EduNLP at BEA 2026 Shared Task 1: Multi-Model Ensemble with Feature-Augmented Transformers for Vocabulary Difficulty Prediction

Avinash Kumar Sharma
Indian Institute of Technology Madras
avics2020@gmail.com

Abstract

We describe our system submitted to the BEA 2026 Shared Task on Vocabulary Difficulty Prediction for English Learners. Our approach combines handcrafted linguistic features with fine-tuned XLM-RoBERTa transformers in a multi-model ensemble, participating in both the closed and open tracks. For the closed track, we ensemble a per-L1 transformer with a LightGBM model trained on 12 features including word frequency, cognate similarity, and part-of-speech indicators. For the open track, we extend this with cross-lingual training across all three L1s and a feature-augmented transformer architecture that concatenates numeric features with the transformer's pooled representation. Our system outperforms the baselines on both tracks across all three L1s (Spanish, German, and Mandarin), with best RMSEs of 1.058 (closed, CN) and 0.992 (open, CN). Post-hoc error analysis on the released test labels reveals that polysemous words in rare senses and nominalized *-ing* forms constitute the primary failure mode, contributing 58% higher RMSE than other words. We additionally report on a negative result where a post-hoc Ridge regression blend overfit on test despite strong development set performance.

1 Introduction

Vocabulary knowledge is a fundamental component of language proficiency, and accurately estimating the difficulty of vocabulary items is essential for adaptive learning systems, automated test construction, and personalized instruction. Traditional approaches to estimating word difficulty rely on large-scale testing with real learners followed by psychometric calibration, a process that is both costly and time-consuming (Schmitt et al., 2024).

The BEA 2026 Shared Task on Vocabulary Difficulty Prediction for English Learners (Felice and Skidmore, 2026) challenges participants to build regression models that predict psychometrically calibrated difficulty scores (GLMM scores) for English vocabulary items, given the learner's first language (L1). The task covers three L1s: Spanish (ES), German (DE), and Mandarin Chinese (CN). It offers two tracks: a **closed track** restricted to the provided data and standard pre-trained transformers, and an **open track** allowing external resources and cross-lingual data combination.

We present a multi-model ensemble system that combines handcrafted linguistic features with fine-tuned transformer models. Our approach is motivated by exploratory data analysis revealing that surface features such as word frequency and cognate similarity capture complementary information to what transformer models learn from text. Our system participates in both tracks and outperforms the baselines on all six track-L1 combinations.

Our contributions are: (1) a systematic comparison of feature-based, transformer-based, and hybrid approaches across three L1s; (2) a feature-augmented transformer architecture concatenating numeric features with the transformer's pooled output; (3) detailed post-hoc error analysis revealing polysemy and nominalization as systematic failure modes; and (4) a report of a Ridge blend that overfit on test despite strong development set performance.

2 Task and Data

The shared task uses the Extended KVL Dataset for NLP (Skidmore et al., 2025), derived from the British Council's Knowledge-based Vocabulary Lists (Schmitt et al., 2024). The dataset was created by testing over 100,000 English learners through a vocabulary recall task: learners see an L1 context sentence, an L1 source word, and a partial English spelling clue (first letter plus blanks), and must produce the English target word. Difficulty scores were estimated using a Generalized Linear Mixed Model (GLMM) with Rasch parame-

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 1024–1028

Split	Items/L1	Total	Item IDs
Train	6,091	18,273	1–6,091
Dev	677	2,031	6,092–6,768
Test	748	2,244	6,769–7,516

Table 1: Dataset statistics. Each L1 (ES, DE, CN) has identical item counts with parallel item IDs.

