CEFR-Cymraeg: A Dataset and Baseline Models for Language Proficiency Assessment in Welsh

Eeshan Waqar[1], **Jonathan Davies**[2], **Dawn Knight**[3], **Fernando Alva-Manchego**[4]
[1]School of Mathematics, Cardiff University, UK
[2]School of Welsh, Cardiff University, UK
[3]School of English, Communication and Philosophy, Cardiff University, UK
[4]School of Computer Science and Informatics, Cardiff University, UK
eeshanwaqar@hotmail.com, {DaviesJW9, KnightD5, AlvaManchegoF}@cardiff.ac.uk

Abstract

We introduce CEFR-Cymraeg, the first dataset annotated with Common European Framework of Reference (CEFR) levels for Welsh. The dataset is built from learning materials for adult learners, carefully extracted from widely used coursebooks and verified by teachers of Welsh as a second language. It spans levels A1 to B2 and includes multiple units of analysis: sentences, dialogues, paragraphs, and documents. In total, 2,658 entries are provided with gold-standard CEFR annotations, making CEFR-Cymraeg a valuable resource for research on language learning and low-resourced Celtic languages. To illustrate its potential applications, we define language proficiency assessment as a multi-class classification task and fine-tune multilingual pre-trained language models. Given the limited size of the dataset, we also experiment with data augmentation. Results show that these models successfully capture proficiency distinctions and generalise well to Welsh, with the best-performing model reaching a weighted F1-score of 0.83. Qualitative analysis confirmed that most apparent errors reflected valid pedagogical variation rather than model inconsistencies. CEFR-Cymraeg establishes a benchmark resource for Welsh and opens new opportunities for educational NLP, corpus linguistics, and multilingual proficiency research.

Keywords: Welsh, language proficiency assessment, low-resource language

1. Introduction

Language proficiency assessment is central to language education and applied linguistics, providing a framework for describing and evaluating learner competence. The Common European Framework of Reference for Languages (CEFR) provides a standardised scale of six proficiency levels (A1–C2) that is widely used in language learning and proficiency assessment worldwide.[1] Within NLP, CEFR-aligned datasets have enabled a range of research on automatic proficiency classification, educational assessment and calibration of reading materials (Xia et al., 2016; Harsch, 2014; Figueras, 2012).

However, such advances have largely benefited high-resource languages such as English, German, and French, while smaller European languages remain underrepresented. The recent UniversalCEFR initiative (Imperial et al., 2025) has sought to standardise CEFR annotations across languages and datasets through unified formatting and metadata guidelines, offering a foundation for multilingual proficiency modeling. Yet, minority and regional languages such as Welsh remain critically under-resourced, appearing only marginally in existing CEFR-aligned corpora.

Developing a predictive tool capable of estimating the CEFR level of a Welsh text has broad implications for education, corpus linguistics, and national language policy. In particular, it aligns with ongoing legislative and strategic efforts to promote the use of Welsh in education and digital domains, such as the Welsh Language Technology Action Plan, and the long-term Cymraeg 2050 strategy.[2] Within the education sector, automatic CEFR prediction can assist teachers and curriculum designers in selecting or developing reading materials appropriate to learners' proficiency levels, supporting adaptive curricula and targeted assessment. This is especially relevant to the Welsh Language and Education (Wales) Act 2025,[3] which emphasises the creation of shared proficiency benchmarks across schools to promote consistent learner progression.

To address the absence of CEFR-annotated resources for Welsh, we introduce **CEFR-Cymraeg**, the first dataset annotated with CEFR levels for Welsh. The corpus comprises texts extracted from adult learner coursebooks spanning levels A1–B2, verified by a qualified teacher of Welsh as a second language, and formatted following the UniversalCEFR guidelines to ensure cross-lingual interoperability. The dataset includes multiple granularities (sentence, dialogue, paragraph, and doc-

[1] https://www.coe.int/en/web/common-european-framework-reference-languages

[2] https://www.gov.wales/sites/default/files/publications/2018-12/cymraeg-2050-welsh-language-strategy.pdf

[3] https://www.legislation.gov.uk/asc/2025/2

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3496–3505
11–16 May 2026.

ument) to enable fine-grained linguistic analysis and model training. We further present baseline experiments on **automatic CEFR level prediction** using fine-tuned multilingual and UK-focused transformer models, complemented by data augmentation through high-similarity back-translation. Quantitative and qualitative analyses demonstrate that large-capacity models such as BritLLM effectively capture proficiency distinctions even under low-resource conditions, while error inspection highlights inconsistencies in source-level labelling and the value of human validation.

Our contributions are threefold:[4]

1. A new open dataset, CEFR-Cymraeg, establishing the first CEFR-aligned benchmark for Welsh;
2. Baseline models and experiments for automatic proficiency classification; and
3. Empirical insights into linguistic progression encoded in CEFR-aligned Welsh materials.

