Cross-Corpus CEFR Classification through Artificial Learners Perplexities

Bernardo Stearns[1] Thomas Gaillat[2]
John P. McCrae[1]
[1] Insight Centre for Data Analytics, Data Science Institute, University of Galway, Ireland
[2] LIDILE / Université de Rennes 2, 35000 Rennes, France
Contact: bernardo.stearns@insight-centre.org

Abstract

The complexity of neural methods for automatic proficiency assessment often sacrifices interpretability and robustness. This paper presents a competitive alternative for CEFR classification using optimized statistical models with a novel perplexity-based feature engineering pipeline. We introduce LLM-derived perplexity features as a proxy for how unexpected a learner's word choices are: native model perplexity measures unexpectedness relative to native language use, while Artificial Learner model perplexity quantifies relative to a specific proficiency level. While recent work favors end-to-end neural architectures, we demonstrate that traditional pipelines enhanced with these interpretable perplexity features can achieve comparable performance on established benchmarks. We evaluate two transfer scenarios: zero-shot (trained on EFCAMDAT, tested on external corpora) and 90-10 split (same features, in-domain classifier training). On KUPA-KEYS, perplexity features achieve RMSE 0.707 (zero-shot) and 0.660 (90-10 split), outperforming fine-tuned BERT and prompt-based LLMs. On CELVA-SP, zero-shot perplexity shows limited generalization (RMSE 1.437 vs. LLM's 1.016), but statistical models close this gap in the 90-10 split (RMSE 0.872). Across all three evaluation datasets, perplexity-based models achieve the best average macro F1 in the 90-10 split (0.446 vs. 0.287 for BERT and 0.175 for prompting), demonstrating that interpretable features paired with domain-adapted classifiers provide the most robust cross-domain representations. We contribute: (1) state-of-the-art KUPA-KEYS results with interpretable models, (2) the first comprehensive CELVA-SP benchmark, and (3) evidence that feature-level transfer outperforms both end-to-end fine-tuning and zero-shot prompting.

Keywords: CEFR classification, artificial learners, perplexity, feature engineering, language proficiency assessment

1. Introduction

Automatic proficiency assessment is fundamental to scaling language education, yet existing approaches struggle to balance three critical requirements: (1) accuracy across diverse learner populations, (2) robustness to domain shifts between educational contexts, and (3) interpretability for pedagogical use. These challenges persist even as the field of natural language processing (NLP) has undergone two major paradigm shifts: first from supervised learning with handcrafted features to pre-training and fine-tuning, and later to prompting-based methods that directly leverage language models capabilities (Liu et al., 2023).

While these advances have revolutionized many NLP tasks, their benefits for educational assessment remain debatable. The very strengths of modern large language models (LLMs), their flexibility and few-shot adaptability, introduce instability in narrow, high-stakes domains like CEFR classification. For instance, GPT-4 exhibits erratic performance (with RMSE fluctuations up to 2.417 across corpora) (Benedetto et al., 2025), highlighting the tension between general-purpose language modeling and the precise, reliable judgments required for educational assessment. Similar limitations also appear in scoring clinical patient notes written by medical students, where specialized models outperform generic LLMs. (Yaneva et al., 2024) showed that a clinically adapted DeBERTa model (F1=0.95) was the best performing model, demonstrating that task-specific tuning drives reliability in high-stakes assessment.

The central challenge lies in balancing the potential of modern NLP with the non-negotiable demands of proficiency assessment: accuracy, robustness, and interpretability. However, a key gap remains: Despite rapid advances in LLMs, few studies systematically compare their performance against optimized statistical models enhanced with state-of-the-art techniques, particularly regarding domain adaptation capabilities.

Traditional approaches, when combined with careful feature engineering, offer inherent advantages in handling corpus variability, perfectly aligning with real-world educational requirements. Yet, recent literature has overlooked them, even as many institutional systems continue to rely on simpler, interpretable models due to their proven reliability.

Our research addresses this gap in systematic comparison of LLM-based and statistical ap-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 828–837
11–16 May 2026.

proaches by introducing two strong statistical models and automatic feature engineering approaches to current CEFR benchmarks. We introduce a novel perplexity-based feature engineering pipeline that repurposes LLMs as sophisticated feature generators rather than black-box classifiers, training native language models and proficiency-specific artificial learner models to derive perplexity features that quantify linguistic expectedness: how unexpected a learner's word choices are relative to expert language use (native model perplexity) and specific proficiency levels (artificial learner model perplexity). This hybrid approach preserves the interpretability of traditional feature-based methods while harnessing the discriminative power of LLM-derived representations. We systematically benchmark our approach against fine-tuned BERT and zero-shot prompting methods on two external datasets (KUPA-KEYS and CELVA-SP), establishing state-of-the-art results on controlled assessments while documenting cross-corpus variance shared with modern prompting approaches.

2. Related Work

2.1. CEFR Classification / Automatic Essay Scoring

CEFR classification is the task of automatically predicting a text's proficiency level according to the Common European Framework of Reference for Languages (CEFR), typically framed as a supervised multiclass classification problem in NLP.

Early approaches relied heavily on feature-based scoring (e.g., lexical complexity indices, syntactic patterns), where researchers used domain knowledge to define and extract numerical features. (Gaillat et al., 2022) demonstrated that microsystem criterial features—linguistically motivated indicators of proficiency—could effectively predict CEFR levels in a machine learning framework, establishing the value of interpretable feature engineering for this task.

The need for automatic extraction of features led to neural approaches that occurred in three distinct phases. Early work employed RNN-based classifiers (e.g., (Kerz et al., 2021)), which introduced end-to-end feature learning but achieved only marginal gains over feature engineering. Subsequent fine-tuned language models (e.g., BERT) marked a paradigm shift by leveraging pre-trained representations, yet studies like (Mayfield and Black, 2020) revealed their surprising limitations: transformer-based fine-tuning yielded minimal improvements (0.05 QWK) over traditional methods across five ASAP datasets, despite 30-100x greater computational costs. Most recently, zero-shot prompting of large language models emerged as a third approach, circumventing fine-tuning altogether by directly querying model knowledge, though this introduced new challenges in prompt sensitivity and proficiency overestimation. (Benedetto et al., 2025) suggest they remain volatile for diagnostic applications, yielding only marginal gains while sacrificing interpretability (e.g., failing to explain why an essay was classified as B1 rather than B2).

