LANG	SENT	PARA	DOC	DIAG
EN	12,826	409,362	1,837	0
ES	0	713	31,355	0
DE	26,244	1,033	5,673	0
NL	0	0	3,596	0
CS	0	441	0	0
IT	0	813	0	0
FR	1,669	0	344	0
ET	0	420	1,277	0
PT	0	1,423	0	0
AR	1,945	215	0	0
HI	1,491	0	0	0
RU	1,758	0	0	0
CY	1,107	109	41	115
Total	**47,040**	**414,529**	**115**	**44,123**

Table 7: Data statistics of **UNIVERSALCEFR-FULL** in terms of levels (sentence, paragraph, document, dialogue) across the 13 target languages.

LANG	A1	A2	B1	B2	C1	C2
EN	173,005	119,335	59,634	20,746	7,122	675
ES	4577	4989	4,051	3,007	1,707	0
DE	273	13,208	12,996	346	108	308
NL	18	93	323	277	84	33
CS	1	92	77	38	2	0
IT	17	261	267	1	0	0
FR	106	302	404	335	210	98
ET	0	266	406	293	215	0
PT	204	62	270	59	80	0
AR	62	207	407	445	285	153
HI	203	219	223	203	182	145
RU	327	234	331	256	192	69
CY	463	332	0	0	0	0
Total	**179,256**	**139,600**	**79,389**	**26,006**	**10,187**	**1,481**

Table 8: Data statistics of **UNIVERSALCEFR-TRAIN** in terms of recognized CEFR levels (A1, A2, B1, B2, C1, C2) across the 13 target languages.

A Full Data Statistics

Tables 7, 9, 11and 13 report the quantity of CEFR-labeled texts across granularity levels per language, and Tables 3, 8, 10 and 12 reflect their counterparts in terms of CEFR level coverage. In forming the TEST split, we randomly sampled CEFR-labeled text instances per language per granularity level, while setting a cap of 200. This allows us to have a sizeable representation of UNIVERSAL-CEFR while maintaining efficiency for inference with LLMs. In total, we have 4,465 CEFR-labeled instances for UNIVERSALCEFR-TEST, which is comparable to the general sizes of benchmark test sets from previous works related to language proficiency (Naous et al., 2024; Zhang et al., 2024; Imperial and Tayyar Madabushi, 2024). For the TRAIN and DEV sets for fine-tuning and feature-based classification, we split the FULL subset (minus the TEST set) into a 90%-10% partition, respectively.

B Coverage of Large Language Models

In Table 14, we map each model's language coverage or language support based on its respective release papers and publications. Language support means what specific languages have been added and in substantial quantities in a model's training data (e.g., multilingual Wikipedia data dumps for pretraining XLM-R (Conneau et al., 2020)).

C Language-Specific Analysis

We provide in-depth analysis of model performances from the experiments in Section 5 across multiple dimensions of UNIVERSALCEFR on results for selected languages that we are qualified to interpret.

English. Analysis of model performance shows that using fine-tuned models and linguistic feature-based classification (62%-75%) obtains the best performance compared to prompting with instruction-tuned LLMs (19%-28%). However, these models tend to provide distinct patterns of specific CEFR labels. For the prompting setup, Gemma1, Gemma3, and EuroLLM models tend to give labels within the A1 and B1 range, while fine-tuned and feature-based models tend to lean towards the B1 and B2 range. For the pre-trained and instruction-tuned models, this finding may be tied to A1 and B2 being the most common CEFR level band of most general-purpose texts found online, where the sources of the data from which these models are trained. For feature-based models, we note the potential effect of training and test data having higher instance counts for these level bands than A1, C1, and C2. Regarding model scale, upgraded versions from similar model families perform better than their previous versions, echoing previous findings in literature (Imperial and Tayyar Madabushi, 2024). This is particularly evident in Gemma3 being 12B in size and trained with massively multilingual data in

LANG	SENT	PARA	DOC	DIAG
EN	12,826	409,362	1,837	0
ES	0	713	31,355	0
DE	26,244	1,033	5,673	0
NL	0	0	3,596	0
CS	0	441	0	0
IT	0	813	0	0
FR	1,669	0	344	0
ET	0	420	1,277	0
PT	0	1,423	0	0
AR	1,945	215	0	0
HI	1,491	0	0	0
RU	1,758	0	0	0
CY	1,107	109	41	115
Total	**47,040**	**414,529**	**115**	**44,123**

Table 9: Data statistics of **UNIVERSALCEFR-TRAIN** in terms of levels (sentence, paragraph, document, dialogue) across the 13 target languages.

LANG	A1	A2	B1	B2	C1	C2
EN	19,449	13,151	6,643	2,384	797	85
ES	1535	1226	904	471	285	0
DE	32	2,494	2,392	60	13	41
NL	6	70	235	230	99	32
CS	0	14	9	6	0	0
IT	3	33	23	1	0	0
FR	13	30	39	43	20	12
ET	0	19	52	21	25	0
PT	61	213	50	144	19	61
AR	7	26	56	53	35	15
HI	22	30	20	16	12	13
RU	34	23	25	34	21	9
CY	67	44	0	0	0	0
Total	**21,229**	**17,373**	**10,448**	**3,463**	**1,326**	**268**

Table 10: Data statistics of **UNIVERSALCEFR-DEV** in terms of recognized CEFR levels (A1, A2, B1, B2, C1, C2) across the 13 target languages.

140+ languages and obtaining 28% in weighted F1 compared to Gemma1, which is 7B in size and English-centric, obtaining 21.8%. We note a potential *default effect* in using these models where additional specific CEFR descriptor information is not needed if the texts being evaluated are in English, due to the majority of data in the context of CEFR that is reflected in the training data being English.