2. Related Work

2.1. CEFR-annotated Datasets

The Common European Framework of Reference for Languages (CEFR) provides a standardised scale for describing language proficiency across six levels (A1-C2).Several CEFR-aligned corpora have been introduced for major languages, supporting research on language learning and language proficiency. Examples include those in repositories such as CEFRLex,[5] Corpora@UCLouvain,[6] CLARIN L2 Learner Corpora,[7] Språkbanken,[8] among others. More recently, UniversalCEFR (Imperial et al., 2025) introduced the first large-scale open multilingual collection of CEFR-labeled texts, covering 13 languages and providing guidelines for data representation and interoperability.

Despite this progress, minority and regional languages remain underrepresented. Welsh, in particular, appears only in small quantities within the UniversalCEFR collection, highlighting the scarcity of open CEFR-annotated resources for Celtic languages. The CEFR-Cymraeg dataset fills this gap by introducing the first CEFR-aligned corpus for Welsh, following the UniversalCEFR formatting guidelines to ensure cross-lingual compatibility and reproducibility.

2.2. Applications of CEFR

Automatic CEFR prediction has a range of applications spanning education, language technology, and public communication. Within education, it can support teachers and curriculum designers in selecting or developing reading materials appropriate to learners' proficiency levels, facilitating adaptive curricula and assessment. This functionality is particularly relevant in the context of the Welsh Language and Education (Wales) Act 2025, which emphasises the establishment of shared proficiency benchmarks in schools to promote consistent learner progression.

For adult learners, such as those enrolled in CEFR-aligned courses provided by the National Centre for Learning Welsh, automated proficiency prediction can enhance learner autonomy. By assessing texts encountered outside formal coursework, such as online articles or community materials, learners can identify resources suited to their current level, reinforcing motivation and self-directed learning (Krashen, 1982; Dörnyei, 1997).At more advanced stages, this functionality assists tutors in sourcing authentic texts that appropriately challenge learners within their proficiency band.

Beyond pedagogy, automatic CEFR annotation of large Welsh corpora enables research on language complexity, cross-level comparison, and typological alignment. Researchers can trace how lexical density, syntactic embedding, discourse markers, and cohesion devices evolve across CEFR strata, contributing to corpus-based models of readability and proficiency progression. Such annotated resources also facilitate cross-linguistic analyses of language learning trajectories and text simplification strategies (Vajjala and Rama, 2018a; Imperial et al., 2025).

Public bodies, publishers, and media organisations would likewise benefit from a CEFR predictor when producing accessible Welsh-language content. Initiatives such as *Cymraeg Clîr* highlight the persistent use of formal registers that hinder comprehension among segments of the Welsh-speaking population. A CEFR-aware text analysis tool could guide content creators toward clearer, more inclusive communication. In line with the Welsh Language (Wales) Measure 2011,[9] which mandates that Welsh be treated "no less favourably" than English, such tools provide practical mechanisms for ensuring linguistic accessibility and parity in public services.

[4] Our dataset and models available at https://huggingface.co/collections/cardiffnlp/welsh-cefr

[5] https://cental.uclouvain.be/cefrlex/

[6] https://corpora.uclouvain.be/catalog/

[7] https://www.clarin.eu/resource-families/L2-corpora

[8] https://spraakbanken.gu.se/en/resources/learner-language

[9] https://www.legislation.gov.uk/mwa/2011/1/contents

2.3. Minority and under-resourced languages

A minority language is often the indigenous language of a particular country or state, but is *not* used by the majority of speakers there. While Welsh is an official language of Wales, it is considered a minority language as according to the 2021 Census, only 17.7% of the population of Wales reported being able to speak Welsh (approximately 538,000 individuals; see ONS, 2022).

Minority languages like Welsh are often under-resourced, either financially, pedagogically, culturally or technologically. The development of digital linguistic tools plays a particularly important role in closing the technological resource gap in under-resourced languages (Scannell, 2007). Efforts to reduce the resource gap have accelerated since the early 2000s, particularly with the increasing availability of natural language processing (NLP) tools and datasets including linguistic corpora (Singh et al., 2024)). Despite a range of initiatives aimed at improving NLP for under-resourced languages, a significant gap remains between the resources available for these languages and those for widely spoken, major languages.

3. The CEFR-Cymraeg Dataset

3.1. Data Sources

The CEFR-Cymraeg dataset was created from a selection of adult learner coursebooks that form part of the "Learn Welsh" series, produced by the National Centre for Learning Welsh. These coursebooks are used across Wales in formal Welsh-as-a-second-language education. They were chosen because they are explicitly organised by CEFR proficiency levels, providing a reliable pedagogical alignment between course content and language proficiency descriptors. All materials are publicly available,[10] and formal permission was obtained to extract text segments from the coursebooks, reformat them, and release them for research use. The coursebooks were available as formatted PDF documents designed for classroom use rather than corpus extraction. As such, the materials required structural segmentation and reformatting prior to inclusion in a machine-readable dataset.

The current version of CEFR-Cymraeg spans four CEFR levels (A1–B2), corresponding to the range covered by the available coursebooks. Higher levels (C1–C2) are not yet included, as suitable teaching materials are not currently available in digital form. Furthermore, two regional variants of the coursebooks exist, one for South Wales and one for North Wales. At this stage, only the South Wales variant has been included. Incorporating the North Wales version in future releases will enable comparative analyses of regional lexical and grammatical variation within Welsh teaching materials.