More recently, (Uchida and Negishi, 2025) combined lexical metrics with generative AI to assign CEFR-J levels to English learners' writing, further illustrating the trend toward hybrid approaches that integrate traditional linguistic features with modern language model capabilities.

This progression reveals a fundamental tension in CEFR classification: while neural methods have reduced the manual effort in feature engineering, they have done so at the cost of interpretability and diagnostic utility, with only marginal performance improvements. The field thus faces a challenge of developing approaches that can leverage modern language models while maintaining the transparency necessary.

2.2. Feature Engineering

Feature engineering in NLP has evolved from manual linguistic features (e.g., n-grams, POS tags) to automated representation learning (e.g., word2vec, BERT), trading interpretability for scalability. However, these dense embeddings sacrifice the transparency needed for diagnostic applications, prompting recent exploration of LLMs as explicit annotation tools rather than embedding generators. Recent work explores LLMs as hybrid tools—generating structured annotations while preserving granularity. For example, (Sung and Kyle, 2024) shows the ability of LLMs to identify linguistic constructions such as the structure of the argument. Similarly, (Mohta et al., 2023) benchmarks LLMs as annotators, showing their potential to produce explainable labels, though their noise limits direct replacement of human annotations.

The possibility of reframing LLMs not merely as black-box classifiers but as nuanced feature annotators could be particularly valuable for language learning applications, where granular pattern analysis (e.g., verb phrase constructions) matters more than raw prediction.

2.3. Artificial Learners

Artificial learners constitute a research paradigm for computationally modeling second language acquisition and learner language production. This encompasses two complementary approaches: (1) *behavioral models* that simulate observable learner language patterns at different proficiency levels,

capturing systematic variation in linguistic production across the developmental trajectory, and (2) *cognitive models* that attempt to represent the internal linguistic knowledge, processing mechanisms, and acquisition dynamics of learners. While various computational frameworks can instantiate artificial learners, language models—and particularly large language models (LLMs)—have emerged as a convenient and powerful tool for modeling learner language behavior due to their probabilistic nature and ability to capture complex linguistic patterns. Early work focused on analyzing narrow linguistic phenomena (Kim, 2024), while recent approaches demonstrate more general modeling capabilities using LLMs (Stearns et al., 2024). Sequential pretraining methods have also shown promise for modeling L1 transfer effects in L2 acquisition. Aoyama and Schneider (2024) employ the TILT (Test of Inductive bias via Language Transfer) method, pretraining GPT2 models on six different L1s (Arabic, Chinese, English, Japanese, Portuguese, Spanish) before continuing training on English with frozen transformer blocks. Their experiments reveal that L1-L2 typological distance correlates with morphosyntactic performance, but surprisingly, matching the model's L1 to human participants' L1 background has minimal effect on predicting L2 reading times. These systems have been applied across diverse educational contexts, from simulating student responses to exam questions (Benedetto et al., 2024) to modeling generative student agents in interactive learning environments (Xu et al., 2024).

A key strength of artificial learners lies in their probabilistic modeling of learner language. For instance, given a word or phrase used by a learner, an artificial learner can estimate how likely it is to appear at specific CEFR levels (e.g., the probability of an A2 learner producing "very happy" versus a B1 learner). This granular approach bridges the gap between traditional feature engineering (e.g., counting lexical sophistication metrics) and modern neural methods, providing interpretable diagnostics while retaining scalability. Similar probabilistic approaches have been used to estimate word production complexity, where perplexity-based features predict semantic errors in learner writing (Strohmaier and Buttery, 2024).

Beyond proficiency assessment, artificial learners serve as proxy populations for data augmentation and evaluation across sparse CEFR levels (Benedetto et al., 2024; Xu et al., 2024), though concerns remain about validity and cognitive plausibility (Aditya Srivatsa et al., 2025).

3. Methodology

3.1. Datasets

We use three corpora for model development and evaluation. We split the EFCAMDAT dataset (Geertzen et al., 2013; Shatz, 2020) into three parts: one for training artificial learners, one for training statistical models on engineered features, and a hold-out set for in-domain validation. CELVA-SP (Mallart et al., 2023) and KUPA-KEYS (Velentzas et al., 2024) are reserved entirely as external test sets, ensuring unbiased evaluation on unseen learners, prompts, and writing conditions.

3.1.1. Training data

EFCAMDAT Corpus - Serving as our large-scale reference dataset, we use the refined version of the original EFCAMDAT corpus (Geertzen et al., 2013), as introduced by (Shatz, 2020). This refined corpus contains 723,282 learner writings from Englishtown language schools, representing a comprehensive record of past learners' linguistic development. We leverage a split of the dataset to train six artificial learner models and another split to fit the weights of the statistical models.

3.1.2. Testing data

CELVA-SP - The external test set is made of learner writings from the CELVA-SP (Mallart et al., 2023), a corpus of Language for Specific Purposes comprising writings from French undergraduates using English for Specific Purposes (ESP). Learners answered one of three question prompts as part of a 45-minute in-class writing task. All their writings were subsequently annotated with the writing competence scale of the CEFR by four expert raters. Pairwise inter-rater agreement was computed on the basis of 60 writings, yielding Cohen's kappa values ranging from .52 to .72.

KUPA-KEYS - The second external test set leverages the KUPA-KEYS corpus (Velentzas et al., 2024), a multimodal dataset capturing L2 English writing processes through keystroke logging. The corpus comprises 1,006 essays authored by both native and non-native speakers (42 L1s represented), collected via a controlled web interface that recorded granular typing dynamics. Each submission was annotated for proficiency using a 12-point CEFR scale (A1–C2) by three expert raters and an automated scorer.

