arabic and Hindi respectively. An additional button to mark transliterations is added.

G License and Use Terms

We provide in Tables 18, 19, and 20 the license or usage term for each data source used in the creation of the corpus as follows:

- License: exact license under which data is avail-

Model	ReadMe++		CEFR-SP		CompDS	
	F1	ρ	F1	ρ	F1	ρ
en → ar						
mBERT	**19.94**	**0.512**	12.38	0.368	1.76	0.099
XLM-R$_B$	**32.63**	**0.645**	9.61	0.068	7.21	0.120
XLM-R$_L$	**31.48**	**0.606**	8.81	0.071	5.99	0.322
en → hi						
mBERT	**15.13**	**0.521**	8.72	0.375	6.45	0.171
XLM-R$_B$	**16.57**	**0.655**	9.87	0.146	9.81	0.398
XLM-R$_L$	**23.87**	**0.702**	13.15	0.267	10.38	0.381
en → fr						
mBERT	**30.63**	**0.751**	10.87	0.490	8.02	0.341
XLM-R$_B$	**33.96**	**0.746**	10.37	0.091	8.97	0.399
XLM-R$_L$	**30.29**	**0.768**	11.06	-0.026	5.92	0.335
en → ru						
mBERT	**16.25**	**0.610**	9.11	0.479	10.9	0.396
XLM-R$_B$	**21.27**	**0.671**	13.16	0.253	12.64	0.404
XLM-R$_L$	**24.60**	**0.760**	15.69	0.173	10.33	0.412
en → it						
mBERT	**12.79**	**0.270**	7.91	0.248	10.37	0.119
XLM-R$_B$	**14.38**	**0.295**	9.66	0.029	12.00	0.137
XLM-R$_L$	**14.68**	**0.239**	9.88	-0.043	10.06	0.099
en → de						
mBERT	**15.98**	**0.672**	12.51	0.595	6.88	0.347
XLM-R$_B$	**27.13**	**0.702**	14.02	0.196	8.68	0.529
XLM-R$_L$	**22.19**	**0.701**	10.00	-0.092	11.84	0.408

Table 11: Zero-shot cross-lingual transfer performance. Models fine-tuned on English data (en) of README++ significantly outperform models fine-tuned with CEFR-SP (Arase et al., 2022) or CompDS (Brunato et al., 2018) for Arabic (ar), Hindi (hi), Italian (it), and German (de).

able (CC BY 4.0 or other).

- Public Domain: data available in the public domain.
- Personal/Non-Commercial: source grants usage permission of data for personal/non-commercial purposes.
- (—): denotes that data needs to be requested from authors.

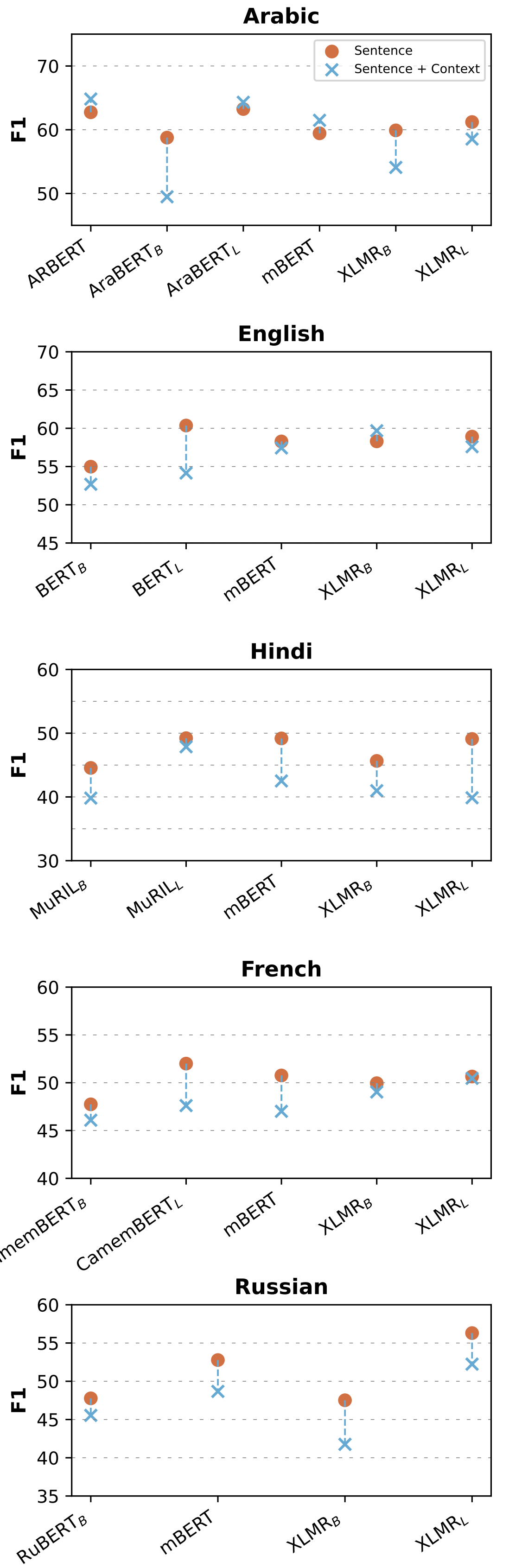

Figure 12: Effect of providing context during fine-tuning.