The "Learn Welsh" programme is currently undertaking a series of planned revisions across its coursebooks. Consequently, most data in CEFR-Cymraeg are drawn from the most recent versions available at the time of collection. Specifically, the A1 (Mynediad), A2 (Sylfaen), and B2(1) (Uwch 1) data were extracted from version 2 of the coursebooks. The B1 (Canolradd) and B2(3) (Uwch 3) levels were sourced from version 1, as their updated editions had not yet been released. The version 2 of the B2(2) (Uwch 2) coursebook was published shortly after data collection; therefore, version 1 was used in this dataset.

3.2. Data Curation

We followed the guidelines of UniversalCEFR (Imperial et al., 2025), ensuring consistency and interoperability with other CEFR-aligned datasets. These guidelines informed both the data extraction and data structuring, including file organisation, metadata fields, and text granularity.

3.2.1. Data Extraction

We combined manual and semi-automatic approaches depending on the amount of suitable data in the coursebooks.

Manual. For the Entry, Foundation, Intermediate, and Advanced 1 coursebooks, extraction was conducted manually by native speakers of Welsh hired specifically for this task. Annotators received detailed instructions on what data to extract and the required format, and regular meetings were held to address questions and resolve issues. The annotators primarily extracted texts from the *Sgwrs* (Dialogue) and *Darllen* (Reading) sections, but were also instructed to include material from other sections that fit one of the four predefined granularities: sentence, dialogue, paragraph, or document. When uncertainties arose, the research team collectively reviewed the cases and agreed on a consistent decision. A sample of manually extracted data was later checked by an experienced teacher of Welsh as a second language, who regularly uses these coursebooks in the classroom.

Semi-Automatic. For the Advanced 2 and Advanced 3 coursebooks, extraction was conducted semi-automatically by the same teacher responsible for data verification. The original materials were available as formatted PDF documents, and automation assisted with structural segmentation

[10] https://learnwelsh.cymru/learning/resource-library/?k=Coursebooks&opt=Tags

and JSON reformatting rather than annotation. The teacher provided ChatGPT with examples of manually extracted JSON entries from the Entry to Intermediate coursebooks and used these as reference to process new materials from Advanced 2. Through several iterative cycles of correction and refinement, ChatGPT was asked to generalise the process and generate a suitable extraction prompt, which was then applied to the remaining units of Advanced 2 and Advanced 3. All resulting outputs were manually reviewed and post-edited to ensure accuracy, completeness, and conformity with the dataset's formatting conventions. Although ChatGPT was occasionally asked to self-assess its extractions, it proved unreliable in identifying minor mistakes, so final quality control was performed manually. Importantly, CEFR labels were not generated by the LLM; all CEFR levels were inherited directly from the source coursebooks.[11]

3.2.2. Data Formatting

Texts are stored in accordance with the UniversalCEFR JSON-based structure. Each entry in the dataset contains the following fields:

- `title`: The unique title of the text retrieved from the corresponding coursework section.
- `lang`: The ISO 639-1 language code of the text (i.e., *cy* for Welsh).
- `source_name`: The name of the coursebook the text was extracted from.
- `format`: The level of textual granularity, selected from the UniversalCEFR-recognised values: document-level, paragraph-level, discourse-level, or sentence-level.
- `category`: The classification of the text based on authorship. All CEFR-Cymraeg entries are labelled as *reference*, since the materials were created by expert teachers and language professionals.
- `cefr_level`: The CEFR level assigned to the text, which is inherited from the source coursebook (A1-B2).
- `license`: The license governing text distribution. CEFR-Cymraeg is released under a *public* license.
- `text`: The text content itself, stored as UTF-8 encoded plain text.

The dataset will be released along with the semi-automatic extraction prompt used during dataset construction, to facilitate reproducibility and reuse.

[11]The full prompt used for semi-automatic extraction is included in the Appendix.

CEFR Level	Total	Sent.	Dial.	Para.	Doc.
A1	764	659	79	18	8
A2	608	448	36	91	33
B1	323	273	33	15	2
B2	963	758	47	124	34
Total	2,658	2,138	195	248	77

Table 1: Number of entries in the CEFR-Cymraeg dataset by CEFR level and four text granularities: sentence, dialogue, paragraph and document.

3.3. Dataset Analysis

Table 1 summarises the composition of the CEFR-Cymraeg dataset. The collection includes 2,658 entries spanning four CEFR levels (A1–B2) and four text granularities (sentence, dialogue, paragraph, document). Sentences dominate the dataset (2,138), followed by paragraphs (248) and dialogues (195), reflecting the structure of the source materials, which emphasise short exchanges and discrete examples at lower levels. B2 contains the largest number of entries overall (963), with a notably higher proportion of paragraphs and documents, suggesting a gradual shift toward extended, context-rich texts at higher proficiency levels.