Spanish. Fine-tuned models outperform other setups, with feature-based approaches, especially Random Forest, achieving reasonable comparative performance. Moreover, multilingual models provide noticeable performance gains when compared to the English-only model. As per prompting strategy, for smaller multilingual models the language-specific prompt seems to play a role in improving the performance as it also does for the Gemma1 English-only model, however, the Gemma3 with 12B parameter is not affected by this, and it has been able to produce the best results of the LLMs (plus more sophisticated prompting strategies). As for the granularity of the input, models perform noticeably better at the document level than at the paragraph level, indicating that longer contexts are easier to classify than short ones. Finally, it is worth reporting a noticeable error of Gemma1: the prediction of C2 grade level, which does not exist in the Spanish dataset.

Hindi. Both the Gemma models perform poorly compared to the fine-tuned XLM-R and the Random Forest variants and tend to classify most Hindi test items as A1 or A2. For example, Gemma1 puts 57% of Hindi test samples as A1, whereas there are only 19% of the test samples labeled as A1 in the gold standard labels. This is in line with the general trend noticed in Section 6.2, as the Hindi subset is entirely sentence-level. The distribution is closer to the Gold distribution for the fine-tuned and feature-engineered models. XLM-R fine-tuned models give the best performance amongst all models for Hindi, both in terms of exact category prediction and in terms of the degree of error (i.e., being within 1 level above or below the correct level). Finally, we looked at the correlation between a simple approximation of text length (calculated as the number of space-separated tokens), a commonly used variable in such automated language assessment approaches in NLP research, and the CEFR gold labels, as well as model-predicted labels, after converting them to a numeric scale. There was a high correlation between text length and the gold labels (0.7), which was also seen with the XLM-R model (0.74) and the Random Forest models (0.77). However, the Gemma models only had correlations of 0.44 and 0.54, respectively, with text length. However, considering that the Hindi subset only has sentence-level annotations without a larger context, it may be challenging to achieve further consistency with the gold standard labels, given the size of the annotated dataset. Future

Lang	Sent	Para	Doc	Diag
EN	1,274	40,980	0	255
ES	0	51	0	4,370
DE	4,168	79	0	785
NL	0	0	0	672
CS	0	29	0	0
IT	0	60	0	0
FR	146	0	0	11
ET	0	19	0	98
PT	0	548	0	0
AR	188	4	0	0
HI	113	0	0	0
RU	146	0	0	0
CY	111	0	0	0
Total	**6,146**	**41,770**	**0**	**6,191**

Table 11: Data statistics of **UniversalCEFR-dev** in terms of levels (sentence, paragraph, document, dialogue) across the 13 target languages.

lang	A1	A2	B1	B2	C1	C2
EN	107	114	132	129	83	35
ES	49	58	140	108	45	0
DE	14	264	238	67	9	8
NL	4	21	69	77	22	7
CS	0	82	79	37	2	0
IT	9	87	104	0	0	0
FR	32	57	132	100	63	16
ET	0	110	130	93	67	0
PT	49	50	47	30	13	11
AR	12	26	162	145	40	15
HI	38	34	42	42	28	16
RU	41	36	52	35	24	12
CY	233	232	0	0	0	0
Total	**588**	**1,171**	**1,327**	**863**	**396**	**120**

Table 12: Data statistics of **UniversalCEFR-test** in terms of recognized CEFR levels (A1, A2, B1, B2, C1, C2) across the 13 target languages.

research should expand the available CEFR-graded resources both in terms of quantity as well as granularity for the language.

Russian. The Russian results follow the broad patterns reported in the paper, but their rich inflectional morphology and their comparatively limited training data amplify several effects. Gemma1 (34.8%) greatly over-predicts texts as beginner-level (only 5% of texts had predictions above B1), confirming the overall trend that small, English-centric LLMs struggle most with morphologically rich languages. Gemma3 (37.4%) partially corrects this, but still massively under-predicts B2 and C2. XLM-R (49.6%) mirrors the gold distribution most faithfully, possibly because its multilingual vocabulary gives it better coverage of Russian inflectional morphology, a pattern also seen for other highly inflected languages such as Czech. The two Random Forest models (47.2% and 47.8%) under-predict A2 and C2 but otherwise match the gold shape, showing that handcrafted lexical and morpho-syntactic features capture useful Russian-specific signals even with limited data. Subword-level multilingual models (XLM-R) or explicit morpho-syntactic features (RF) are best suited to capture the meanings and relations between Russian words. Text length appears to be a false friend; although it does correlate highly with readability (r=0.65), it also appears to be the source of many errors; top-performing model outputs had text length correlations as high as 0.73. Since this experiment with Russian is limited to sentence-level readability, comparison with previous research on Russian readability assessment is not straightforward. However, the weighted F1 (49.6%) of the best-performing model (XLM-R) is below state-of-the-art results for longer texts, including 67% (Reynolds, 2016), 74% (Solnyshkina et al., 2018), and 78% (Blinova and Tarasov, 2022). Most likely, this difference is partly due to the absence of Russian-specific morphosyntactic features that have been highly informative in previous studies' models.

Portuguese. Comparing the different setups, we can see that the results for Portuguese follow the global tendency, with fine-tuned models achieving the highest performance, followed by feature-based models, and with prompting taking the last place. Although this study only covers paragraph-level learner data for Portuguese, similar patterns were observed on reference data (Ribeiro et al., 2024b). However, comparing the results with those of other languages and, particularly, those with paragraph-level learner data, we can see that Portuguese is the language with the lowest performance (≈33.5%). Several factors may contribute to this outcome. For instance, Portuguese is one of the languages with the least available training data, and the distribution of proficiency labels is right-skewed (especially in COPLE2). Furthermore, the data consists of texts written by learners from a wide range of