Domain	en → ar		en → fr		en → hi		en → ru	
	ReadMe++	**CEFR-SP**	**ReadMe++**	**CEFR-SP**	**ReadMe++**	**CEFR-SP**	**ReadMe++**	**CEFR-SP**
Captions	**0.545**	0.165	**0.551**	0.179	**0.336**	0.028	**0.644**	0.202
Dialogue	0.126	**0.269**	**0.635**	-0.387	**0.438**	0.122	**0.150**	-0.220
Dictionaries	-0.274	**0.000**	—	—	—	—	—	—
Entertainment	**0.374**	0.107	0.000	0.000	**0.657**	0.099	**0.397**	0.288
Finance	—	—	**0.784**	-0.013	—	—	**0.352**	-0.084
Forums	**0.440**	0.161	**0.564**	0.000	**0.603**	0.281	**0.737**	-0.109
Guides	**0.534**	0.024	**0.388**	-0.030	**0.362**	0.041	**0.438**	0.011
Legal	**0.277**	-0.093	**0.557**	-0.190	**0.362**	0.261	**0.782**	-0.220
Letters	—	—	**0.794**	0.000	—	—	**0.892**	0.214
Literature	**0.692**	0.081	**0.709**	-0.368	**0.561**	0.168	**0.498**	0.059
News	**0.447**	0.000	—	—	—	—	—	—
Poetry	0.000	0.000	**0.339**	-0.068	**0.202**	-0.347	**0.779**	0.112
Policies	**0.835**	0.009	**0.727**	-0.070	**0.551**	-0.427	**0.703**	0.144
Research	**0.562**	-0.021	**0.564**	0.154	**0.501**	-0.112	**0.647**	0.262
Social Media	**0.620**	0.313	**0.489**	-0.677	**0.341**	0.036	**0.452**	-0.106
Speech	**0.337**	-0.147	**0.618**	0.291	**0.668**	0.200	**0.583**	0.118
Statements	**0.374**	-0.019	**0.592**	-0.193	**0.331**	-0.013	**0.602**	-0.130
Textbooks	**0.600**	0.569	—	—	**0.427**	-0.201	—	—
User Reviews	**0.570**	0.240	—	—	**0.375**	-0.018	**0.000**	-0.196
Wikipedia	**0.644**	0.111	**0.625**	0.097	**0.630**	0.110	**0.715**	0.109

Table 12: Pearson Correlation per domain when performing cross lingual transfer to Arabic, French, Hindi, and Russian using XLMR$_L$ fine-tuned with README++ (en) vs CEFR-SP-WikiAuto (Arase et al., 2022).

LITERATURE - Novels
Over the river men were at work with spades and sieves on the sandy foreshore, and on the river was a boat, also diligently employed for some mysterious end. An electric tram came rushing underneath the window. No one was inside it, except one tourist; but its platforms were overflowing with Italians, who preferred to stand. Children tried to hang on behind, and the conductor, with no malice, spat in their faces to make them let go. Then soldiers appeared–good-looking, undersized men–wearing each a knapsack covered with mangy fur, and a great-coat which had been cut for some larger soldier. Beside them walked officers, looking foolish and fierce, and before them went little boys, turning somersaults in time with the band. The tramcar became entangled in their ranks, and moved on painfully, like a caterpillar in a swarm of ants. One of the little boys fell down, and some white bullocks came out of an archway. Indeed, if it had not been for the good advice of an old man who was selling button-hooks, the road might never have got clear.
MEDICAL - Clinical Reports
The patient underwent a flex sigmoidoscopy on Friday , 11-02 , which showed old blood in the rectal vault but no active source of bleeding. Given this , it was advised that the patient have a colonoscopy to rule out further bleeding
TEXTBOOKS - Engineering
The script might email information about the target user to the attacker, or might attempt to exploit a browser vulnerability on the target system in order to take it over completely. The script and its enclosing tags will not appear in what the victim actually sees on the screen.
FORUMS - StackOverflow
What's the best way to convert a string to an enumeration value in C# ?
USER REVIEWS - Product
First of all the package was shoved into my mail box and was basically crushed when I pulled it out. In addition there are deep marks and scrapes that show the wallet was used or pre-owned before getting to me..
STATEMENTS - Quotes
I may not have gone where I intended to go, but I think I have ended up where I needed to be.
WIKIPEDIA - Philosophy
Monarchies are associated with hereditary reign, in which monarchs reign for life and the responsibilities and power of the position pass to their child or another member of their family when they die.

Table 13: English Examples from several domains of READMe++. The sentence annotated for readability is highlighted in blue within the paragraph it belongs to, if applicable. Up to three preceding sentences of context to the sentence are highlighted in green if applicable.

LITERATURE - History

بل لقد كانت بدر بمثابة العَلَم الخفَّاق الذي يُرفِرف على ممتلكات الإسلام في قابل السنين والأعوام، كانت بدايةَ فتح خير دِين سمت مبادؤه، وتلألأت أضواؤه
حتى بلغت جبال الألب والبيرنيه غربًا، والصين واليابان شرقًا، وصار معتنقوه خمسمائة مليون من النفوس بعد أن كانوا نفرًا قليلًا، محمدًا وصحبه الأكرمين الأولين

Translation: Rather, Badr was like a fluttering flag that flutters over the possessions of Islam in the face of years and years. It was the beginning of the conquest of the best religion whose principles were elevated, and its lights sparkled. It reached the Alps and the Pyrenees in the west, and China and Japan in the east, and its adherents became five hundred million souls after they were a small number; Muhammad and his first noble companions.

NEWS ARTICLES - Sports

يستضيف ملعب كامب نو اليوم السبت انطلاقا من الساعة مساء نهائي كأس الملك بين برشلونة وأتلتيك بلباو فيما يلي التشكيلة المتوقعة بحسب صحيفة موندو ديبرتيفو

Translation: Today, Saturday, the Camp Nou stadium will host the King's Cup final between Barcelona and Athletic Bilbao. The following is the expected line-up, according to the Mundo Deportivo newspaper.