Table 2 presents mean linguistic statistics per CEFR level, calculated using Proffiliadur (Gutiérrez-Rolón et al., 2026), a text profiling toolkit for Welsh. The values illustrate the expected increase in structural and lexical complexity. Text length and sentence length both rise steadily from A1 to B2, accompanied by higher averages in token count, number of verbs, and syntactic tree depth. The number of clauses per text and the frequency of mutations also grow substantially at higher levels, reflecting the introduction of multi-clause structures, subordinate constructions, and more natural morphosyntactic variation. A slight irregularity at B1 can be attributed to granularity bias, since this level is dominated by sentence-level entries with almost no document-level data. These shorter units contain fewer tokens and simpler syntactic structures, which lowers mean values relative to neighbouring levels. Overall, the trends align with CEFR descriptors of linguistic range and complexity, confirming that CEFR-Cymraeg captures a clear progression in grammatical and lexical sophistication across proficiency levels.

4. CEFR Level Prediction

This section defines CEFR-level prediction as a multiclass classification task over four proficiency bands (A1–B2). The goal is to evaluate how well multilingual and UK-focused language models can capture Welsh proficiency distinctions using the

CEFR	Sent.	Tokens	Sent. Len.	Verbs	Synt. Tree Depth	Mutations	Clauses
A1	2.62	20.95	6.84	3.65	2.55	1.53	1.18
A2	3.20	34.20	8.74	5.90	3.32	3.22	2.41
B1	2.83	29.06	8.68	5.42	3.22	2.56	2.08
B2	3.47	55.48	11.83	9.24	4.09	6.11	3.84

Table 2: Mean linguistic statistics per CEFR level in the CEFR-Cymraeg dataset, including average number of sentences (Sent.), tokens (Tokens), sentence length (Sent. Len.), number of verbs (Verbs), syntactic tree depth, number of mutations (Mutations), and number of clauses (Clauses). All values represent per-text means within each CEFR level.

CEFR-Cymraeg dataset.

CEFR-Cymraeg are inherited from discrete coursebook classifications rather than continuous proficiency scores. Treating the task as multi-class classification therefore aligns with the categorical nature of the annotation process. Finally, modelling the task as multi-class maintains comparability with prior CEFR prediction studies (Vajjala and Rama, 2018b) (Vajjala and Lucic, 2018) and provides a stable baseline for future work exploring hybrid approaches that incorporate ordinal-aware regularisation while preserving categorical supervision.

4.1. Fine-Tuned Models

We experimented with two transformer-based architectures: **(i) EuroBERT-210M** (Boizard et al., 2025),[12] a multilingual encoder model trained on a diverse set of European languages, and **(ii) BritLLM-3B**,[13] a 3-billion-parameter causal large language model that supports English, Welsh, and other UK languages. Both models were fine-tuned using stratified K-fold cross-validation, preserving proportional label distributions across CEFR levels. This setup facilitated a comparison between a parameter-efficient multilingual encoder (EuroBERT) and a large-capacity decoder-style model (BritLLM).[14]

4.2. Data Augmentation

To mitigate class imbalance and expand lexical diversity, we experimented with **back-translation** augmentation to generate semantically equivalent paraphrases of existing texts. This pipeline employed (Junczys-Dowmunt et al., 2018) from the Helsinki-NLP[15] family to sequentially utilise both Welsh → English, and English → Welsh, to produce high-fidelity back-translated Welsh sentences.[16] to produce high-fidelity back-translated Welsh sentences.[17]

To ensure semantic fidelity, each original/back-translated pair was evaluated using three complementary metrics: (i) **BLEU** (Papineni et al., 2002) for lexical overlap, (ii) **chrF** (Popović, 2015) for character-level correspondence, and (iii) **cosine similarity** between Sentence-BERT (Reimers and Gurevych, 2019) embeddings for semantic preservation. Only back-translated samples exceeding a cosine similarity threshold of 0.8 were retained for augmentation. This filtering strategy prioritised semantic equivalence and grammatical integrity over maximal data expansion. All augmented instances were recorded with their original Welsh source, intermediate English translation, back-translated Welsh text, CEFR label, and the corresponding BLEU, chrF, and cosine scores. High-similarity samples were compiled into a consolidated dataset and merged with the original CEFR corpus for downstream fine-tuning.

4.3. Evaluation Metrics

We performed model evaluation under a stratified K-fold cross-validation scheme, with performance tracked both at the overall level and separately for each CEFR label (A1-B2). For each fold, we recorded weighted precision, recall, and F1 scores to account for mild class imbalance, along with per-class metrics for each CEFR level. After all the folds were completed, per-fold predictions were concatenated to compute aggregate performance statistics across the entire dataset.

To provide a more interpretable view of classification behaviour, an **overall confusion matrix** was constructed by combining predictions from all folds. This matrix captured absolute prediction counts distributions, highlighting the tendency of models to confuse adjacent proficiency levels (e.g. A2–B1 or B1–B2).

[12] https://huggingface.co/EuroBERT/EuroBERT-210m

[13] https://huggingface.co/britllm/britllm-3b-v0.1

[14] Details on the training process and hyper-parameters for each model are included in the Appendix.