3.1.3. Experimental Scenarios

We evaluate our approach under two experimental scenarios to assess both zero-shot generalization and in-domain adaptation capabilities across two external test sets (CELVA-SP and KUPA-KEYS)

Distribution of proficiency levels across datasets

Dataset	A1	A2	B1	B2	C1	C2
EFCAMDAT						
Train	37,777 (47.2%)	23,764 (29.7%)	12,932 (16.2%)	4,385 (5.5%)	1,140 (1.4%)	0 (0.0%)
Test	9,438 (47.2%)	5,937 (29.7%)	3,240 (16.2%)	1,112 (5.6%)	275 (1.4%)	0 (0.0%)
Held-out	293,940 (47.2%)	185,643 (29.8%)	100,367 (16.1%)	34,741 (5.6%)	8,591 (1.4%)	0 (0.0%)
External Test Sets						
KUPA-KEYS	0 (0.0%)	0 (0.0%)	109 (10.8%)	570 (56.7%)	312 (31.0%)	15 (1.5%)
CELVA-SP	157 (9.0%)	511 (29.3%)	609 (35.0%)	353 (20.3%)	100 (5.7%)	12 (0.7%)

Table 1: Distribution of proficiency levels across datasets

and one held-out EFCAMDAT subset (EFCAMDAT-test).

Zero-shot scenario - Both feature extraction models (artificial learners) and statistical classifiers are trained exclusively on EFCAMDAT and tested on the complete external corpora without any exposure to target domain data. This scenario evaluates true cross-corpus generalization, where perplexity features are computed using EFCAMDAT-trained artificial learner models, and statistical models (Logistic Regression, XGBoost) are trained on EFCAMDAT-derived features. We evaluate on: (1) the entire CELVA-SP corpus (1,742 essays), (2) the entire KUPA-KEYS corpus (1,006 essays), and (3) a held-out EFCAMDAT test set (EFCAMDAT-test). This setup mirrors real-world deployment where labeled data from the target domain is unavailable.

90-10 split scenario - The same zero-shot perplexity features (computed from EFCAMDAT-trained artificial learners) are used, but statistical models are trained on 90% of each external test set and evaluated on the remaining 10%. We apply this scenario to: (1) CELVA-SP (90% train: 1,567 essays, 10% test: 175 essays) and (2) KUPA-KEYS (90% train: 905 essays, 10% test: 101 essays). This scenario isolates the contribution of in-domain statistical model training while keeping feature representations constant. By comparing these scenarios, we can disentangle the effects of domain-specific feature learning versus domain-specific classifier training, revealing whether performance gains come from better feature representations or better-calibrated decision boundaries. To assess result stability, we repeat each 90-10 split 30 times with stratified random sampling and report mean $\pm$ standard deviation across runs.

3.2. Model Training and Baselines

3.2.1. Artificial Learner Language Models

We pre-train artificial learner language models using the held-out EFCAMDAT collection to generate perplexity-based features for CEFR classification. We pre-train two categories of models: (1) a **General Artificial Learner** pre-trained on the complete held-out EFCAMDAT data spanning all proficiency levels (A1–C1), establishing a baseline understanding of typical developmental trajectories across the full CEFR spectrum, and (2) five **proficiency-specific Artificial Learners**, each pre-trained exclusively on CEFR level-filtered subsets of the same held-out EFCAMDAT collection (A1, A2, B1, B2, C1), as EFCAMDAT contains no C2 texts. This design allows us to compare general cross-level perplexity signals against level-specific linguistic patterns in our feature engineering pipeline.

All artificial learner models are pre-trained in the Next Word Prediction task using the nanoGPT framework[1] with the GPT-2 architecture (124M parameters: 12 layers, 12 heads, 768 embedding dimensions), trained on the held-out EFCAMDAT subset for the Next Word Prediction task with cross-entropy loss using the following hyperparameters: learning rate 6×10^{-4} with linear warmup over 2,000 steps, context window (block size) of 1,024 tokens, batch size 32 on a single 16GB GPU, using the AdamW optimizer . The Native language model refers to the pre-trained GPT-2 model.

3.2.2. Statistical Classification Models

We train two interpretable statistical models using the automatically engineered features: **Logistic Regression** (LR) and **XGBoost** (XGB). Both models are optimized through hyperparameter search using Optuna[2]. For Logistic Regression, we search over the inverse regularization strength C in [0.001, 100] on a logarithmic scale, penalty types including ℓ_2 regularization (L2), ℓ_1 regularization (L1), and elasticnet combinations, the saga solver, maximum iterations in [500, 3000], and class weight balancing strategies (balanced or unweighted). For XGBoost, we optimize the number of boosting rounds in [50, 300], tree depth in [3, 10] levels, learning rate in [0.01, 0.3] on a logarithmic scale, minimum child weight in [1, 10], subsampling and column subsampling rates in [0.6, 1.0], and regularization

[1] nanoGPT: https://github.com/karpathy/nanoGPT. A minimal implementation of GPT-2 in PyTorch.

[2] Optuna: https://github.com/optuna/optuna

parameters (gamma, reg_alpha, and reg_lambda) each in [0.0, 5.0], using the multi-class softmax objective and multiclass logloss evaluation metric. Both models are trained with four distinct feature configurations: (1) **native perplexities** using only native language model perplexities, (2) **native + general AL perplexities** combining native perplexities with the General Artificial Learner perplexities, (3) **all models perplexities** incorporating perplexities from native models, the General AL, and all five proficiency-specific ALs, and (4) **TF-IDF features** using term frequency-inverse document frequency representations extracted from the training corpus.

3.2.3. Baseline Models

We compare our perplexity-based statistical models against three categories of baselines: (1) **fine-tuned BERT**, domain-adapted on EFCAMDAT training data using standard transformer fine-tuning procedures, (2) **zero-shot LLMs** including Mistral 7B, Gemma 2B/7B, and LLaMA-3 8B with two automated essay scoring prompts (AES1 and AES2), identically as used in (Benedetto et al., 2025), with temperature set to 0 for deterministic outputs, following the exact inference parameters reported in (Benedetto et al., 2025), and (3) **test set majority class** (oracle), which predicts the most frequent CEFR level in each test set, serving as a reasonable upper bound baseline for imbalanced datasets.

