Field	**Description**
`title`	The unique title of the text retrieved from its original corpus (`NA` if there are no titles such as CEFR-assessed sentences or paragraphs).
`lang`	The source language of the text in ISO 638-1 format (e.g., `en` for English).
`source_name`	The source dataset name where the text is collected as indicated from their source dataset, paper, and/or documentation (e.g., `cambridge-exams` from Xia et al. (2016)).
`format`	The format of the text in terms of level of granularity as indicated from their source dataset, paper, and/or documentation. The recognized formats are the following: [`document-level`, `paragraph-level`, `discourse-level`, `sentence-level`].
`category`	The classification of the text in terms of who created the material. The recognized categories are `reference` for texts created by experts, teachers, and language learning professionals and `learner` for texts written by language learners and students.
`cefr_level`	The CEFR level associated with the text. The six recognized CEFR levels are the following: [`A1, A2, B1, B2, C1, C2`]. A small fraction (<1%) of text in UNIVERSALCEFR contains unlabelled text, texts with plus signs (e.g., `A1+`), and texts with no level indicator (e.g., `A, B`).
`license`	The licensing information associated with the text (`Unknown` if not stated).
`text`	The actual content of the text itself.

Table 15: The structured JSON fields with descriptions and examples used as the standardized uniform format for building the UNIVERSALCEFR dataset. All instances validated from the collection of CEFR-labelled corpora conform to this format.

Category	Feature Name
Length	doc_num_sents
	doc_num_tokens
	num_characters
	num_characters_per_sentence
	num_characters_per_word
	num_syllables_in_sentence
	num_syllables_per_sentence
	num_syllables_per_word
	num_words
Lexical	average_pos_in_freq_table
	lexical_complexity_score
Morphosyntactic	ratio_Tense_Past
	ratio_of_determiners
	ratio_of_numerals
	ratio_of_pronouns
Psycholinguistic	concreteness
	imagebility
Readability	sentence_fkgl
	sentence_fre
Syntactic	avg_distance_between_words
	average_length_VP
	parse_tree_height
	ratio_of_coordinating_clauses

Table 16: List of linguistic features occurring in the top 10 of at least three languages. We use this list for the TopFeatures subset used in the experiment result in Table 4.

A1 level, particularly for psycholinguistic features such as imageability ($\rho = 0.48$) and concreteness ($\rho = 0.46$), as well as punctuation-related measures. This suggests that certain surface-level and lexical-semantic features may be especially informative at the lowest proficiency level. A notable case is the feature of word length in characters, which shows a negative correlation at A1 ($\rho = -0.45$), becomes neutral at A2, and shifts to a positive correlation at B1 and higher levels. This pattern may reflect increasing lexical complexity with proficiency. Similarly, features related to syntactic structure, such as the ratio of past tense verbs and phrase length, generally shift from weak negative to weak positive correlations as proficiency increases, indicating progressive syntactic development. Overall, the directionality of several features suggests dynamic usage patterns across CEFR bands, even if the correlation strengths remain modest.

F Hyperparameter Values

We detail the hyperparameter values used for fine-tuning pretrained (ModernBERT, EuroBERT, and XLM-R) and instruction-tuned language models (Gemma1, Gemma3, and EuroLLM) in Tables 17 and 18, respectively.

Hyperparameter	Value
Learning rate	3.6×10^{-5}
Train batch size	2
Evaluation batch size	3
Random seed	42
Gradient accumulation steps	16
Total effective batch size	32
Optimizer	`adamw_torch_fused`
Betas	$(0.9, 0.999)$
Epsilon	10^{-8}
Learning-rate scheduler	linear
Warm-up ratio	0.1

Table 17: Hyperparameter values used for fine-tuning pretrained language models.

Hyperparameter	Value
Sampling	False
Max New Tokens	10
Data Type	`torch.bfloat16`
GPU	4 x NVIDIA RTX A5000 (24GB)

Table 18: Hyperparameter values and GPU information used for prompting instruction-tuned models.

Lang	Sent	Para	Doc	Overall
AR	**55.7**	32.6	-	43.1
CY	**86.9**	72.5	61.5	72.7
CS	-	68.8	-	68.8
DE	65.4	71.1	**83.4**	73.2
EN	68.3	**100.0**	57.6	75.5
ES	-	40.6	**98.0**	69.69
ET	-	**93.6**	84.0	88.9
FR	**57.6**	-	44.2	51.7
HI	52.9	-	-	52.9
IT	-	83.3	-	83.3
NL		-	59.0	59.0
PT	-	29.2	-	29.2
RU	49.6	-	-	49.6

Table 19: Weighted F1 scores for the fine-tuned XLM-R (top model across all setups) performance on the UniversalCEFR-test, classified by the granularity levels of the data.

G Additional Context on Restrictions of GDPR-Protected Datasets

The critical aspect of the GDPR is that it gives data subjects (e.g., L2 learners of CEFR) the right to withdraw their personal information from processing, which requires data processors to store both the signed consents and the ID mappings (i.e., mappings between the names of the real people and their IDs in a released corpora). As long as these documents exist and reidentification is theoretically possible, the data falls under the scope of the GDPR. Further complicating factors are national legislations and ethical regulations, such as archival laws, that treat any data produced at universities—including those used for language proficiency assessment such as essays, recorded dialogues, and written texts from personal experiences—as the property of the state (and hence making destruction of the ID mappings a non-trivial act) (European Parliament and Council, 2016).

Yet another upcoming challenge is the EU AI Act (European Parliament and Council, 2024) that implies that AI models trained on personal data should inherit the same license as the data they have been trained on, meaning that the models will be under the scope of the GDPR. We hypothesize that the non-restricted datasets included in UNIVERSALCEFR either do not contain personal information or were collected before the GDPR, since they are already openly accessible to the public. We further hypothesize that the datasets currently under the GDPR will eventually have their ID mappings destroyed and will no longer be subject to the GDPR. This may mean that the learner corpora that can be added to UNIVERSALCEFR will grow with time.

H Full Dataset Directory of UniversalCEFR

We provide the complete information of qualified corpora included in the current UNIVERSALCEFR collection to form a directory of datasets. Aside from eight per-instance information included in the standardized JSON format in Table 15, we also report five per-corpus information as listed below:

- Annotation method used (manual, computer-assisted, or NA).
- Total number of expert annotators.
- Distinct L1 learners per language for learner corpora.
- Inter-annotator agreement (IAA) metric and score.
- Reference to published paper or repository.

I Prompt Templates

We provide the complete copies of the prompt templates used in prompting experiments with instruction-tuned LLMs as described in Section 5. The prompt templates are categorized by color based on the setup: **BASE**, **EN-READ**, **LANG-READ**, **EN-WRITE**, **LANG-WRITE**.

J Welsh Data Collection

One of the contributions of UNIVERSALCEFR is the release of the first-ever open dataset for the Welsh language (CY) with gold-standard CEFR labels for A1 and A2. To obtain this data, we corresponded with data maintainers from Learn Welsh (`https://learnwelsh.cymru/`), which is a compilation of expert-created books (reference texts) and acquired PDF versions. This resource can be shared in any format for non-commercial research, which fits the goal of UNIVERSALCEFR. We then manually extracted qualified texts according to the four levels of granularity: sentence, paragraph, dialogue, and document. The distribution of CEFR levels and text granularity for this new Welsh dataset can be found in Table 3 and 7, respectively.