[15] https://huggingface.co/models?other=marian

[16] Full model configuration and translation details are included in the Appendix.

[17] Full model configuration and translation details are included in the Appendix.

5. Results

5.1. Model Comparison

To establish a baseline for Welsh CEFR prediction, we first compared the two fine-tuned models, **EuroBERT-210M** and **BritLLM-3B**, using the Welsh-only dataset without data augmentation. Table 3 summarises the results across all CEFR levels (A1–B2), reporting weighted Precision (P), Recall (R), and F1 averaged across folds.

Metric	EuroBERT	BritLLM
All CEFR Levels		
Precision (P)	0.784	**0.830**
Recall (R)	0.782	**0.830**
F1 Score	0.781	**0.829**
A1		
Precision (P)	0.835	**0.869**
Recall (R)	0.871	**0.905**
F1 Score	0.852	**0.886**
A2		
Precision (P)	0.713	**0.819**
Recall (R)	0.739	**0.788**
F1 Score	0.725	**0.803**
B1		
Precision (P)	0.713	**0.794**
Recall (R)	0.667	**0.711**
F1 Score	0.684	**0.749**
B2		
Precision (P)	0.812	**0.818**
Recall (R)	0.777	**0.836**
F1 Score	0.793	**0.826**

Table 3: Comparison of EuroBERT-210M and BritLLM-3B without data augmentation. Scores show weighted Precision (P), Recall (R), and F1 averaged across folds. Best results per metric are shown in bold.

Overall performance. BritLLM continued to achieve the strongest overall performance following retraining, with an average weighted F1 score of approximately 0.83 across folds. This represents an improvement of around five percentage points over EuroBERT, which attained an average weighted F1 of approximately 0.78. The performance gap remains consistent with expectations given BritLLM's substantially larger parameter count (3B) and its pre-training on UK-focused corpora that include Welsh-language data. Nevertheless, EuroBERT, despite its smaller size (210M parameters), demonstrates competitive performance and stable generalisation, highlighting the efficiency of multilingual encoder-based architectures for CEFR classification.

Per-level analysis. At the individual CEFR level, both models broadly preserve ordinal structure, with many misclassifications occurring between adjacent proficiency bands (e.g., A2–B1, B1–B2). However, some cross-band confusion is also observed, particularly between A2 and B2, which may reflect class imbalance, granularity effects, and structural overlap between sentence-level segments drawn from different course levels. EuroBERT exhibits comparatively strong performance at the lower proficiency levels, particularly at A1, where it achieves high recall and balanced precision–recall trade-offs. However, its performance degrades at intermediate levels, most notably A2, where reduced recall indicates difficulty distinguishing lower-intermediate from intermediate texts. BritLLM demonstrates more balanced behaviour across all CEFR levels, achieving consistently higher recall at A2, B1, and B2, with especially robust performance at the upper bands (B1–B2). This suggests that BritLLM more effectively captures syntactic, lexical, and discourse-level complexity associated with higher proficiency texts.

Confusion matrix analysis. Figure 1 presents the overall confusion matrices for BritLLM and EuroBERT.

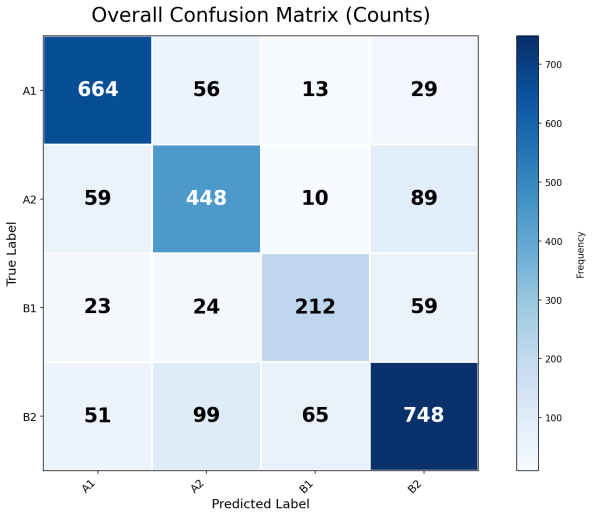
EuroBERT exhibits a predominantly diagonal structure, with correct classifications at A1 = 664, A2 = 448, B1 = 212, and B2 = 748. However, it shows increased cross-level confusion compared to BritLLM, particularly at the intermediate and upper levels, including A2→B2 (89) and B2→A2 (99). These patterns suggest greater difficulty in distinguishing between mid-level and high-level texts, consistent with EuroBERT's lower recall at A2 and B1 observed in the quantitative results.

BritLLM demonstrates strong diagonal dominance, correctly classifying the majority of instances at each CEFR level (A1 = 690, A2 = 478, B1 = 226, B2 = 805). Misclassifications are most notably A2→B2 (86) and B1→B2 (59), indicating residual ambiguity at higher proficiency boundaries while maintaining clear separation overall.