Zero-shot LLMs produce identical predictions in both scenarios since they perform inference via prompting without any training. In the 90-10 split scenario, we use these same predictions but evaluate them on the same 30 random stratified 10% test splits used for the statistical models, reporting mean $\pm$ standard deviation to enable direct comparison under identical evaluation conditions. For BERT, the zero-shot scenario uses a model fine-tuned on the EFCAMDAT held-out training split, while for the 90-10 split scenario, BERT is fine-tuned on only the 90% training split of each target dataset, resulting in different predictions between the two scenarios.

3.3. Feature Engineering

We focus exclusively on automatic feature engineering strategies that dynamically encode corpus-specific and contextual information without manual linguistic annotation. This approach offers several advantages: (1) scalability across different learner corpora without expert-designed features, (2) adaptability to corpus-specific distributions and lexical patterns, and (3) interpretability through probabilistic representations that capture linguistic expectedness. By avoiding hand-crafted linguistic features, we ensure that our methods can generalize to new assessment contexts while maintaining the transparency necessary for educational applications.

3.3.1. LLM-based Perplexity Features

Our work is inspired by the possibility of reframing LLMs not merely as black-box classifiers but as nuanced feature annotators. We operationalize this approach by using LLM-derived perplexity scores as structured numeric features, where each score reflects a proficiency-tuned model's degree of "surprise" for each LLM subtoken in a target learner text. Concretely, for each learner text we compute the per-position perplexity at each of the first 512 BPE subtoken positions, producing a fixed-length vector of 512 features per language model. Texts shorter than 512 subtokens are zero-padded, while longer texts are truncated. Each feature dimension corresponds to the perplexity value at a specific subtoken position, preserving the sequential structure of model surprisal across the text. With 7 language models (1 native GPT-2, 1 general artificial learner, and 5 proficiency-specific artificial learners), this yields up to $7 \times 512 = 3{,}584$ positional perplexity features per text in the "all models perplexities" configuration. We expect that such perplexity features could implicitly encode fluency and proficiency signals while remaining compatible with traditional statistical analysis. We note that perplexity is computed at the subtoken level, which does not directly align with word-level linguistic constructs used in pedagogy. Future work can automatically align such perplexity scores to human tokens for better interpretability.

3.3.2. TF-IDF Features

As a complementary automatic feature engineering approach, we extract Term Frequency-Inverse Document Frequency (TF-IDF) representations from learner texts. TF-IDF features capture corpus-specific lexical distributions, weighting terms by their discriminative power across proficiency levels. These features provide a traditional baseline for comparison with our perplexity-based representations, while still maintaining the automatic extraction property that enables scalability and cross-corpus adaptation.

3.4. Evaluation Metrics

We evaluate our models across two experimental scenarios—**zero-shot** (models trained on EFCAMDAT, tested on external corpora) and **90-10 split** (cross-domain features with in-domain model training)—using multiple complementary metrics:

Numerical CEFR prediction (Tables 2 and 3): We report (1) **RMSE** (Root Mean Square Error)

measuring prediction accuracy with penalty for large deviations, (2) **Within1** quantifying the proportion of predictions within one CEFR level of the reference, (3) **Spearman** ρ assessing rank-order correlation between predicted and reference scores, and (4) **AC2** (Gwet's agreement coefficient) evaluating inter-rater reliability while accounting for chance agreement, computed using the irrCAC Python library[3].

Multiclass F1 scores (Table 4): We compute macro-averaged F1 scores across CEFR levels on three test datasets (CELVA-SP, KUPA-KEYS, EFCAMDAT-test), with macro-averaging ensuring balanced evaluation across proficiency levels regardless of class distribution.

All metrics are computed on complete external test datasets with no overlap with EFCAMDAT training data, enabling genuine evaluation of cross-corpus generalization.

4. Results

We evaluate our perplexity-based feature engineering pipeline on two external datasets (KUPA-KEYS and CELVA-SP), comparing statistical models (Logistic Regression and XGBoost) against fine-tuned BERT and zero-shot prompting methods (Gemma, LLaMA 3, Mistral).

4.1. KUPA-KEYS

On KUPA-KEYS, our perplexity-based statistical models substantially outperform existing benchmarks. Table 2 shows that **XGBoost with native + general AL perplexities** achieves RMSE 0.707 with 98.4% within-one-level accuracy and AC2 agreement of 0.906, outperforming the feature-based baseline (RMSE 1.570, AC2 0.911) reported by (Benedetto et al., 2025) by 55% and fine-tuned BERT (RMSE 0.892, AC2 0.818) by 21%. **XGBoost with native perplexities** achieves RMSE 0.723 (97.8% within-one-level), while **Logistic Regression with all models perplexities** reaches RMSE 0.744 (97.0% within-one-level).

Perplexity-based features demonstrate robust performance across configurations, with RMSEs ranging from 0.707 to 0.893, all substantially outperforming both BERT and zero-shot prompting baselines. Notably, TF-IDF features also achieve competitive performance (XGBoost: RMSE 0.982; LR: RMSE 0.838), though they underperform perplexity-based representations. The consistent superiority of perplexity features across different statistical classifiers suggests that probabilistic modeling of linguistic expectedness effectively captures proficiency signals in process-logged writing assessments.

[3] https://pypi.org/project/irrCAC/

In contrast, prompt-based LLMs exhibited high variance and poor calibration, with RMSEs ranging from 1.058 (Gemma 7B AES1) to 1.481 (Mistral 7B AES2), highlighting their instability for controlled assessment contexts despite some configurations achieving moderate correlations (e.g., Mistral 7B AES1: $\rho = 0.485$). This 1.5–2.1× performance gap demonstrates that zero-shot prompting fails to leverage proficiency-diagnostic patterns without domain-specific training.

4.2. CELVA-SP

We establish the first comprehensive benchmark on CELVA-SP across multiple modeling approaches, revealing contrasting generalization patterns from KUPA-KEYS. Table 3 shows that **prompt-based LLMs substantially outperform statistical and fine-tuned models** on this context-specific academic writing corpus. **LLaMA 3 8B (AES1)** achieves RMSE 1.016 with 87.7% within-one-level accuracy and AC2 agreement of 0.910, followed by **Mistral 7B (AES1)** (RMSE 1.037, $\rho = 0.564$) and **Gemma 7B** (RMSE 1.220–1.359).