POLICIES - Contracts

جميع المصاريف والأتعاب الناشئة عن مماطلة أيّ من الطرفين في سداد الأقساط أو سداد مصاريف الصيانة، أو إزالة الضرر الناشئ بسببه
تعتبر جزءًا من التزاماته الأصلية، ويتعهد الطرف المماطل بدفعها

Translation: All expenses and fees arising from the delay of either party in paying the installments or paying the maintenance expenses, or removing the damage arising because of it, are considered part of their original obligations, and the party that caused the delay undertakes to pay them

GUIDES - Online Tutorials

يجب أن تضع الطائر بعيدًا عن الأطفال الصغار أو أي حيوانات أخرى قد تهاجمه أو تصيبه بإصابة أخرى دون قصد

Translation: You should keep the bird away from small children or other animals that might attack or otherwise inadvertently injure it

DICTIONARIES

ألا إن شر الروايا روايا الكذب

Translation: Verily, the most evil of stories are false stories

STATEMENTS - Quotes

العاقل لا يستقبل النعمة ببطر ولا يودعها بجزع

Translation: The wise person does not welcome a blessing with arrogance, nor does he become impatient when he loses it

POETRY

أَرِقتُ لَهُ وَالبَرقُ دونَ طَمِيَّةٍ

Table 14: Arabic sentence examples from README++. Note that a sentence in Arabic could be translated into multiple sentences in English.

Literature - Children's Stories

हाथी सियार की चापलूसी भरी बातों में आ गया.

Translation: The elephant got caught in the jackal's flattering words.

Entertainment - Jokes

चिंटू से एक आदमी ने पूछा- बेटा, आपके पापा का क्या नाम है?चिंटू- अंकल, अभी उनका नाम नहीं रखा मैंने, बस प्यार से पापा ही कहता हूं.

Translation: A man asked Chintu - Son, what is your father's name? Chintu - Uncle, I have not named him yet, I just call him father with love.

Speech - Ted Talks

नई टेक्नोलॉजी, क्षमता बढ़ाने के साथ नई ज़रूरतें उत्पन्न करता है, जिसमें और संसाधन लगते हैं.

Translation: New technology, along with increasing capacity, creates new needs, which take up more resources.

Research - Law

इन्हीं दो सवालों के इर्द-गिर्द देश में भ्रामक वातावरण तैयार करने का प्रयास इन राजनीतिक दलों द्वारा किया जा रहा है और यह साबित किया जा रहा है कि यह कानून मुस्लिम-विरोधी है.

Translation: Efforts are being made by these political parties to create a misleading atmosphere in the country around these two questions and it is being proved that this law is anti-Muslim.

Wikipedia - Health

इनके अतिरिक्त विटामिन और खनिज तत्व पोषण के आवश्यक हैं.

Translation: Apart from these, vitamins and minerals are essential for nutrition.

Statements - Rumours

एमनेस्टी इंटरनेशनल पेगासस प्रोजेक्ट पर अपनी पहली रिपोर्ट से पीछे हट गया है.

Translation: Amnesty International has retracted its first report on the Pegasus project.

Wikipedia - Technology

एनटीपी-1999 के अनुसार ग्लोबल मोबाइल निजी संचार उपग्रह (जीएमपीसीएम) के लिए लाइसेंस प्रदान करने संबंधी नीति को 2 नवम्बर 2001 को अंतिम रूप दिया गया और इसकी घोषणा की गई.

Translation: The policy for grant of licenses for Global Mobile Private Communication Satellites (GMPCM) as per NTP-1999 was finalized and announced on 2 November 2001.

Figure 13: Hindi sentence examples from ReadMe++.

Wikipedia – History

Ce renouveau des dons va alors satisfaire la population égyptienne, et les temples, assurant la loyauté de ce contre-pouvoir durant les guerres des Diadoques.

Translation: This renewal of donations will then satisfy the Egyptian population, and the temples, ensuring the loyalty of this counter-power during the wars of the Diadochi.

Policies - Contracts

Pour le calcul de la durée de travail effectif hebdomadaire, les heures de présence responsable de jour sont prises en compte après leur conversion en heures de travail effectif.

Translation: To calculate the actual weekly working time, the hours of responsible presence during the day are taken into account after their conversion into actual working hours.

Statements - Quotes

Les mariages sont écrits dans le ciel.

Translation: Marriages are written in the sky.

Letters

Plusieurs fois elle fut recherchée en mariage; mais elle chérissoit trop l'indépendance pour contracter un pareil engagement.

Translation: She was sought in marriage several times; but she cherished independence too much to enter into such a commitment.

Forums - Reddit

Les syndicats enseignants ont fait part de leurs inquiétudes et de leur surprise face à ces annonces.

Translation: The teachers' unions have expressed their concerns and surprise at these announcements.

Research - Science & Engineering

Certains champignons (biotrophes) vivent sur la cellule vivante, d'autres (nécrotrophes) la tuent avant de s'en nourrir.

Translation: Some fungi (biotrophs) live on the living cell, others (necrotrophs) kill it before feeding on it.

Figure 14: French sentence examples from ReadMe++.

Wikipedia – Mathematics

Несмотря на видимую простоту многих из них, такие доказательства используют свойства площадей фигур, доказательства которых сложнее доказательства самой теоремы Пифагора.

Translation: Despite the seeming simplicity of many of them, these types of proofs use properties of areas of figures, proofs of which are harder than the proofs of the Pythagorean theorem itself

Literature - Poetry

И на суше и водах, веслом и посохом, будут песнь и молитва бездны пасти, и осядет прахом взметенное порохом, и домой вернется пропавший без вести.

Translation: And on the land and sea, oar and staff, there will be songs and prayers at the abyss' mouth, and that which was thrown up by gunpowder will settle into dust, and the missing in action will return home.