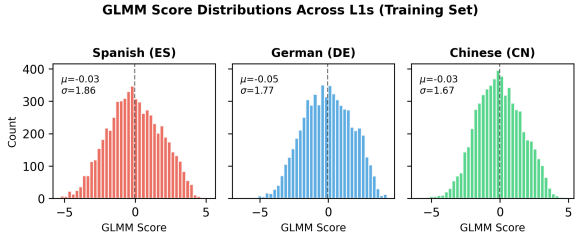

Figure 1: GLMM score distributions across L1s (training set). All three L1s follow approximately normal distributions with similar shape but different spread.

terization, simultaneously estimating learner ability and item difficulty (Schmitt et al., 2021). Lower GLMM scores indicate harder words.

Each item consists of seven fields: item ID, L1 code, English target word, part of speech (POS), partial spelling clue, L1 source word, and L1 context sentence. The training data GLMM scores are approximately normally distributed with mean near zero and standard deviation of 1.67 (CN) to 1.86 (ES), ranging from approximately -6.5 to $+5.1$, as shown in Figure 1.

Notably, only 17.4% of test words appear in the training or development sets, meaning 82.5% of test predictions must generalize to entirely unseen words.

3 System Description

3.1 Exploratory Analysis and Motivation

We conducted exploratory data analysis prior to model development, which informed several key design decisions. Word frequency on the Zipf scale (Van Heuven et al., 2014) shows the strongest correlation with GLMM scores among simple features ($r = +0.43$ for ES). Word length correlates negatively with difficulty ($r = -0.33$ for ES, -0.44 for CN), with the stronger effect for Mandarin consistent with the absence of orthographic overlap between Chinese and English. Figure 2 illustrates these two key relationships.

Cross-L1 GLMM score correlations are moderate (ES–DE: $r = 0.684$, ES–CN: $r = 0.634$, DE–CN: $r = 0.662$), indicating approximately 40–47% shared variance across L1s.

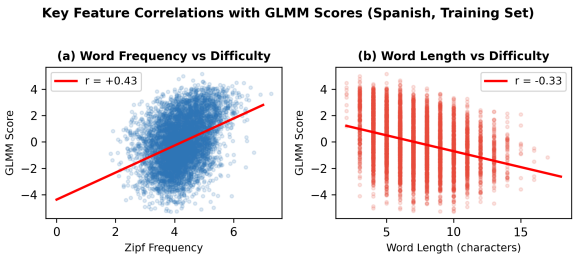

Figure 2: Key feature correlations with GLMM scores (Spanish training set). (a) Word frequency and (b) word length are the strongest individual predictors of difficulty.

These findings motivated three design choices: (1) including word frequency and cognate similarity as explicit numeric features; (2) combining per-L1 and cross-lingual training strategies; and (3) ensembling feature-based and transformer-based models.

3.2 Feature Engineering

We extract 12 numeric features from each vocabulary item, organized into four groups:

Word frequency. Zipf-scale frequency (Speer et al., 2017) for the English target word (`en_zipf`) and L1 source word (`l1_zipf`).

Cognate similarity. Normalized Levenshtein edit distance (`edit_distance`) and character-level Jaccard overlap (`jaccard_overlap`) between the English target and L1 source words.

Surface features. Word length (`word_len`), source word length (`source_len`), syllable count via vowel group counting (`syllable_count`), and clue ratio (`clue_ratio`).

POS indicators. Binary features for noun, verb, adjective, and adverb.

3.3 Model Architectures

We develop four complementary models:

Model 1: LightGBM (Ke et al., 2017). A gradient boosting model trained per-L1 on the 12 features with 1,000 estimators, learning rate 0.05, max depth 6, and subsampling 0.8. While this model alone does not match transformer baselines (dev RMSE 1.39–1.56), it produces more accurate predictions than the transformer for approximately 27% of individual words on the development set.