Among statistical models, **TF-IDF features outperform perplexity-based representations**: XGBoost with TF-IDF achieves RMSE 1.200 (79.1% within-one-level, AC2 0.838), while Logistic Regression with TF-IDF reaches RMSE 1.255 (76.1% within-one-level, AC2 0.828). In contrast, perplexity-based features show degraded performance, with best results from **XGBoost with native perplexities** (RMSE 1.442, 67.7% within-one-level, AC2 0.759) and **Logistic Regression with native perplexities** (RMSE 1.437, 68.4% within-one-level, AC2 0.765). Adding artificial learner perplexities generally degrades performance, with **Logistic Regression with native + general AL** reaching RMSE 1.950 (48.3% within-one-level), suggesting poor cross-corpus generalization of AL-derived features to domain-specific contexts.

One possible explanation for the superior performance of TF-IDF on CELVA-SP is the corpus's specialized nature: as an English for Specific Purposes (ESP) collection from French undergraduates, CELVA-SP may contain domain-specific lexical cues (e.g., academic and professional vocabulary) that TF-IDF could directly capture but that general perplexity measures might not differentiate from proficiency-related variation. If this is the case, the artificial learner perplexities, trained on the general-purpose EFCAMDAT corpus, could be capturing stylistic and register differences rather than pure proficiency signals when applied to this academic writing context. This suggests that perplexity-based features are most effective when the training and target domains share similar registers.

BERT achieves RMSE 1.728 (54.4% within-one-level, AC2 0.610), underperforming both TF-IDF

Table 2: Numerical CEFR grading results on KUPA-KEYS across experimental scenarios

		Zero-shot					90-10 Split (×30)				
Model	Features/Prompts	Pred.	RMSE↓	Within1↑	Spearman ρ ↑	AC2↑	Pred.	RMSE↓	Within1↑	Spearman ρ ↑	AC2↑
1. Statistical Models											
Test set Majority Class (oracle)*	–	1006	0.692	0.985	-	0.919	3030	0.707 ± 0.015	0.980	-	0.914 ± 0.008
Logistic Reg.	native perplexities	1006	0.886	0.938	0.152	0.824	3030	0.853 ± 0.024	0.948	0.228 ± 0.034	0.835 ± 0.011
Logistic Reg.	native + general AL perplexities	1006	0.893	0.935	0.136	0.819	3030	0.860 ± 0.025	0.945	0.215 ± 0.035	0.832 ± 0.012
Logistic Reg.	all models perplexities	1006	0.744	0.970	0.418	0.878	3030	0.720 ± 0.018	0.975	0.415 ± 0.028	0.887 ± 0.010
Logistic Reg.	tf-idf	1006	0.838	0.953	0.497	0.914	3030	0.810 ± 0.022	0.965	0.485 ± 0.032	0.915 ± 0.009
XGBoost	native perplexities	1006	0.723	0.978	0.453	0.898	3030	0.704 ± 0.020	0.970	0.217 ± 0.038	0.898 ± 0.011
XGBoost	**native + general AL perplexities**	**1006**	**0.707**	**0.984**	**0.503**	**0.906**	3030	0.704 ± 0.020	0.970	0.286 ± 0.041	0.895 ± 0.010
XGBoost	**all models perplexities**	1006	0.778	0.964	0.403	0.857	**3030**	$\mathbf{0.660 \pm 0.019}$	**0.980**	$\mathbf{0.330 \pm 0.042}$	$\mathbf{0.910 \pm 0.009}$
XGBoost	tf-idf	1006	0.982	0.880	0.319	0.869	3030	0.697 ± 0.021	1.000	0.244 ± 0.039	0.897 ± 0.010
2. Text Classifiers											
BERT	–	1006	0.892	0.931	0.131	0.818	3030	1.583 ± 0.056	0.565	0.318 ± 0.077	-0.471 ± 0.109
3. Prompt Models											
Gemma 2B*	AES1	1006	1.452	0.643	0.191	0.310	3030	1.472 ± 0.055	0.623	-0.017 ± 0.096	0.255 ± 0.038
Gemma 2B*	AES2	1006	1.163	0.798	0.236	0.651	3030	0.813 ± 0.045	0.941	-0.006 ± 0.087	0.726 ± 0.031
Gemma 7B*	AES1*	1006	1.058	0.849	0.216	0.842	3030	0.904 ± 0.076	0.919	0.346 ± 0.143	0.409 ± 0.128
Gemma 7B*	AES2*	1006	1.230	0.770	0.262	0.854	3030	1.147 ± 0.043	0.821	0.271 ± 0.100	-0.174 ± 0.097
LLaMA 3 8B*	AES1*	1006	1.442	0.672	0.464	0.803	3030	1.421 ± 0.049	0.670	0.248 ± 0.110	-0.730 ± 0.126
LLaMA 3 8B*	AES2*	1006	1.322	0.731	0.306	0.827	3030	1.257 ± 0.052	0.757	0.303 ± 0.085	-0.465 ± 0.132
Mistral 7B*	AES1*	1006	1.473	0.659	0.485	0.797	3030	1.433 ± 0.058	0.668	0.270 ± 0.098	-0.675 ± 0.127
Mistral 7B*	AES2*	1006	1.481	0.644	0.367	0.645	3030	1.404 ± 0.042	0.670	0.394 ± 0.068	-0.818 ± 0.122

Table 3: Numerical CEFR grading results on CELVA-SP across experimental scenarios