Entertainment - Jokes

Сколько гостя не корми, он все равно напьется.

Translation: No matter how much you feed a guest, he will still get drunk.

Speech - TED Talks

В начале истории Америки характеру лидера придавалось большое значение, и мы ценили людей с богатым внутренним миром и высокой нравственностью.

Translation: At the start of American history, the character of the leader was given more value, and we valued people with rich inner peace and high morality.

Guides - Cooking Recipes

Кислота кваса должна приятно дополнять пресный или солоноватый вкус рыбы, а не противоречить ему.

Translation: The acidity of the Kvass should pleasantly add to the fresh or salty taste of the fish, not counteract it.

Research - Law

Русское общество с огромным интересом следило за первыми шагами нового суда.

Translation: Russian society watched the first steps of the new court with immense interest.

Figure 15: Russian sentence examples from ReadMe++.

Domain / Sub-Domain	ar	en	fr	hi	ru
WIKIPEDIA					
History	50	50	50	22	50
Geography	50	50	50	31	50
Philosophy	49	47	50	34	50
Technology	43	50	50	19	50
Mathematics	43	50	32	23	50
Art & Culture	49	50	50	35	50
Social Sciences	48	50	50	41	50
Natural Sciences	49	49	50	38	50
Health & Fitness	49	49	50	40	50
NEWS ARTICLES					
Sports	46	46	—	—	—
Politics	13	44	—	—	—
Culture	50	50	—	—	—
Economy	41	50	—	—	—
Technology	36	50	—	—	—
RESEARCH					
Law	36	19	—	13	50
Politics	19	22	—	19	50
Medical	—	30	31	—	50
Literature	—	39	—	28	—
Economics	26	46	—	31	50
Science & Engineering	—	30	47	—	50
LITERATURE					
Novels	50	50	50	48	50
History	40	45	50	47	—
Biographies	26	47	—	46	—
Children's Books	50	49	50	44	—
TEXTBOOKS					
Business	35	50	—	47	—
Psychology	—	50	—	47	—
Agriculture	—	50	—	—	—
Engineering	—	50	—	—	—
USER REVIEWS					
Products	50	40	—	33	49
Books	50	47	—	—	—
Movies	—	50	—	43	—
Hotels	50	48	—	—	—
Restaurants	50	47	—	—	—
DICTIONARIES	40	40	—	—	—
FORUMS					
Reddit	39	50	50	49	50
QA Websites	28	48	50	47	50
StackOverflow	—	50	—	—	—
SOCIAL MEDIA					
Twitter	41	47	50	44	49
POLICIES					
Contracts	27	34	45	—	41
Olympic Rules	40	50	50	—	50
Code of Conduct	—	50	—	50	—
GUIDES					
User Manuals	50	46	50	28	50
Online Tutorials	51	47	50	44	50
Cooking Recipes	40	48	50	47	50
Code Documentation	—	49	—	—	—
CAPTIONS					
Images	50	50	47	48	44
Videos	—	50	50	50	—
Movies	27	41	50	46	—
YouTube	—	42	—	—	—
MEDICAL TEXT					
Clinical Reports	—	39	—	—	—
ENTERTAINMENT					
Jokes	50	50	—	46	49
SPEECH					
Ted Talks	49	43	50	48	50
Public Speech	35	47	—	45	30
STATEMENTS					
Rumours	20	40	—	39	—
Quotes	50	50	50	49	50
DIALOGUE					
Open-domain	39	44	50	39	49
Negotiation	—	45	—	—	—
Task-oriented	39	50	50	50	—
LEGAL					
Constitutions	43	30	50	34	50
Judicial Rulings	—	21	—	35	47
UN Parliament	39	43	50	—	50
FINANCE	—	50	50	—	50
POETRY	46	50	50	49	50
LETTERS	—	22	50	—	50

Table 15: Dataset Statistics. (—) denotes that no public resource was found in the particular language.