Model 2: Per-L1 XLM-RoBERTa (closed track). We fine-tune `xlm-roberta-base` (Conneau et al., 2020) separately for each L1. Input: `{source} </s> {context} </s> {clue} </s>`

System	Track	ES	DE	CN
Glite (1st)	Closed	0.903	0.885	0.776
EduNLP	**Closed**	**1.176**	**1.124**	**1.058**
Baseline	Closed	1.257	1.258	1.140
Sakura (1st)	Open	0.742	0.723	0.630
EduNLP	**Open**	**1.143**	**1.071**	**0.992**
Baseline	Open	1.198	1.166	1.034

Table 2: Official test results (RMSE ↓). Best EduNLP run per track shown. Double line separates tracks.

`{target}`. We use mean pooling over the final hidden states followed by dropout ($p = 0.1$) and a linear head.

Model 3: Cross-lingual XLM-RoBERTa (open track). A single model trained on all three L1s combined (18,273 examples), with an L1 identifier prefix (e.g., `[ES]`). This triples the training data.

Model 4: Feature-augmented XLM-RoBERTa (open track). The transformer's mean-pooled output (768 dims) is concatenated with the 12 standardized features, and fed through an MLP ($780 \to 256 \to 1$) with ReLU and dropout. This directly combines text representations with surface features.

All transformers use AdamW with linear warmup (10%), learning rate 2×10^{-5}, weight decay 0.01, batch size 16, and 5 epochs. We average predictions across 3 seeds (42, 123, 456).

3.4 Ensemble and Calibration

Closed track. Weighted average of per-L1 transformer and LightGBM, with weights optimized on the dev set (65–70% transformer, 30–35% LightGBM).

Open track. Three-model ensemble of cross-lingual transformer, feature-augmented transformer, and LightGBM, with per-L1 optimized weights. We also experiment with a cross-notebook ensemble blending open and closed track models.

Calibration. We apply conservative scaling (1.04–1.06×) around the mean to partially address prediction compression (Section 5.3).

For final predictions, we retrain on combined train+dev data (6,768 items per L1), as permitted by the organizers.

4 Results

4.1 Official Test Results

Table 2 presents our official test results alongside baselines and winning systems.

Model	ES	DE	CN
LightGBM (features only)	1.559	1.475	1.393
XLM-R per-L1	1.322	1.232	1.098
XLM-R cross-lingual	1.219	1.157	1.045
Feature-Aug XLM-R	1.169	1.164	1.014
4-model ensemble	**1.158**	**1.127**	**1.018**
Baseline (closed)	1.357	1.328	1.175
Baseline (open)	1.206	1.149	1.021

Table 3: Ablation study on development set (RMSE ↓). Each row adds capability. Double line separates our models from baselines.

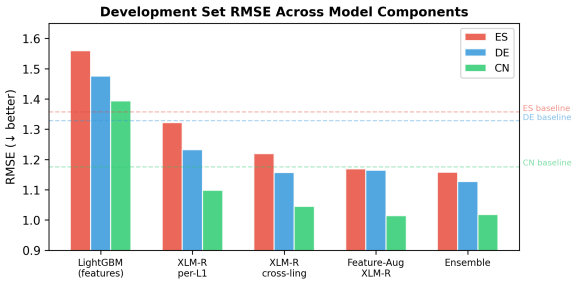

Figure 3: RMSE progression across model components on the development set. Dashed lines indicate organizer baselines (closed track).

Our system outperforms the baseline on both tracks across all L1s, with improvements of 0.081–0.134 (closed) and 0.042–0.095 (open) RMSE. A gap remains to the winning systems.

4.2 Ablation Study

To understand the contribution of each component, we evaluate four individual models and their ensemble on the development set (Table 3). LightGBM uses only the 12 handcrafted features described in Section 3.2. XLM-R per-L1 is a fine-tuned transformer trained separately for each language. XLM-R cross-lingual trains a single transformer on all three L1s combined. Feature-Aug XLM-R concatenates the transformer output with the 12 numeric features before prediction. The 4-model ensemble computes a weighted average of all four models, with weights optimized on the development set.

Table 3 shows the contribution of each component on the development set.