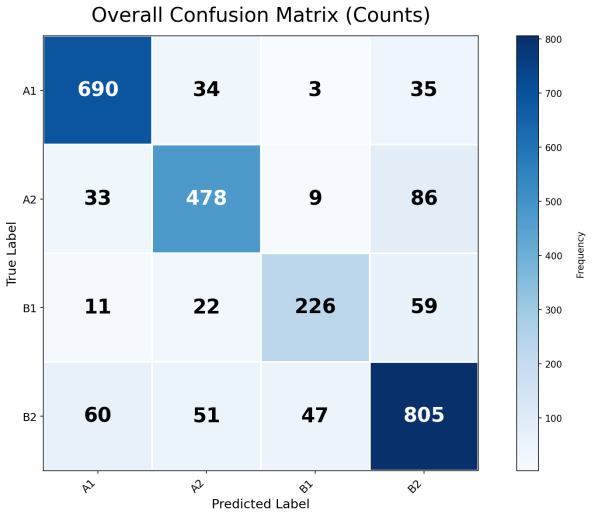
To sum up, both models exhibit strong ordinal coherence, with most errors confined to adjacent levels. BritLLM achieved superior overall accuracy and clearer separation across CEFR bands, while EuroBERT offers a more parameter-efficient alternative that maintains competitive performance on a low-resource language.

5.2. Impact of Data Augmentation

To assess the effect of data augmentation on model performance, we compared each model's results on the original Welsh-only dataset with those obtained on the augmented corpus. The augmented dataset expanded all CEFR levels through high-similarity back-translation, resulting in larger and more balanced class distributions. Specifically, the



(a) EuroBERT (rows: true label; columns: predicted)



(b) BritLLM (rows: true label; columns: predicted)

Figure 1: Overall confusion matrices for EuroBERT and BritLLM (common scale).

number of instances increased from A1 = 762 to 964, A2 = 606 to 719, B1 = 318 to 387, and B2 = 963 to 1185 sentences. This corresponds to an approximate 19–27% increase in available training data across proficiency levels, with the most pronounced gains observed for A1 and B2. The additional data introduced greater lexical and syntactic diversity, particularly benefiting the lower and intermediate CEFR bands, which are more sensitive to data sparsity.

EuroBERT. The effect of data augmentation on EuroBERT was mixed and generally negative at the aggregate level. The overall weighted F1 score decreased from 0.781 on the original dataset to 0.701 following augmentation, indicating reduced generalisation performance when trained on the expanded corpus. At the level of individual CEFR categories, A1 performance declined slightly (F1: 0.825 to 0.803), while A2 experienced a more pronounced reduction (F1: 0.725 to 0.599), suggesting increased confusion at the lower–intermediate boundary. The most substantial degradation was observed at B1, where F1 dropped from 0.684 to 0.443, reflecting heightened overlap between intermediate and upper-level texts. B2 performance also decreased moderately (F1: 0.793 to 0.765). Overall, these results suggest that EuroBERT is sensitive to the inclusion of synthetic data, with augmentation introducing stylistic noise that disproportionately affects discrimination at intermediate and higher CEFR levels.

BritLLM. exhibited consistent but moderate improvements under data augmentation than EuroBERT. The overall weighted F1 score increased from 0.829 on the original dataset to 0.842 following augmentation, corresponding to an absolute improvement of approximately 1.3 percentage points. At the level of individual CEFR categories, the largest gains were observed at B2 (F1: 0.826 to 0.853) and B1 (0.749 to 0.763), indicating improved discrimination at higher proficiency levels. A1 performance showed a modest increase (0.886 to 0.890), while A2 performance remained largely stable (0.803 to 0.803). Recall patterns remained broadly consistent across augmentation, with small improvements at A2 and B2 and minor fluctuations at A1 and B1, suggesting that the additional synthetic data did not induce systematic overfitting. Overall, these results indicate that BritLLM's larger representational capacity enables it to more effectively exploit the increased lexical and structural diversity introduced through back-translation, particularly at the upper CEFR bands where discourse-level complexity is more prominent.

6. Qualitative Error Analysis

We manually examined 250 instances of disagreement between the prescribed CEFR level of the source material and the CEFR level prediced by BritLLM. The misclassified items were reviewed by the same Welsh teacher who contributed to data extraction. Each case was re-evaluated against the teacher's expert judgement of the text's actual CEFR level in Welsh.