Domain	Source		
Sub-Domain	ar	en	hi
WIKIPEDIA	wikipedia.com	wikipedia.com	wikipedia.com
NEWS ARTICLES	(Alfonse and Gawich, 2022)	(Misra, 2022)	—
RESEARCH			
Law	spu.sharjah.ac.ae	elgaronline.com	library.bjp.org
Politics	jcopolicy.uobaghdad.edu.iq	tandfonline.com	journal.ijarms.org
Medical	—	onlinelibrary.wiley.com	—
Literature	—	jstor.org/journal/jmodelite	hindijournal.com
Economics	asjp.cerist.dz/index.php/en	aeaweb.org	journal.ijarms.org
Science & Engineering	—	arxiv.org	—
LITERATURE	hindawi.org/books/	gutenberg.org	Public Domain Books
TEXTBOOKS	hindawi.org/books/	open.umn.edu	ncert.nic.in
LEGAL			
Constitutions	presidency.gov.lb	constitutioncenter.org	legislative.gov.in
Judicial Rulings	—	law.cornell.edu/supremecourt	HLDC (Kapoor et al., 2022)
UN Parliament	United Nations Parallel Corpus (Ziemski et al., 2016)		—
USER REVIEWS			
Products	(ElSahar and El-Beltagy, 2015)	MARC (Keung et al., 2020)	(Akhtar et al., 2016)
Books	LABR (Aly and Atiya, 2013)	(Wan et al., 2019)	—
Movies	—	JMURv1 (Chatterjee et al., 2021)	(HindiMovieReviews)
Hotels	(ElSahar and El-Beltagy, 2015)	(Ray et al., 2021)	—
Restaurants	(ElSahar and El-Beltagy, 2015)	(TripAdvisor)	—
DIALOGUE			
Open-domain	ArabicED (Naous et al., 2020)	DailyDialog (Li et al., 2017)	MDIA (Zhang et al., 2022)
Negotiation	—	CraigslistBargain (He et al., 2018)	—
Task-oriented	xSID (van der Goot et al., 2021)	xSID (van der Goot et al., 2021)	HDRS (Malviya et al., 2021)
FORUMS			
Reddit	Reddit Dump		
QA Websites	CQA-MD (Nakov et al., 2016)	quora.com (Quora.com, 2017)	(Howard et al., 2021)
StackOverflow	—	(Tabassum et al., 2020)	—
SOCIAL MEDIA			
Twitter	Stanceosaurus (Zheng et al., 2022)		
POLICIES			
Contracts	ejar.sa	honeybook.com	—
Olympic Rules	resources.specialolympics.org/translated-resources		—
Code of Conduct	—	fatimafellowship.com	lonza.com
GUIDES			
User Manuals	samsung.com/us/support/downloads		
Online Tutorials	ar.wikihow.com	wikihow.com	hi.wikihow.com
Cooking Recipes	ar.wikibooks.org	en.wikibooks.org	—
Code Documentation	—	mathworks.com	—
CAPTIONS			
Images	(ElJundi et al., 2020)	Flikr30K (Plummer et al., 2015)	(Rathi, 2020)
Videos	—	Vatex (Wang et al., 2019)	(Singh et al., 2022)
Movies	OpenSubtitles2016 (Lison and Tiedemann, 2016)		
YouTube	—	youtube.com	—
MEDICAL TEXT			
Clinical Reports	—	i2b2/VA (Uzuner et al., 2011)	—
DICTIONARIES	almaany.com	dictionary.com	—
ENTERTAINMENT			
Jokes	(Al-Khalifa et al., 2022)	(Weller and Seppi, 2019)	123hindijokes.com
FINANCE	—	(Malo et al., 2014)	—
SPEECH			
Ted Talks	ted.com/talks	ted.com/talks	ted.com/talks
Public Speech	state.gov/translations/arabic	whitehouse.gov	—
STATEMENTS			
Rumours	Stanceosaurus (Zheng et al., 2022)		
Quotes	arabic-quotes.com	goodreads.com/quotes	storyshala.in
POETRY	aldiwan.net	poetryfoundation.org	hindionlinejankari.com
LETTERS	—	oflosttime.com	—

Table 16: Dataset Sources (1/2). (—) denotes that no resource was found in the particular language.

Domain	Source	
Sub-Domain	fr	ru
WIKIPEDIA	wikipedia.com	wikipedia.com
RESEARCH	hal.science	ruscorpora.ru
LITERATURE	gutenberg.org	gutenberg.org
LEGAL		
Constitutions	legifrance.gouv.fr	constitution.ru
Judicial Rulings	—	supcourt.ru
UN Parliament	United Nations Parallel Corpus (Ziemski et al., 2016)	
USER REVIEWS		
Products	—	RuReviews (Smetanin and Komarov, 2019)
DIALOGUE		
Open-domain	MDIA (Zhang et al., 2022)	MDIA (Zhang et al., 2022)
Task-oriented	M-CID (Arora et al., 2020)	—
FORUMS		
Reddit	Reddit Dump	
QA Websites	(d'Hoffschmidt et al., 2020)	(Efimov et al., 2020)
SOCIAL MEDIA		
Twitter	(Kozlowski et al., 2020)	RuSentiTweet (Smetanin, 2022)
POLICIES		
Contracts	cesu.urssaf.fr	blanker.ru
Olympic Rules	resources.specialolympics.org/translated-resources	
GUIDES		
User Manuals	samsung.com/us/support/downloads	manuals.plus/ru
Online Tutorials	wikihow.com	
Cooking Recipes	wikibooks.org	
CAPTIONS		
Images	(Schamoni et al., 2018)	
Videos	citevideo-captions-fr	—
Movies	OpenSubtitles2016 (Lison and Tiedemann, 2016)	
ENTERTAINMENT		
Jokes	—	(Jokes)
FINANCE	(Daudert and Ahmadi, 2019)	ruscorpora.ru
SPEECH		
Ted Talks	ted.com/talks	ted.com/talks
Public Speech	—	ruscorpora.ru
STATEMENTS		
Quotes	evene.lefigaro.fr	infoselection.ru
POETRY	poesie-francaise.fr	ruscorpora.ru
LETTERS	gutenberg.org	runivers.ru

Table 17: Dataset Sources (1/2). (—) denotes that no resource was found in the particular language.