The progression (Figure 3) shows consistent improvement: transformers reduce RMSE by 0.15–0.30 over features, cross-lingual training adds 0.05–0.10, and the ensemble consistently outperforms all individual models. Among the 12 features, word frequency (Zipf scale) and cognate similarity (edit distance) contribute most to the feature-augmented model's gains. On the development set, remov-

Zone	N	RMSE	%Err	$\bar{\epsilon}$
Very Easy (>3)	44	0.920	3.8%	+0.66
Easy (1 to 3)	194	0.943	17.7%	+0.41
Medium (−1 to 1)	275	0.901	22.9%	−0.10
Hard (−3 to −1)	197	1.225	30.3%	−0.75
Very Hard (<−3)	38	2.552	25.4%	−2.19

Table 4: Error analysis by GLMM score zone (ES, open ensemble, test set). %Err = percentage of total squared error. $\bar{\epsilon}$ = mean signed error.

Word	POS	True	Pred.	Err	Issue
dining	noun	−4.67	+2.22	−6.89	Nominal.
baking	noun	−3.79	+0.77	−4.56	Nominal.
clear	adv	−4.14	+0.11	−4.25	Rare POS
amount	verb	−4.04	−0.55	−3.49	Rare POS
very	adj	−2.01	+1.27	−3.28	Rare POS
received	adj	−5.90	−2.26	−3.63	Rare POS

Table 5: Selected worst predictions (ES test). All involve polysemy or rare POS usage.

ing frequency alone increases RMSE by 0.08–0.12 across L1s, while removing cognate features increases RMSE by 0.04–0.07. POS indicators and word length provide smaller but consistent contributions.

5 Analysis and Discussion

5.1 Error Analysis by Difficulty Zone

Using the released test labels, we analyze prediction errors stratified by GLMM score zones for Spanish (Table 4).

The "very hard" zone contributes 25.4% of total error despite comprising only 5.1% of items, with systematic under-prediction of difficulty (mean error −2.19).

5.2 Polysemy and Nominalization Failures

Our worst predictions involve common words in rare senses (Table 5).

Words ending in "-ing" used as nouns are particularly problematic: these 39 words exhibit RMSE of 1.735 compared to 1.101 for all other words, a 58% increase. The model sees "dining" and associates it with the common verb "to dine," predicting it as easy. However, producing "dining" as a noun in a vocabulary test requires knowledge of the nominalized form, which is substantially harder. We observe similar patterns across all three L1s. The same 18 nominalized *-ing* words appear as nouns in each test set. However, difficulty varies substantially across L1 backgrounds. For example, "dining" as a noun has GLMM scores of −4.67 (ES), −2.00 (DE), and −1.03 (CN), indicating that Spanish speakers find this usage most challenging. Similarly, "received" as an adjective is consistently difficult across all L1s (ES: −5.90, DE: −5.06, CN: −3.88), while "clear" as an adverb shows strong L1 variation (ES: −4.14, DE: +2.64, CN: −3.12), suggesting that German speakers find this usage relatively easy compared to other L1 groups.

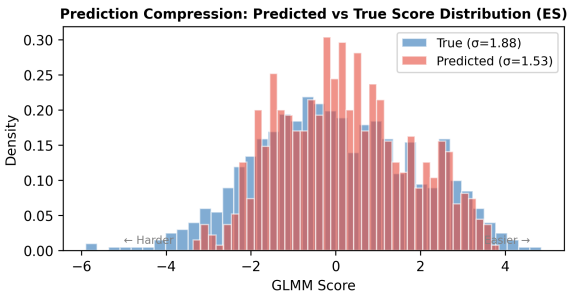

Figure 4: Predicted vs. true GLMM score distributions (ES test set). Predictions are compressed relative to the true distribution.

