Model & Setup	EN	ES	DE	NL	CS	IT	FR	ET	PT	AR	HI	RU	CY	Avg
Baseline														
Most Frequent Class	7.39	18.1	26.8	21.4	23.8	35.5	16.3	15.9	10.0	23.3	7.28	10.7	33.4	19.3
Gemma1-7B (English)														
Base	21.8	26.0	40.6	32.1	44.0	57.3	32.2	39.0	14.0	28.9	25.0	34.8	48.7	34.2
En-Read	20.5	28.3	31.0	23.5	53.6	41.0	22.7	24.9	27.2	29.5	8.4	18.0	55.7	29.6
En-Write	19.8	24.5	34.5	29.3	51.9	57.7	27.7	42.7	22.2	20.8	14.0	27.6	52.1	32.9
Lang-Read	20.5	29.3	35.1	37.8	55.3	48.0	27.1	44.6	20.2	32.2	12.8	26.2	52.8	34.0
Lang-Write	19.8	29.8	32.6	34.0	49.9	61.7	26.3	46.3	21.2	36.7	12.7	26.9	53.6	**34.7**
Gemma3-12B (Multi)														
Base	28.8	35.0	42.2	47.0	42.6	65.2	38.1	39.5	24.6	41.8	28.7	29.7	40.9	38.8
En-Read	19.3	25.5	35.8	25.5	18.5	22.9	29.3	26.0	9.8	33.3	14.8	21.2	20.5	23.3
En-Write	26.6	36.7	46.4	46.7	50.1	77.4	40.5	43.8	27.3	48.6	24.0	37.4	52.4	**43.2**
Lang-Read	19.3	28.1	35.2	37.6	50.9	64.8	35.0	30.4	26.1	29.5	20.5	32.5	61.6	36.3
Lang-Write	26.6	33.2	38.3	39.6	55.0	76.4	37.7	42.4	25.4	38.0	24.6	31.5	53.7	40.2
EuroLLM-9B (Multi)														
Base	18.6	25.4	28.0	29.1	25.0	39.9	25.9	32.0	16.4	34.3	12.7	15.1	14.4	24.4
En-Read	23.1	26.9	38.1	30.2	33.3	41.9	24.5	33.6	19.9	33.8	18.0	21.8	26.4	**28.6**
En-Write	21.5	26.2	29.8	32.0	32.4	33.1	26.8	32.8	21.1	31.8	17.7	17.5	24.5	26.7
Lang-Read	23.1	27.0	32.7	31.8	29.8	32.9	28.3	28.6	16.8	32.4	14.3	16.2	17.3	25.5
Lang-Write	21.5	28.5	35.1	30.1	30.8	30.6	27.6	29.9	16.5	35.2	21.0	16.1	8.80	25.5
fine-tuned Models														
ModernBERT (English)	75.8	71.8	72.1	54.2	66.9	82.7	47.2	88.3	33.5	30.8	51.6	48.9	73.2	61.3
EuroBERT (Multi)	74.6	72.0	70.6	53.2	63.9	79.7	42.0	86.6	32.1	35.4	44.7	45.9	79.9	60.0
XLM-R (Multi)	75.5	69.6	73.2	59.0	68.8	83.2	51.6	88.8	29.2	43.0	52.8	49.6	72.6	**62.8**
Feature-Based Models														
RandForest (TopFeats)	62.0	57.6	64.9	54.5	69.5	79.9	44.1	84.2	27.8	43.8	44.1	47.2	72.9	57.9
RandForest (AllFeats)	63.4	60.6	65.4	53.0	69.2	79.3	41.4	84.2	26.4	42.8	46.8	47.8	78.2	**58.3**
LogRegr (AllFeats)	32.1	28.2	50.9	47.1	62.9	81.9	41.7	67.5	23.1	34.1	47.8	41.1	63.8	47.9
LogRegr (TopFeats)	30.4	29.7	52.5	44.1	62.7	82.7	40.3	67.5	22.7	33.5	48.4	41.1	59.2	47.3

Table 4: Full weighted F1 performance results from the multilingual and English-centric model evaluation experiments using three setups (feature-based, fine-tuning, and prompting) and using UniversalCEFR-test split across the 13 languages. **Boldfaced** values indicate the highest scores overall per model setup, while underlined values highlight the highest scores for each model setup within each language.

for the class imbalance in CEFR level distribution and granularity across language subsets in UniversalCEFR-test. Using accuracy in the experiments would produce misleading performance in favor of any majority class.