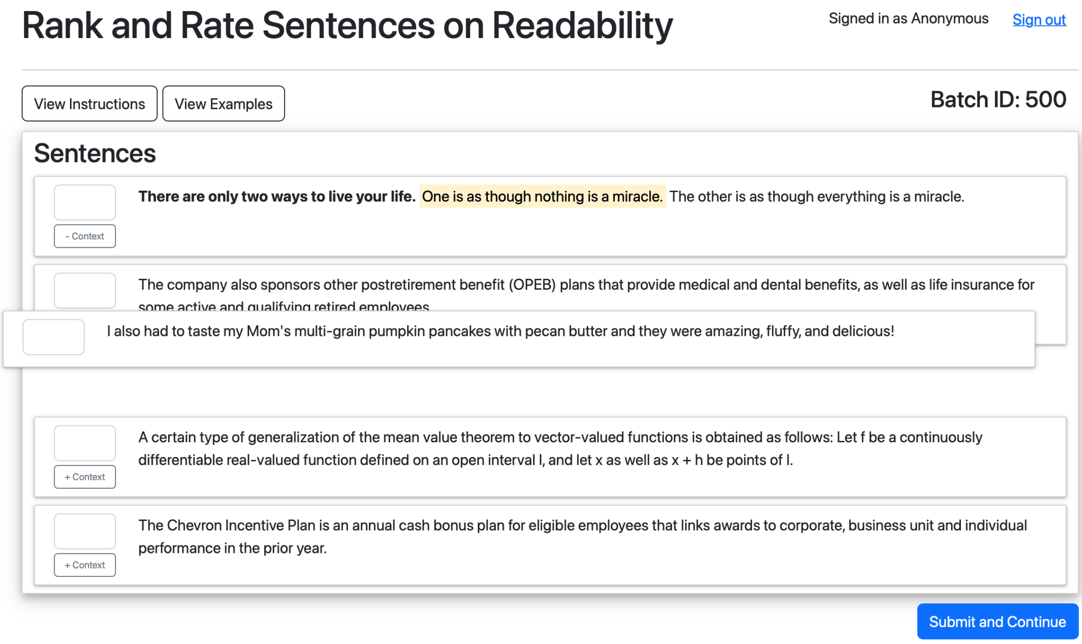

Figure 16: Screenshot of the developed annotation interface for rating English readability sentences. Annotators first rank sentences according to their readability level by simply dragging the box as shown in the figure. An optional Context button if available to show the context of a sentence if available.

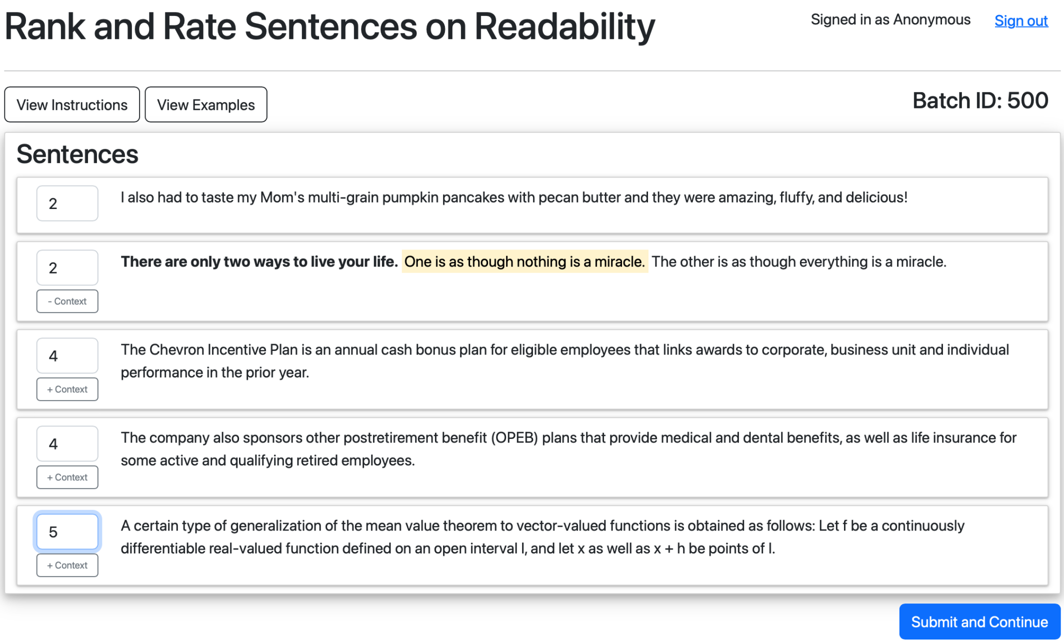

Figure 17: After ranking, annotators then assign a score for each sentence on a scale of 1 to 6 that corresponds to the CEFR levels. When done, annotators submit their scores and proceed to another batch of 5 sentences.

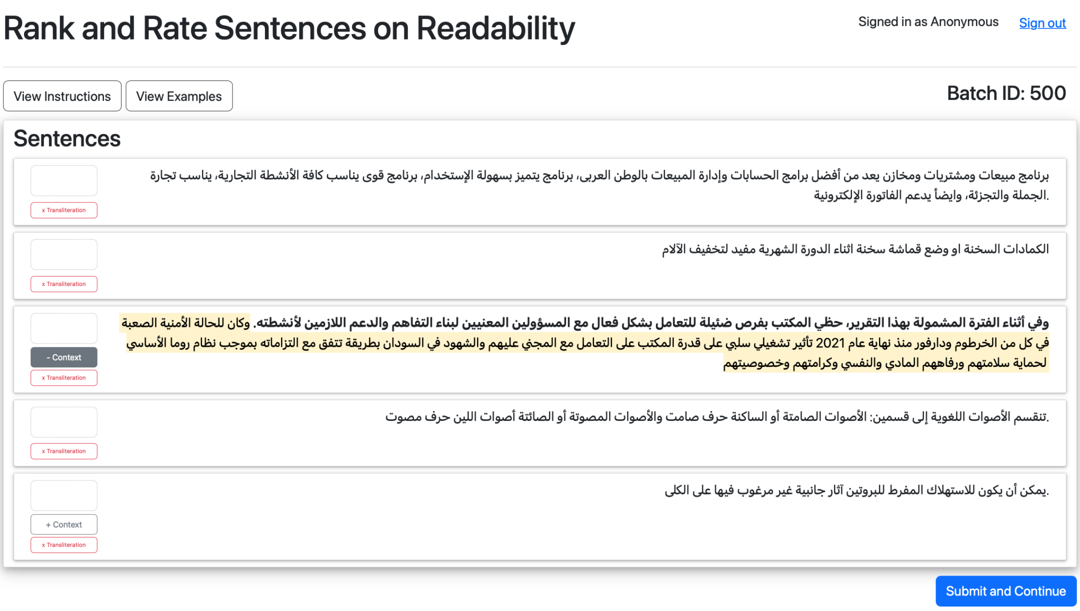

Figure 18: Screenshot of the developed annotation interface for Arabic sentences. An additional button to mark whether a sentence contains transliterations is provided.