Despite a large number of perceived misclassifications, many of the 'true-labels' did not accurately reflect the CEFR level of the segment which has been assessed. As discussed above, the sample material was taken from a range of Learn Welsh coursebooks that are CEFR aligned. However, although many of the sentences within the B1 coursebook were at a B1 level, that is not to say that every sentence within the sample text was also at that level. Many of the classification errors accurately predicted the CEFR level of the segmented text, even when it does not align with the sample mate-

Metric	EuroBERT	BritLLM
All CEFR Levels		
Precision (P) — DA	0.704	**0.843**
Recall (R) — DA	0.714	**0.843**
F1 Score — DA	0.701	**0.842**
A1		
Precision (P) — DA	0.757	**0.879**
Recall (R) — DA	0.860	**0.902**
F1 Score — DA	0.803	**0.890**
A2		
Precision (P) — DA	0.613	**0.811**
Recall (R) — DA	0.595	**0.796**
F1 Score — DA	0.599	**0.803**
B1		
Precision (P) — DA	0.585	**0.832**
Recall (R) — DA	0.388	**0.706**
F1 Score — DA	0.443	**0.763**
B2		
Precision (P) — DA	0.754	**0.837**
Recall (R) — DA	0.783	**0.869**
F1 Score — DA	0.765	**0.853**

Table 4: Impact of data augmentation (DA) on EuroBERT-210M and BritLLM-3B. Scores report weighted Precision (P), Recall (R), and F1 averaged across folds. Best results per metric are shown in bold.

rials' prescribed CEFR level. For example, a highlighted error is the sentence *'dyn ni wedi bod yn siarad amdanoch chi'*; the original material was at a B1 level, however, the model accurately classified this sentence at the A1 level as the language in this sentence would be immediately intelligible to someone studying at that level. Similarly, the sentence *'dw i'n nabod rhywun oedd yn chwarae pel droed'* was originally classified as B2, but this structure is taught in the A1 course and is revisited in unit 2 of the A2 course.

Furthermore, a number of perceived errors were fragments of longer texts. In this sense, where the CEFR level is lower than the 'true-level', this would accurately reflect the fragment of text or word which the model has analysed; despite the longer text being of a higher CEFR level. The initial error analysis highlighted a number of lone words which the model had misclassified; however, beyond prescribed word lists of high frequency words, ascribing a specific CEFR level to a given word presents a number of challenges, especially when considering a learner's individual circumstances, which may affect when they first learn a given word. To support this, the model classified *'tafodiaeth'* (dialect) at level B1; however, *'tafodiaeth'* is a word that could be understood by a learner of Welsh at an earlier level, especially where they are learning in areas where dialectic choices are relevant.

There was a small number of instances where the model had classified a text at a higher level than the 'true-level'. In some instances, this CEFR level was accurate to the predicted level where the course material had been supplemented with native materials such as poetry, which includes a language pattern that is much higher than the 'true-level' of the course book. One example is where the term *'buoch chi'* is used in the B1 coursebook within a larger text. However, this language pattern is not taught explicitly until B2, and is a largely literary form that would only be used at at least high B2.

In addition, a small number of 'true-labels' were incorrectly labelled. In some of these instances, the model accurately predicted the correct CEFR level of the text despite the 'true-level' being inaccurate. Sgwrs 2 from unit 2 of the A1 course (page 21) was originally mislabelled as B1, and the model accurately predicted the CEFR level as A1. However, there were a number of instances where the 'true-label' and the 'predicted-label' were incorrect. For example, 'sgwrs 1' from the A1 course book (page 9) was mislabelled as A2 in the 'true-level' and as B2 by the model.

Quantitatively, comparison against the verified CEFR levels established by the reviewer confirmed that **26.9%** of the model predictions were correct. Crucially, **68.3%** of the true labels themselves were misaligned with these verified levels, providing clear evidence that a substantial portion of the dataset contained labeling inconsistencies. This supports the qualitative observation that many of BritLLM's so-called "errors" were in fact reflective of more accurate assessments of linguistic difficulty. Further analysis of the *direction of error* revealed that **51.8%** of the model's misclassifications were **underestimations**, assigning a lower CEFR level than prescribed in the course material, while **47.8%** represented overestimations. In terms of agreement metrics, the model's accuracy against the verified CEFR levels **(27.6%)** exceeded its accuracy against the original true labels **(0.4%)**, highlighting improved alignment once human verification was introduced. Finally, an examination of the textual characteristics revealed that **13.3%** of the sample consisted of short fragments ($\leq$4 words), with all such items (**100%**) being misclassified. Similarly, both single-word entries (0.8%) were misclassified.

7. Conclusion and Future Work

We introduced **CEFR-Cymraeg**, the first CEFR-aligned dataset for Welsh, alongside baseline models for automatic language proficiency classification. The results show that multilingual and regional transformer models can successfully distinguish between CEFR levels despite limited data availability. Data augmentation through high-similarity

back-translation improved class balance and lexical diversity, with the larger BritLLM model achieving the most consistent performance gains. These findings confirm that model capacity and exposure to domain-relevant text play key roles in adapting large language models to low-resource educational contexts.

Future research can extend this work along several complementary directions. First, expanding **CEFR-Cymraeg** to include higher proficiency levels (C1–C2) and the North Wales regional variant would allow a fuller evaluation of model generalisability across dialectal and stylistic variation. In the future, we could explore **cross-lingual transfer learning** by leveraging typologically related languages and high-resource corpora (e.g., English and other Celtic languages). Furthermore, beyond back-translation, data augmentation could be extended through **synthetic text generation** guided by CEFR descriptors or **contrastive paraphrasing** methods to capture inter-level linguistic boundaries better. Finally, incorporating **human-in-the-loop validation** of model predictions would enable refinement of annotation quality and strengthen the reliability of Welsh CEFR benchmarks in future iterations of this dataset.