6 Results

6.1 Model-Based Performance Comparison

Table 4 shows that, in terms of overall average performance across languages, the fine-tuned setup with ModernBERT, EuroBERT, and XLM-R achieved the highest weighted F1 score range (≈60%-62.8%) outperforming feature-based models (≈47%-58%) and prompting (≈23%-43%). Among the LLM-based approaches—prompting and fine-tuning—models trained on broader multilingual corpora generally performed better. For instance, XLM-R, which supports 100 languages, was the top performer, followed by EuroBERT (15 languages) and ModernBERT (English-only). A similar trend was observed in prompting: Gemma 3, trained on 140+ languages, outperformed EuroLLM (15 languages) and the English-centric Gemma 1, achieving the best prompting score of 43.2. These findings are consistent with previous work (Naous et al., 2024; Shardlow et al., 2024; Colla et al., 2023; Yuan and Strohmaier, 2021), reinforcing the usefulness of multilingual models for language proficiency assessment tasks. One limitation of our experimental setup, however, is that we did not include language-specific pre-trained models for languages other than English, which may have further improved performance for low-

Model	Sent	Para	Doc	All
Gemma1	19.41	**42.74**	30.81	33.63
Gemma3	38.71	**43.12**	39.62	42.33
XLM-R	62.67	66.38	**71.12**	65.92
RandForest-All	56.88	62.77	**64.58**	61.38
RandForest-Top	53.89	62.98	**64.94**	60.50

Table 5: Weighted F1 scores for top-performing unique model evaluation setups across granularities available for all languages.

Language	Learner	Reference
AR	41.92†	54.69
DE	71.14	**74.39**
EN	83.41	58.24
ES	**97.99**	42.72

Table 6: Average performances of the best models on learner text versus reference text across languages.† indicates performance with Gemma 3, and the rest refer to performance of the XLM-R model. Only these four languages have both learner and reference texts.

and mid-resource languages.

6.2 Granularity-Level Comparison

Table 5 highlights clear performance differences across text granularities (sentence, paragraph, and document) for all models, but more prominently for the Gemma models under prompting. Gemma 1, in particular, tends to over-predict lower CEFR levels (A1–B1) on sentence-level data, whereas its predictions on document-level subsets are more evenly distributed and better aligned with ground truth distributions. This suggests that prompt-based methods may require longer texts to make more accurate predictions, unlike models trained or fine-tuned on the respective datasets. Other models, such as XLM-R and Random Forest, show better results on document (≈64%-71%) and paragraph-level data (≈62%-66%) than sentence-level data (≈53%-62%), which was shown to be a more difficult task in previous work on readability (Dell'Orletta et al., 2011; Vajjala and Meurers, 2014). Regarding language-specific differences, among English, German, and Welsh, the best performance is seen with the paragraph-level dataset for English, the document-level dataset for German, and the sentence-level dataset for Welsh and French with the fine-tuned XLM-R model. Similar variations can be observed for other languages with more than one level of granularity (see Table 19). No single granularity or model shows consistently better performance across all tested languages. These results are likely due to the distribution of excerpts across granularity levels in each language (see Table 7 in Appendix A).

6.3 Learner-Reference Comparison

Four languages in UniversalCEFR contain both learner and reference texts: Arabic, German, English, and Spanish. Table 6 reports the average weighted F1 performance difference between the two categories across the four languages. For German, performance is comparable between learner and reference texts (≈71–74%). In contrast, English and Spanish show higher performance on learner texts (83% and 98%) than on reference texts (58% and 42%, respectively). Arabic displays the opposite trend: results on reference texts (54%) are much higher than those of learner texts, where the best results were obtained by Gemma 3 (41%). One possible explanation is that Gemma 3 may have been exposed to more Arabic content in its pre- and post-training phases.

7 Discussion

We discuss potential pathways through which UniversalCEFR can serve as a model, and offer key considerations for advancing data accessibility in language proficiency research.