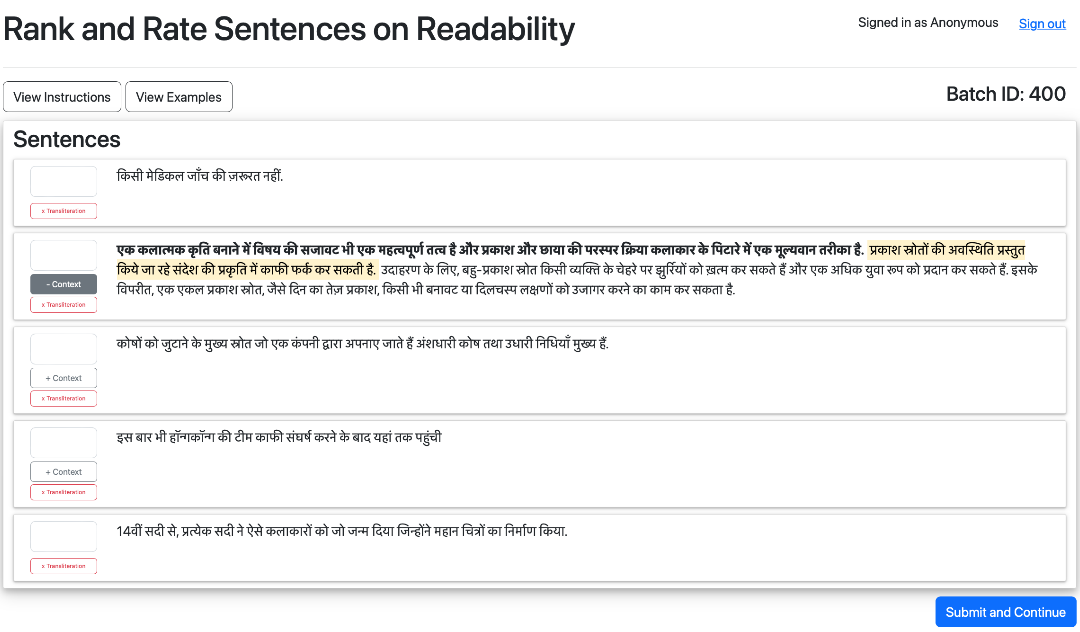

Figure 19: Screenshot of the developed annotation interface for Hindi sentences. An additional button to mark whether a sentence contains transliterations is provided.

Domain Sub-Domain	Source	Type	License
WIKIPEDIA	wikipedia.com	Web Article	CC BY-SA 3.0
NEWS ARTICLES	(Misra, 2022)	Public Dataset	CC BY 4.0
	(Alfonse and Gawich, 2022)	Public Dataset	CC BY 4.0
RESEARCH			
Law	spu.sharjah.ac.ae	Research Article	CC BY 4.0
	elgaronline.com	Research Article	CC BY 4.0
	library.bjp.org	Research Article	CC
Politics	jcopolicy.uobaghdad.edu.iq	Research Article	CC BY 4.0
	tandfonline.com	Research Article	CC BY 4.0
	journal.ijarms.org	Research Article	CC
Medical	onlinelibrary.wiley.com	Research Article	CC BY-NC
Literature	jstor.org/journal/jmodelite	Research Article	CC
	hindijournal.com	Research Article	CC
Economics	asjp.cerist.dz/index.php/en	Research Article	CC
	aeaweb.org	Research Article	CC BY 4.0
	journal.ijarms.org	Research Article	CC BY 4.0
Science & Engineering	arxiv.org	Research Article	CC BY 4.0
	hal.science	Research Article	CC
	ruscorpora.ru	Research Article	Personal/Non-Commercial
LITERATURE	hindawi.org/books/	Book	Public Domain
	gutenberg.org	Book	Public Domain
TEXTBOOKS	hindawi.org/books/	Book	Public Domain
	open.umn.edu	Book	CC BY 4.0
	ncert.nic.in	Book	Public Domain
LEGAL			
Constitutions	presidency.gov.lb	Document	Public Domain
	constitutioncenter.org	Document	CC BY-NC-ND 4.0
	legifrance.gouv.fr	Document	Public Domain
	legislative.gov.in	Document	Public Domain
	constitution.ru	Document	Public Domain
Judicial Rulings	law.cornell.edu/supremecourt	Document	CC BY-NC-SA 2.5
	HLDC (Kapoor et al., 2022)	Public Dataset	Public Domain
	supcourt.ru	Document	Public Domain
UN Parliament	UN Parallel Corpus (Ziemski et al., 2016)	Public Dataset	Public Domain

Table 18: License or term of use per source (1/3)

