egorized 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 (Arase et al., 2022)	**WikiAuto**	0.66	0.73
	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 Corre-

lation (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.

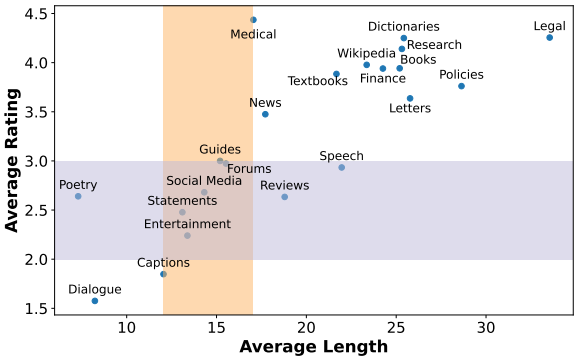

Figure 3: Average readability rating and sentence length per domain in the English portion of README++. Domain diversity presents additional challenges for readability assessment. Certain domains may be within the same readability range (e.g. [2, 3] that corresponds to A2 and B1 levels) but have varying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

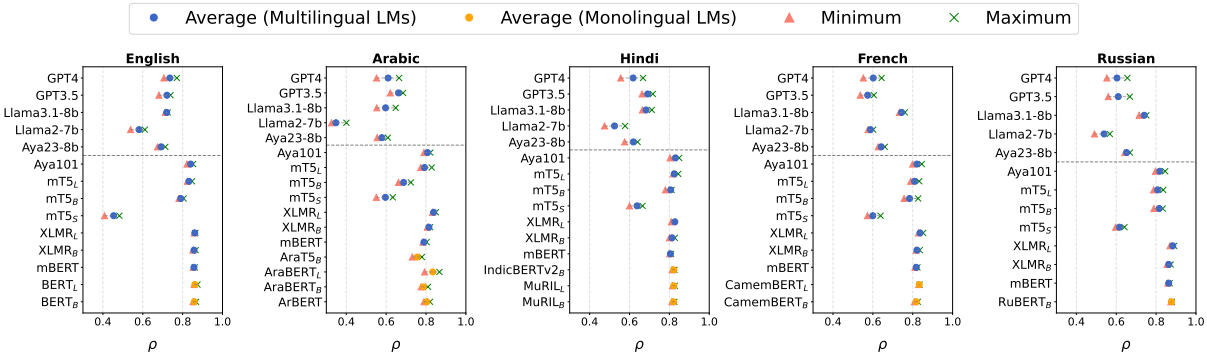

Figure 4: Pearson correlation (ρ) of **fine-tuned** multilingual and monolingual LMs, as well as **prompted** GPT3.5, GPT4, Aya23-8b, Llama2-7b, and Llama3.1-8b models with 5-shot examples, on the test set of README++. The small ($_S$), base ($_B$), and large ($_L$) sizes of the models are used. We report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

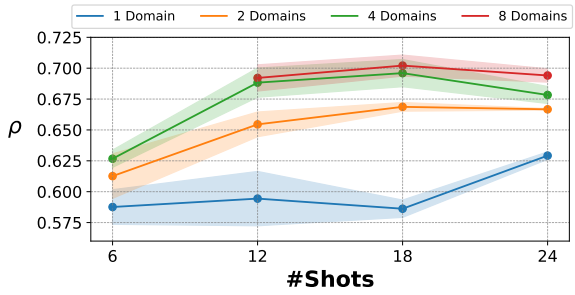

Figure 5: Effect of domain diversity of in-context examples on Llama2-7b performance on README++ (*en*). Correlation is greatly improved when examples are sampled from an increasing number of domains.

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