Ethics Statement

All materials were extracted from publicly available educational resources under explicit permission for academic use. The dataset represents pedagogical content, not learner writing, thus minimising privacy concerns. The dataset will be released for research use under a CC BY-NC 4.0 license, accompanied by metadata and documentation.

8. Bibliographical References

Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte Miguel Alves, Andre Martins, Ayoub Hammal, Caio Corro, CELINE HUDELOT, Emmanuel Malherbe, Etienne Malaboeuf, Fanny Jourdan, Gabriel Hautreux, João Alves, Kevin El Haddad, Manuel Faysse, Maxime Peyrard, Nuno M Guerreiro, Patrick Fernandes, Ricardo Rei, and Pierre Colombo. 2025. EuroBERT: Scaling multilingual encoders for european languages. In *Second Conference on Language Modeling*.

Zoltán Dörnyei. 1997. Psychological processes in cooperative language learning: Group dynamics and motivation. *The Modern Language Journal*, 81(4):482–493.

Neus Figueras. 2012. The impact of the CEFR. *ELT journal*, 66(4):477–485.

Nicolás Gutiérrez-Rolón, Jonathan Davies, Tomos Williams, Dawn Knight, and Fernando Alva-Manchego. 2026. Proffiliadur: Welsh language text profiling toolkit. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, Palma, Mallorca. European Language Resources Association.

Claudia Harsch. 2014. General Language Proficiency Revisited: Current and Future Issues. *Language Assessment Quarterly*, 11(2):152–169.

Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Munoz Sanchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R. Jablonkai, Ekaterina Kochmar, Robert Reynolds, Eugenio Ribeiro, Horacio Saggion, Elena Volodina, Sowmya Vajjala, Thomas Francois, Fernando Alva-Manchego, and Harish Tayyar Madabushi. 2025. Universalcefr: Enabling open multilingual research on language proficiency assessment.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. *CoRR*, abs/1804.00344.

Stephen D. Krashen. 1982. Acquiring a second language. *World Englishes*, 1(3):97–101.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Kevin P. Scannell. 2007. The crúbadán project: Corpus building for under-resourced languages.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning.

Sowmya Vajjala and Ivana Lucic. 2018. Onestopenglish corpus: A new corpus for automatic readability assessment and text simplification. pages 297–304.

Sowmya Vajjala and Taraka Rama. 2018a. Experiments with universal CEFR classification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 147–153, New Orleans, Louisiana. Association for Computational Linguistics.

Sowmya Vajjala and Taraka Rama. 2018b. Experiments with universal CEFR classification. *CoRR*, abs/1804.06636.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics.

A. Fine-tuning and Optimization Details

EuroBERT was fine-tuned with a sequence-classification head for three epochs per fold. The model checkpoints were selected based on the highest validation weighted F1 score, mitigating the effects of a mild class imbalance between proficiency levels. Our training arguments used the fused AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$) with a maximum learning rate of 3.6×10^{-5}, a linear learning rate scheduler with a warm-up ratio of 10%, and accumulation of gradients over 16 steps (effective batch size ≈ 32). The batch sizes per device were set to 2 for training and 3 for evaluation. Furthermore, checkpoints were saved at each epoch, and the best performing model was retained for downstream evaluation. The model performance report consisted of weighted precision, recall, and F1 metrics, both overall and per CEFR level (A1–B2).

BritLLM-3B is a decoder-only causal model and was fine-tuned for four epochs per fold using the same learning-rate schedule and optimiser configuration as EuroBERT, supplemented with additional optimisation techniques suited to large-scale architectures. These included `bfloat16` precision, gradient checkpointing, and scaled dot-product attention (SDPA) to reduce GPU memory consumption. Since BritLLM is generative, key–value caching was disabled to conserve memory during training. To prevent overfitting, regularization was applied via weight decay (0.01) and label smoothing (0.1), and model selection was guided by validation loss rather than weighted F1. Evaluation was performed after each fold, and metrics included overall and per-class precision, recall, and F1, along with confusion matrices.

B. Model Configuration and Translation Process

Model configuration. Both translation models were loaded using mixed precision (`bfloat16` on A100 GPUs, otherwise `float16`) with low CPU memory usage. The pipeline was also infused with optional optimisations including Scaled Dot-Product Attention (SDPA), `BetterTransformer` fusion, and `torch.compile()` acceleration. The generation employed deterministic decoding (`do_sample = False`, `num_beams = 1`, `max_new_tokens = 128`) to ensure stable paraphrase quality. Furthermore, the back-translation pipeline operated entirely in inference mode, with `torch.autocast` providing automatic mixed-precision control.

Translation process. Each Welsh text segment was first translated into English and then back-translated into Welsh. We implemented batch processing through the `translate_batch()` function, which tokenised inputs (padding and truncating to 512 tokens) and generated outputs on GPU under auto-cast precision. To handle out-of-memory events, dynamic error handling was incorporated by reducing batch size adaptively, ensuring stable execution.