		Zero-shot					90-10 Split (×30)				
Model	Features/Prompts	Pred.	RMSE↓	Within1↑	Spearman ρ ↑	AC2↑	Pred.	RMSE↓	Within1↑	Spearman ρ ↑	AC2↑
1. Statistical Models											
Test set Majority Class (oracle)*	–	1742	1.072	0.846	-	0.913	5250	1.072 ± 0.035	0.845	-	0.913 ± 0.010
Logistic Reg.	native perplexities	1742	1.437	0.684	0.404	0.765	5250	1.000 ± 0.055	0.897	0.567 ± 0.045	0.874 ± 0.018
Logistic Reg.	native + general AL perplexities	1742	1.950	0.483	0.061	0.525	5250	1.129 ± 0.065	0.840	0.446 ± 0.052	0.837 ± 0.022
Logistic Reg.	all models perplexities	1742	1.551	0.653	0.364	0.693	5250	1.121 ± 0.063	0.829	0.432 ± 0.050	0.840 ± 0.020
Logistic Reg.	**tf-idf**	1742	1.255	0.761	0.490	0.828	**5250**	$\mathbf{0.872 \pm 0.048}$	**0.926**	$\mathbf{0.593 \pm 0.038}$	$\mathbf{0.915 \pm 0.015}$
XGBoost	native perplexities	1742	1.442	0.677	0.430	0.759	5250	0.910 ± 0.052	0.920	0.558 ± 0.046	0.905 ± 0.016
XGBoost	native + general AL perplexities	1742	1.538	0.632	0.434	0.723	5250	0.932 ± 0.054	0.914	0.543 ± 0.048	0.901 ± 0.017
XGBoost	all models perplexities	1742	1.452	0.672	0.465	0.742	5250	0.962 ± 0.058	0.897	0.509 ± 0.051	0.895 ± 0.018
XGBoost	tf-idf	1742	1.200	0.791	0.474	0.838	5250	0.929 ± 0.053	0.897	0.540 ± 0.049	0.904 ± 0.016
2. Text Classifiers											
BERT	–	1742	1.728	0.544	0.326	0.610	5250	1.152 ± 0.051	0.817	0.384 ± 0.061	0.240 ± 0.070
3. Prompt Models											
Gemma 2B*	AES1	1742	2.806	0.231	0.160	-0.209	5250	2.839 ± 0.052	0.232	-0.011 ± 0.079	-1.794 ± 0.063
Gemma 2B*	AES2	1742	2.273	0.418	0.160	0.319	5250	1.598 ± 0.027	0.607	-0.056 ± 0.072	-0.538 ± 0.048
Gemma 7B*	AES1*	1742	1.220	0.782	0.418	0.838	5250	1.233 ± 0.040	0.771	0.353 ± 0.069	-0.261 ± 0.090
Gemma 7B*	AES2*	1742	1.359	0.732	0.152	0.805	5250	1.116 ± 0.032	0.832	0.348 ± 0.065	0.067 ± 0.063
LLaMA 3 8B	**AES1**	**1742**	**1.016**	**0.877**	**0.539**	**0.910**	5250	1.010 ± 0.031	0.879	0.328 ± 0.067	0.450 ± 0.031
LLaMA 3 8B	AES2	1742	1.282	0.764	0.172	0.824	5250	1.025 ± 0.043	0.876	0.402 ± 0.065	0.248 ± 0.069
Mistral 7B*	AES1*	1742	1.037	0.871	0.564	0.899	5250	1.061 ± 0.047	0.858	0.340 ± 0.063	0.453 ± 0.046
Mistral 7B	AES2	1742	1.166	0.816	0.274	0.873	5250	0.955 ± 0.024	0.901	0.435 ± 0.050	0.494 ± 0.025

baselines and prompt-based LLMs, highlighting the limitations of supervised fine-tuning when training and test distributions diverge significantly. The 90-10 split results (Table 4) show improved performance: **Logistic Regression with all models perplexities** achieves F1 0.346 and **XGBoost with all models perplexities** reaches F1 0.345, demonstrating that perplexity features transfer effectively when paired with domain-adapted classifiers.

4.3. Cross-Corpus Generalization Analysis

Table 4 reveals striking patterns in how different feature representations generalize across corpora. Our macro F1 analysis demonstrates that perplexity-based models exhibit remarkable stability across diverse assessment contexts, maintaining consistent performance hierarchies despite domain shifts.

In the zero-shot scenario, perplexity models show robust cross-corpus generalization. XGBoost with tf-idf achieves the highest average F1 (0.390) among statistical models, with strong performance on EFCAMDAT-test (0.785) and moderate but consistent results on external corpora (CELVA-SP: 0.197, KUPA-KEYS: 0.188). XGBoost with native perplexities follows closely (avg 0.383), demonstrating particularly strong performance on in-domain EFCAMDAT-test (0.741) while maintaining reasonable generalization to KUPA-KEYS (0.238) and CELVA-SP (0.169).

The stability of perplexity features becomes even more apparent in the 90-10 split scenario. When paired with domain-adapted classifiers, XGBoost with all models perplexities achieves the highest average F1 (0.446), with balanced performance across all three datasets (CELVA-SP: 0.345, KUPA-KEYS: 0.320, EFCAMDAT-test: 0.673). This represents a 35% relative improvement over its zero-shot performance (0.329), demonstrating that perplexity features effectively capture transferable proficiency signals that can be calibrated to new domains through classifier adaptation.

Notably, perplexity models maintain consistent relative performance rankings across scenarios. Models using native perplexities consistently outperform those with only artificial learner features, while combinations of all perplexity features show the strongest domain adaptation capacity. This consistency contrasts sharply with prompt-based LLMs, which exhibit high variance across datasets (F1 ranging from 0.022 to 0.199) and fail to maintain stable performance hierarchies.