5.3 Prediction Compression

Our predictions exhibit variance compression: the predicted standard deviation is only 81% of the true value for Spanish ($\sigma_{pred} = 1.53$ vs. $\sigma_{true} = 1.88$; Figure 4). We attempted post-hoc calibration by scaling predictions, but this provided minimal improvement, indicating the gap to top systems lies in prediction quality (Pearson $r = 0.802$ vs. winner's $r = 0.920$) rather than calibration alone.

5.4 Ridge Blend Failure

One of our open track runs incorporated a Ridge regression with interaction features (frequency × word length, cognate × frequency), blended with our ensemble at 85%/15%. While this improved development set performance, it failed on test: RMSE of 1.397 (DE) and 1.421 (CN), worse than the baseline. The interaction features captured patterns in the 677-item development set that did not generalize to the test set. This outcome highlights the risk of incorporating models validated only on a single held-out split without cross-validation. We did not perform cross-validation of the blend weights prior to submission, which we identify as a methodological limitation.

5.5 Gap to Top Systems

Our system explains 64.3% of score variance ($r^2 = 0.643$ for ES) compared to the winners' 84.6% ($r^2 = 0.846$). We hypothesize that this

gap is driven by three factors: (1) model scale (`xlm-roberta-base` has 270M parameters, while larger variants such as `xlm-roberta-large` or DeBERTa-v3-large have 430M–560M); (2) model diversity across multiple architectures; and (3) the use of generative large language models fine-tuned for regression in the open track, in addition to encoder models like ours. These three factors likely account for the bulk of the 0.27–0.40 RMSE gap.

6 Conclusion

We presented a multi-model ensemble for vocabulary difficulty prediction, combining linguistic features with XLM-RoBERTa transformers across both tracks. Our system outperforms baselines on all six track-L1 combinations. Error analysis reveals polysemous words in rare senses as the primary failure mode, with nominalized *-ing* forms showing 58% higher RMSE. The gap to top systems is driven primarily by model scale and diversity. For future work, we identify larger models, multi-architecture stacking, and sense-aware features as the most promising directions.

Limitations

Our system has several limitations. We use only `xlm-roberta-base` (270M parameters), limiting model capacity. Our multi-seed averaging uses only 3 seeds. Feature engineering operates at the word level without sense-level or morphological features. The Ridge blend was validated on a single split rather than with cross-validation, leading to overfitting. We do not leverage LLMs for the open track. Finally, while we have extended our error analysis to include cross-lingual comparisons, a more systematic analysis across all L1s remains future work.

Ethics Statement

This work uses the publicly available BEA 2026 shared task dataset. The data contains vocabulary items and aggregated difficulty scores; no individual learner data is accessible. Our system is intended to support educational applications. Automated difficulty prediction should complement, not replace, expert judgment in high-stakes assessment.

Acknowledgements

We thank Dr. Shabana K M, Senior Project Scientist at the Wadhwani School of Data Science and AI, Indian Institute of Technology Madras, for her guidance and feedback on this work. We also thank the shared task organizers for organizing the task and providing the dataset. Computing resources were provided by Kaggle.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.

Mariano Felice and Lucy Skidmore. 2026. Findings of the bea 2026 shared task on vocabulary difficulty prediction for english learners. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, San Diego, California. Association for Computational Linguistics.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Norbert Schmitt, Karen Dunn, Barry O'Sullivan, Laurence Anthony, and Benjamin Kremmel. 2021. Introducing knowledge-based vocabulary lists (kvl). *Tesol Journal*, 12(4).

Norbert Schmitt, Karen Dunn, Barry O'Sullivan, Laurence Anthony, and Benjamin Kremmel. 2024. *Knowledge-based vocabulary lists*. University of Toronto Press.

Lucy Skidmore, Mariano Felice, and Karen Dunn. 2025. Transformer architectures for vocabulary test item difficulty prediction. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 160–174, Vienna, Austria. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Walter JB Van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. Subtlex-uk: A new and improved word frequency database for british english. *Quarterly journal of experimental psychology*, 67(6):1176–1190.