Domain	Source	Type	License
Sub-Domain			
USER REVIEWS			
Products	(ElSahar and El-Beltagy, 2015)	Public Dataset	Public Domain
	MARC (Keung et al., 2020)	Public Dataset	Public Domain
	(Akhtar et al., 2016)	On Request Dataset	—
	RuReviews (Smetanin and Komarov, 2019)	Public Dataset	Apache-2.0 License
Books	LABR (Aly and Atiya, 2013)	Public Dataset	GPL-2.0
	(Wan et al., 2019)	Public Dataset	Public Domain
Movies	JMURv1 (Chatterjee et al., 2021)	Public Dataset	Public Domain
	(HindiMovieReviews)	Public Dataset	CC BY-SA 4.0
Hotels	(ElSahar and El-Beltagy, 2015)	Public Dataset	Public Domain
	(Ray et al., 2021)	Public Dataset	CC BY 4.0
Restaurants	(ElSahar and El-Beltagy, 2015)	Public Dataset	Public Domain
	(TripAdvisor)	Public Dataset	Apache 2.0
DIALOGUE			
Open-domain	ArabicED (Naous et al., 2020)	Public Dataset	MIT License
	DailyDialog (Li et al., 2017)	Public Dataset	CC BY-NC-SA 4.0
	MDIA (Zhang et al., 2022)	Public Dataset	CC BY 4.0
Negotiation	CraigslistBargain (He et al., 2018)	Public Dataset	MIT License
Task-oriented	xSID (van der Goot et al., 2021)	Public Dataset	CC BY 4.0
	M-CID (Arora et al., 2020)	Public Dataset	Public Domain
	HDRS (Malviya et al., 2021)	Public Dataset	CC BY-NC 4.0
FINANCE	(Malo et al., 2014)	Public Dataset	CC BY-NC-SA 3.0
	CoFiF (Daudert and Ahmadi, 2019)	Public Dataset	CC BY-NC 4.0
	ruscorpora.ru	Document	Personal/Non-Commercial
FORUMS			
Reddit	files.pushshift.io/reddit	User Posts	Public Domain
QA Websites	CQA-MD (Nakov et al., 2016)	Public Dataset	Public Domain
	quora.com (Quora.com, 2017)	Public Dataset	Public Domain
	FQuAD (d'Hoffschmidt et al., 2020)	Public Dataset	Personal/Non-Commercial
	(Howard et al., 2021)	Public Dataset	Public Domain
	SberQuAD (Efimov et al., 2020)	Public Dataset	Apache-2.0 License
StackOverflow	(Tabassum et al., 2020)	Public Dataset	MIT License
SOCIAL MEDIA			
Twitter	Stanceosaurus (Zheng et al., 2022)	Public Dataset	Developer Agreement and Policy
	(Kozlowski et al., 2020)	Public Dataset	CC BY-NC 4.0
	RuSentiTweet (Smetanin, 2022)	Public Dataset	Public Domain
POLICIES			
Contracts	ejar.sa / hud.gov	Document	Public Domain
	cesu.urssaf.fr	Document	Public Domain
	blanker.ru	Document	Public Domain
	honeybook.com	Document	Public Domain
Olympic Rules	resources.specialolympics.org	Document	Personal/Non-Commercial
Code of Conduct	fatimafellowship.com	Web Article	Personal/Non-Commercial
	lonza.com	Document	Personal/Non-Commercial
GUIDES			
User Manuals	samsung.com/us/support/downloads	Document	Personal/Non-Commercial
	manuals.plus/ru	Web Article	Personal/Non-Commercial
Online Tutorials	wikihow.com	Web Article	CC BY-NC-SA 3.0
Cooking Recipes	wikibooks.org	Web Article	CC BY-SA 3.0
	narendramodi.in	Web Article	Personal/Non-Commercial
Code Documentation	mathworks.com	Documentation	Personal/Non-Commercial

Table 19: License or term of use per source (2/3)

Domain	Source	Type	License
Sub-Domain			
CAPTIONS			
Images	(ElJundi et al., 2020)	Public Dataset	Public Domain
	Flikr30K (Plummer et al., 2015)	Public Dataset	CC0
	WikiCaps (Schamoni et al., 2018)	Public Dataset	CC BY 4.0
	(Rathi, 2020)	Public Dataset	Public Domain
Videos	Vatex (Wang et al., 2019)	Public Dataset	CC BY 4.0
	MultiCapCLIP (Yang et al., 2023)	Public Dataset	BSD-3-Clause license
	(Singh et al., 2022)	Public Dataset	Public Domain
Movies	OpenSubtitles2016 (Lison and Tiedemann, 2016)	Public Dataset	Public Domain
YouTube	youtube.com	Captions	CC
MEDICAL TEXT			
Clinical Reports	i2b2/VA (Uzuner et al., 2011)	On Request Dataset	—
DICTIONARIES			
	almaany.com	Web Article	CC
	dictionary.com	Web Article	CC
ENTERTAINMENT			
Jokes	(Al-Khalifa et al., 2022)	Public Dataset	Public Domain
	(Weller and Seppi, 2019)	Public Dataset	MIT License
	(Jokes)	Public Dataset	Public Domain
	123hindijokes.com	Web List	Public Domain
SPEECH			
Ted Talks	ted.com/talks	Video Transcription	CC BY-NC-ND 4.0
Public Speech	state.gov/translations/arabic	Web Article	Public Domain
	ruscorpora.ru	Document	Personal/Non-Commercial
	whitehouse.gov	Web Article	CC BY 3.0 US
STATEMENTS			
Rumours	Stanceosaurus (Zheng et al., 2022)	Public Dataset	Public Domain
Quotes	arabic-quotes.com	Web List	Public Domain
	goodreads.com/quotes	Web List	Public Domain
	evene.lefigaro.fr	Web List	Personal/Non-Commercial
	storyshala.in	Web List	Public Domain
	infoselection.ru	Web List	Personal/Non-Commercial
POETRY	aldiwan.net	Web List	Public Domain
	poetryfoundation.org	Web List	Public Domain
	poesie-francaise.fr	Web List	Public Domain
	hindionlinejankari.com	Web List	Public Domain
	ruscorpora.ru	Document	Personal/Non-Commercial
LETTERS	oflosttime.com	Web Article	Public Domain
	gutenberg.org	Document	Public Domain
	runivers.ru	Document	Personal/Non-Commercial

Table 20: License or term of use per source (3/3)