BERT achieves a zero-shot average of 0.254, with KUPA-KEYS performance of 0.253, but shows

Table 4: Results across test datasets by experimental scenario (macro F1 scores)

	Zero-shot				90-10 Split (×30)			
Model & Setup	CELVA-SP	KUPA-KEYS	EFCAMDAT-test	Avg	CELVA-SP	KUPA-KEYS	EFCAMDAT-test	Avg
Statistical Models								
Test set Majority Class (oracle)*	0.086	0.121	0.107	0.105	0.086 ± 0.012	0.180 ± 0.025	0.128 ± 0.018	0.131 ± 0.015
Logistic Reg. native perplexities	0.157	0.228	0.538	0.308	0.286 ± 0.032	0.255 ± 0.035	0.513 ± 0.045	0.351 ± 0.028
Logistic Reg. native + general AL perplexities	0.153	0.197	0.657	0.336	0.286 ± 0.032	0.250 ± 0.036	0.640 ± 0.038	0.392 ± 0.025
Logistic Reg. all models perplexities	0.156	0.285	0.499	0.313	0.346 ± 0.038	0.305 ± 0.040	0.520 ± 0.044	0.390 ± 0.026
Logistic Reg. tf-idf	0.193	0.192	0.731	0.372	0.283 ± 0.031	0.278 ± 0.038	0.635 ± 0.039	0.399 ± 0.024
XGBoost native perplexities	0.169	0.238	0.741	0.383	0.314 ± 0.035	0.315 ± 0.039	0.575 ± 0.042	0.401 ± 0.027
XGBoost native + general AL perplexities	0.173	0.228	0.736	0.379	0.314 ± 0.035	0.310 ± 0.039	0.570 ± 0.043	0.398 ± 0.027
XGBoost all models perplexities	0.177	0.269	0.541	0.329	$\mathbf{0.345 \pm 0.038}$	$\mathbf{0.320 \pm 0.041}$	$\mathbf{0.673 \pm 0.036}$	$\mathbf{0.446 \pm 0.024}$
XGBoost tf-idf	**0.197**	**0.188**	**0.785**	**0.390**	0.294 ± 0.033	0.282 ± 0.037	0.638 ± 0.038	0.405 ± 0.025
Text Classifiers								
BERT	0.128	0.253	0.380	0.254	0.141 ± 0.025	0.248 ± 0.028	0.472 ± 0.050	0.287 ± 0.021
Prompt Models								
Gemma 2B AES1	0.063	0.174	0.022	0.086	0.056 ± 0.010	0.179 ± 0.021	0.022 ± 0.004	0.086 ± 0.008
Gemma 2B* AES2	0.063	0.199	0.022	0.095	0.057 ± 0.001	0.175 ± 0.006	0.022 ± 0.002	0.085 ± 0.002
Gemma 7B AES1	0.171	0.160	0.149	0.160	0.168 ± 0.025	0.263 ± 0.053	0.150 ± 0.010	0.194 ± 0.020
Gemma 7B* AES2*	0.118	0.104	0.149	0.124	0.165 ± 0.027	0.159 ± 0.027	0.148 ± 0.007	0.157 ± 0.013
LLaMA 3 8B AES1	0.183	0.050	0.198	0.144	0.170 ± 0.029	0.058 ± 0.020	0.199 ± 0.007	0.142 ± 0.012
LLaMA 3 8B AES2	0.142	0.093	0.198	0.144	0.202 ± 0.028	0.126 ± 0.018	0.197 ± 0.006	0.175 ± 0.011
Mistral 7B AES1	0.178	0.047	0.193	0.139	0.161 ± 0.019	0.047 ± 0.006	0.192 ± 0.004	0.133 ± 0.007
Mistral 7B AES2	0.132	0.068	0.193	0.131	0.186 ± 0.025	0.050 ± 0.008	0.194 ± 0.005	0.143 ± 0.009

limited generalization to CELVA-SP (0.128) and lacks the adaptability demonstrated by perplexity models in the 90-10 split scenario. This highlights a critical trade-off: while BERT may excel on specific datasets similar to its fine-tuning domain, perplexity-based approaches offer more consistent and adaptable performance across diverse assessment contexts.

The results establish that **feature-level transfer outperforms model-level transfer** for cross-domain proficiency assessment. Maintaining EFCAMDAT-trained perplexity features while adapting only the classifier (90-10 split) yields superior and more stable results than either pure zero-shot transfer or end-to-end fine-tuning, providing a practical framework for deploying CEFR classification systems across new educational contexts.

An important observation across both datasets is that native perplexity alone constitutes a strong baseline, with AL perplexities providing only marginal improvements. On KUPA-KEYS, XGBoost with native perplexities (RMSE 0.723) performs comparably to the best AL-augmented configuration (RMSE 0.707), while on CELVA-SP, adding AL features sometimes degrades performance. This could potentially reflect the fact that artificial learners would benefit from more extensive training data or from being implemented using more modern language model architectures. Alternatively, this raises the question of whether a stronger pretrained language model (e.g., LLaMA or Mistral) used as a perplexity source, rather than as a zero-shot classifier, could yield more discriminative features without requiring task-specific artificial learner training. Nevertheless, AL perplexities remain theoretically motivated: they capture level-specific linguistic expectations that native perplexity cannot, and the 90-10 split results show that the full set of AL features provides the strongest domain adaptation capacity (average F1 0.446), suggesting their value may be most apparent when combined with domain-adapted classifiers.

5. Conclusion

Our perplexity-based feature engineering pipeline demonstrates strong performance on CEFR classification tasks while revealing important insights about cross-corpus generalization. On KUPA-KEYS, XGBoost with native + general AL perplexities establishes a new state-of-the-art benchmark with RMSE 0.707, substantially outperforming fine-tuned BERT and all prompting methods. Notably, all perplexity-based models achieve strong results on this dataset (RMSE 0.707–0.893), demonstrating the robustness of our perplexity feature approach for process-logged writing assessments.

We also establish the first comprehensive benchmark on CELVA-SP across multiple modeling approaches, revealing different generalization patterns. Unlike KUPA-KEYS, prompt-based LLMs outperform statistical methods on this context-specific academic writing corpus, with LLaMA 3 8B achieving the best prompt performance (RMSE 1.016) compared to our best statistical model (XGBoost TF-IDF, RMSE 1.200). This cross-corpus variance mirrors the behavior of zero-shot prompting methods: our approach excels on controlled assessments but shows reduced generalization to specialized academic contexts.

A key contribution of our work lies in the interpretability of native model perplexity features, which quantify linguistic expectedness and provide transparent signals that enable strong performance while maintaining explainability. However, proficiency-specific perplexities from artificial learner models did not consistently improve prediction accuracy, suggesting that further research is needed to effectively integrate these features into CEFR classification systems.

