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.

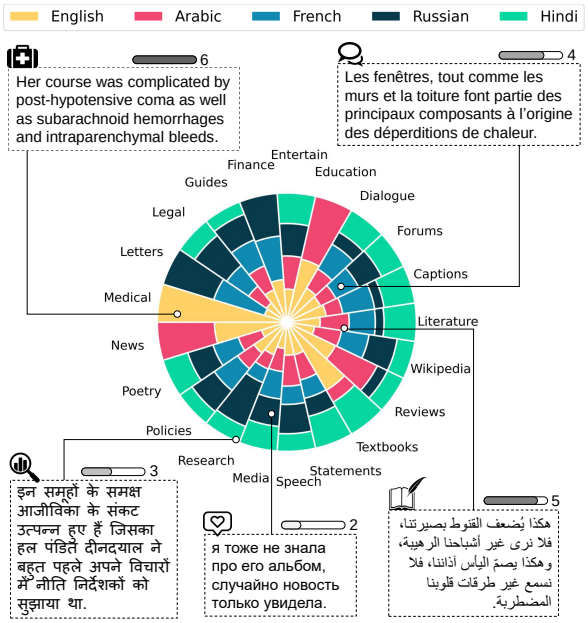

Figure 1: Language distribution per each domain in README++. Example sentences from each language are shown along with their human-annotated readability levels on a 6-point scale (1: easiest, 6: hardest).

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-

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).

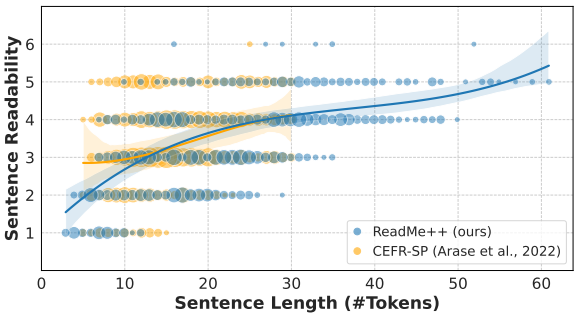

Figure 2: Distribution of sentence lengths across readability levels in the English portion of README++, compared with CEFR-SP (Arase et al., 2022). README++ offers a wider coverage of lengths and readability levels.

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 cat-