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.

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++.

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.

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.

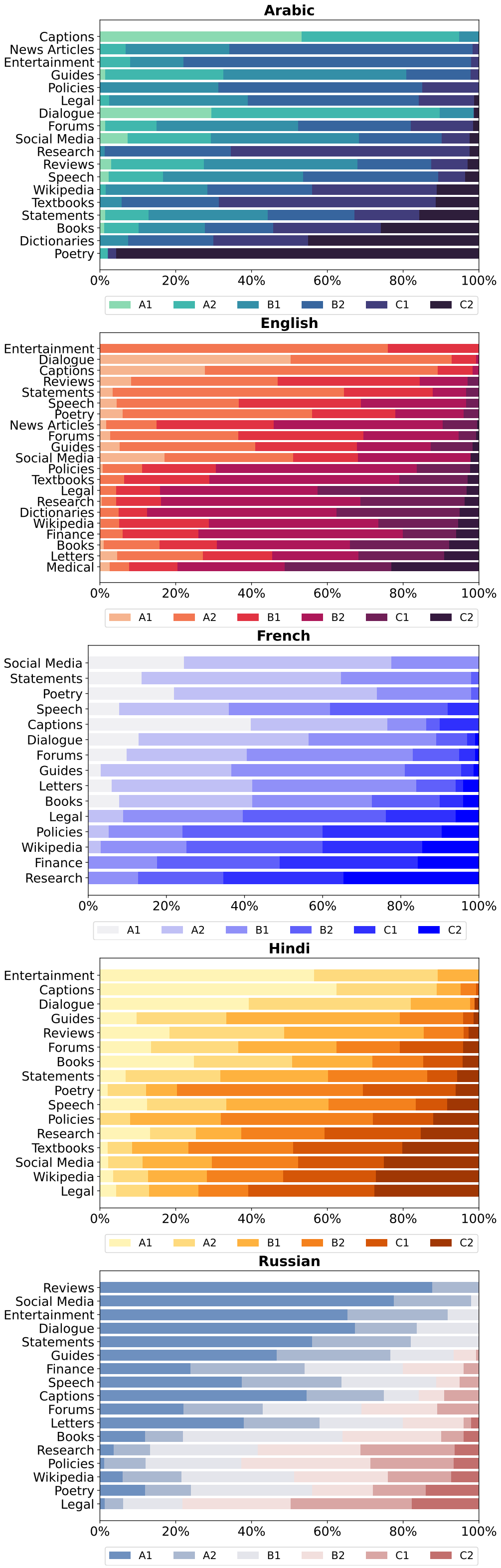

Figure 8: The readability levels vary greatly across domains and languages in README++, highlighting the importance to consider diversity of data sources.

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.