Future work should explore using perplexity from stronger pretrained models (e.g., LLaMA, Mistral)

as feature sources, which may yield more discriminative proficiency signals without task-specific training. Additionally, aligning subtoken-level perplexity scores to human word boundaries through aggregation strategies (e.g., averaging subtoken perplexities per word) would improve pedagogical interpretability and enable direct mapping to word-level linguistic constructs used in CEFR descriptors. The similar performance between native perplexity alone and native + artificial learner configurations suggests that the artificial learners may not have been sufficiently pre-trained to capture level-specific linguistic patterns distinct from the native model. A dedicated investigation into efficient pre-training strategies for artificial learners, including larger model architectures and longer training schedules, is a promising direction for improving the discriminative power of proficiency-specific perplexity features.

Limitations

Our study has several limitations. First, perplexity features are computed at the LLM subtoken level, which does not directly correspond to human word-level units used in language pedagogy. While the standard perplexity formula normalizes across sequence length, enabling cross-text comparison, the subtoken granularity may obscure pedagogically meaningful patterns at the word or phrase level. Second, our artificial learner models are trained exclusively on EFCAMDAT, a general-purpose learner corpus, which may limit their effectiveness on domain-specific corpora such as CELVA-SP where register and genre differences confound proficiency signals. Third, our artificial learner models were trained on a single 16GB GPU, which constrained model size and training duration; larger models with more extensive pre-training may better capture proficiency-specific patterns but require significantly more computational resources.

Ethics Statement

In accordance with the curators of the EFCAMDAT corpus, we have planned to make our models pre-trained on the EFCAMDAT accessible on the web server hosting the EFCAMDAT data.

Bibliographical References

KV Aditya Srivatsa, Kaushal Kumar Maurya, and Ekaterina Kochmar. 2025. Can LLMs Reliably Simulate Real Students' Abilities in Mathematics and Reading Comprehension? *arXiv e-prints*, page arXiv:2507.08232.

Tatsuya Aoyama and Nathan Schneider. 2024. Modeling nonnative sentence processing with L2 language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4927–4940, Miami, Florida, USA. Association for Computational Linguistics.

Luca Benedetto, Giovanni Aradelli, Antonia Donvito, Alberto Lucchetti, Andrea Cappelli, and Paula Buttery. 2024. Using LLMs to simulate students' responses to exam questions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11351–11368.

Luca Benedetto, Gabrielle Gaudeau, Andrew Caines, and Paula Buttery. 2025. Assessing how accurately large language models encode and apply the common european framework of reference for languages. *Computers and Education: Artificial Intelligence*, 8:100353.

Thomas Gaillat, Andrew Simpkin, Nicolas Ballier, Bernardo Stearns, Annanda Sousa, Manon Bouyé, and Manel Zarrouk. 2022. Predicting CEFR levels in learners of English: The use of microsystem criterial features in a machine learning approach. *ReCALL*, 34(2):130–146.

Jeroen Geertzen, Theodora Alexopoulou, Anna Korhonen, et al. 2013. Automatic linguistic annotation of large scale l2 databases: The ef-cambridge open language database (efcamdat). In *Proceedings of the 31st Second Language Research Forum. Somerville, MA: Cascadilla Proceedings Project*, pages 240–254.

Elma Kerz, Daniel Wiechmann, Yu Qiao, Emma Tseng, and Marcus Ströbel. 2021. Automated classification of written proficiency levels on the CEFR-scale through complexity contours and RNNs. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 199–209.

Wonbin Kim. 2024. Let's make an artificial learner to analyze learners' language! , (70):167–193.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9):1–35.

Cyriel Mallart, Andrew Simpkin, Rémi Venant, Nicolas Ballier, Bernardo Stearns, Jen Yu Li, and Thomas Gaillat. 2023. A new learner language data set for the study of English for Specific Purposes at university level. In *Proceedings of the*

4th Conference on Language, Data and Knowledge - LDK 2023, volume 1, pages 281–287, Vienna, Austria.

Elijah Mayfield and Alan W Black. 2020. Should you fine-tune BERT for automated essay scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 151–162, Seattle, WA, USA → Online. Association for Computational Linguistics.

Jay Mohta, Kenan Ak, Yan Xu, and Mingwei Shen. 2023. Are large language models good annotators? In *Proceedings on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models" at NeurIPS 2023 Workshops*, volume 239 of *Proceedings of Machine Learning Research*, pages 38–48. PMLR.

Itamar Shatz. 2020. Refining and modifying the EFCAMDAT: Lessons from creating a new corpus from an existing large-scale English learner language database. *International Journal of Learner Corpus Research*, 6(2):220–236.

Bernardo Stearns, Nicolas Ballier, Thomas Gaillat, Andrew Simpkin, and John P McCrae. 2024. Evaluating the generalisation of an artificial learner. In *Swedish Language Technology Conference and NLP4CALL*, pages 199–208.

David Strohmaier and Paula Buttery. 2024. Semantic error prediction: Estimating word production complexity. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 209–225, Rennes, France. LiU Electronic Press.

Hakyung Sung and Kristopher Kyle. 2024. Leveraging pre-trained language models for linguistic analysis: A case of argument structure constructions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7302–7314, Miami, Florida, USA. Association for Computational Linguistics.

Satoru Uchida and Masashi Negishi. 2025. Assigning CEFR-J levels to English learners' writing: An approach using lexical metrics and generative AI. *Research Methods in Applied Linguistics*, 4(1):100199.

Georgios Velentzas, Andrew Caines, Rita Borgo, Erin Pacquetet, Clive Hamilton, Taylor Arnold, Diane Nicholls, Paula Buttery, Thomas Gaillat, Nicolas Ballier, et al. 2024. Logging keystrokes in writing by English learners. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10725–10746.

Songlin Xu, Xinyu Zhang, and Lianhui Qin. 2024. EduAgent: Generative student agents in learning.

Victoria Yaneva, King Yiu Suen, Le An Ha, Janet Mee, Milton Quranda, and Polina Harik. 2024. Automated scoring of clinical patient notes: Findings from the Kaggle competition and their translation into practice. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 87–98, Mexico City, Mexico. Association for Computational Linguistics.