Critical Reflections of Current Practices. The multiregional and multidisciplinary effort behind UniversalCEFR exposed significant inconsistencies and critical gaps in building CEFR-labeled language proficiency assessment corpora. Upon examination of annotation practices, *there appears to be no standard method for conducting expert annotations, including inconsistent use of inter-annotator agreement metrics and unclear guidelines on the number of annotators required to achieve reliable agreement*. This is reflected in the UniversalCEFR dataset itself, where nearly half of the corpora lack information on the annotators involved and their agreement scores. We posit that this may be due to diverse judgments of what constitutes high-quality data that does not require further human annotations.

In terms of language coverage, UniversalCEFR includes nine (EN, ES, DE, NL, CS, IT, FR, ET, PT) of the 24 recognized European languages. As a result, researchers working on these nine languages now have access to open, standard-

ized data for CEFR-based language proficiency assessment. The remaining 15 languages represent valuable opportunities for future expansion through collaborative efforts. While our open data and standardization initiative is a step towards addressing current challenges in interoperability and accessibility of resources, similar parallel efforts are needed in areas such as annotation and evaluation practices to ensure sustained progress in the language proficiency assessment community.

Need for Pro-Research Data Sharing Policies. As generative AI, particularly LLMs, becomes more ubiquitous, organizations that create valuable data for language proficiency assessment, such as publishers, educational institutions, and media outlets, are growing more cautious about how their resources are used. A major concern is the risk of data being used to train proprietary generative models, especially when such models are only accessible via commercial APIs that require transferring evaluation corpora to external servers. An example is the TCFLE-8 corpus (Wilkens et al., 2023) containing CEFR-labeled essays hosted by France Education International. Researchers seeking access to this dataset must explicitly specify that the resource will not be processed through commercial APIs to prevent potential data harvesting. To address these concerns, we believe *the community needs to agree on a unified pro-research data sharing policy with clear usage guidelines* for academic, non-commercial studies that require analysis of protected data with generative AI models without training on them.

Linguistic Features and Fine-tuning Still Matter. While recent advances in LLMs keep transforming NLP research, *our multilingual and multidimensional experiments in Section 6 reaffirm the continued value of linguistic features for traditional ML classifiers and fine-tuning pre-trained models* in language proficiency assessment. We observe common patterns where higher distribution and instance count lead to better results using these two setups (see performances on Spanish, English, and German subsets in Table 4) over prompting with CEFR descriptors. Moreover, using linguistic features in language proficiency assessment allows deeper analysis of language interactions with variables such as complexity, as seen in Appendix C. Given these insights, we encourage further efforts in the expansion of existing but low-resource language datasets with CEFR labels, as well as the exploration of features to better model morphologically-rich languages (e.g., Estonian and Portuguese). Together, these recommendations bridge current observed model failures to practical approaches in improving multilingual CEFR proficiency assessment.

8 Conclusion and Future Directions

In this work, we introduced UNIVERSALCEFR, a large-scale, open, multilingual, multidimensional dataset comprising 505,807 CEFR-annotated texts across 13 languages developed through global collaboration. Our findings from diverse model experiments with CEFR level prediction provide strong support for the utility of linguistic features and fine-tuning multilingual models in language proficiency assessment. Similarly, our critical analysis of the current data and resource-building practices emphasized the need for similar initiatives from the community, and pro-research data sharing policies in the advent of generative AI to remove barriers to accessibility without compromising data privacy and intellectual property.

Beyond its data and technical contributions, UNIVERSALCEFR also carries broader sociolinguistic significance. UNIVERSALCEFR addresses the growing linguistic inequality in modern AI development by focusing on underrepresented languages alongside English. We hope this initiative can lead to more responsible AI development that actively resists the growing linguistic centralization around English in global AI research—a modern *Matthew effect* (Merton, 1988)—where well-resourced languages receive disproportionate technological attention while smaller languages (like Czech or Welsh) are left behind (Masciolini et al., 2025). The UNIVERSALCEFR is a strong step towards mitigating the Matthew effect in language proficiency assessment research.

Limitations

We discuss several limitations of our work on UNIVERSALCEFR and how researchers can consider these directions to develop the resource further.

Natural Data Disparity in Experiments. From the statistics presented in Tables 3 and 7 for UNIVERSALCEFR, it is expected that not all languages have the exact same distribution of data across dimensions, including formats (sentence-, paragraph-, document-, and dialogue-level) and category (reference and learner texts). Hence, our main experi-