Data Asgardians at BEA 2026 Shared Task 1: A Hybrid Transformer–Feature Ensemble for L1-Aware English Vocabulary Difficulty Prediction

Adrián Pineda Sánchez[1], Sabur Butt[2], Héctor Gibrán Ceballos Cancino[2]

[1]School of Engineering and Sciences, Tecnológico de Monterrey, Mexico
[2]Institute for the Future of Education (IFE), Tecnológico de Monterrey, Mexico
A00834710@tec.mx, saburb@tec.mx, ceballos@tec.mx

Correspondence: Sabur Butt, saburb@tec.mx

Abstract

This paper presents our system for the BEA 2026 Shared Task on Vocabulary Difficulty Prediction for English Learners (Felice and Skidmore, 2026). The task requires predicting psychometrically calibrated GLMM difficulty scores for English vocabulary items across three learner first-language (L1) backgrounds: Spanish (ES), German (DE), and Mandarin Chinese (CN). Our approach studies how hand-crafted linguistic features can complement contextual multilingual transformer representations. We engineer 33 phonological, morphological, semantic, contextual, and cross-lingual features, and evaluate feature-only regressors, Solo transformer models, Hybrid transformer models, and prediction-level ensembling. Our official Closed Track submissions were generated with `XLM-RoBERTa-large` Solo and Hybrid models, which improved over the official baseline for all three L1 groups, achieving test RMSEs of 1.182 (ES), 1.117 (DE), and 1.006 (CN) with a mean of 1.103. We then conducted a post-submission refinement using `mDeBERTa-v3-base` components and a Ridge stacking ensemble, which further reduced test RMSE to 1.037 (ES), 0.997 (DE), and 0.913 (CN), with a mean of 0.982, a mean improvement of 0.121 over our best `XLM-RoBERTa-large` system.

1 Introduction

Predicting the difficulty of English vocabulary items for second-language learners has direct applications in adaptive assessment, personalised content generation, and computer-adaptive testing. In educational measurement, item difficulty is commonly estimated with Generalised Linear Mixed Models (GLMM) (Baayen et al., 2008), which provide reliable psychometric calibration from learner response data. However, this process is costly to scale, since each new item bank typically requires extensive response collection. The BEA 2026 Shared Task on Vocabulary Difficulty Prediction (Felice and Skidmore, 2026) addresses this limitation by framing difficulty estimation as a supervised NLP task over the British Council's Knowledge-based Vocabulary Lists (KVL) (Skidmore et al., 2025), a multilingual resource that conditions vocabulary difficulty on the learner's first language (L1).

Recent lexical complexity systems commonly fine-tune pre-trained language models directly on the regression objective (Shardlow et al., 2021). While effective, this approach may underuse linguistic factors known to shape vocabulary acquisition, such as frequency, phonological complexity, polysemy, morphological structure, and the distance between English and the learner's L1 (Laufer and Goldstein, 2004). These factors suggest that L1-aware difficulty prediction should benefit from combining contextual transformer representations with explicit linguistic features.

Our system follows this feature-aware modelling perspective. We combine contextual multilingual representations with 33 engineered features covering phonological, morphological, semantic, contextual, and cross-lingual properties. Our official Closed Track submissions were based on `XLM-R Large` Solo and Hybrid models. We then developed a refined post-submission system based on `mDeBERTa-v3-base`, feature-only regressors, and a cross-validated Ridge stacking ensemble.

This paper makes three contributions. First, we present an L1-aware system for vocabulary difficulty prediction that integrates multilingual transformer representations with linguistically motivated features. Second, we compare feature-only regressors, Solo transformers, Hybrid feature-augmented transformers, and prediction-level ensembling across Spanish, German, and Mandarin Chinese. Third, we show that the strongest performance among our systems is obtained through prediction-level stacking, indicating that comple-

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 1137–1145

mentary model signals are more effective than relying on a single transformer or on direct feature concatenation alone.

2 Task and Data

The task data come from the Extended KVL Dataset (Skidmore et al., 2025), which provides GLMM-calibrated difficulty scores for English vocabulary items across three L1 backgrounds: Spanish (ES), German (DE), and Mandarin Chinese (CN). Each instance includes the English target word, its part of speech, a letter-based clue, an L1 translation, and an L1 context sentence. GLMM scores are continuous logit values, with higher values indicating easier items.

Table 1 shows representative examples across the three L1 groups, including the clue format, translation, context, and difficulty score.

As shown in Table 2, the dataset is balanced across L1 groups, with 6,091 training instances, 677 development instances, and 748 test instances per language. The Closed track restricts systems to the provided data and standard NLP tools, while the Open track allows external resources. Systems are ranked by RMSE, with Pearson's r used as a secondary metric.

3 Related Work

The BEA 2026 Shared Task on vocabulary difficulty prediction (Felice and Skidmore, 2026) frames item difficulty estimation as a supervised regression problem over psychometrically calibrated GLMM scores. Unlike general lexical complexity benchmarks, the task requires predicting difficulty across multiple learner first-language (L1) groups, making the problem dependent not only on the English target word and its context, but also on cross-linguistic effects between English and the learner's L1.

Recent lexical complexity systems have largely relied on pre-trained transformer models. In the SemEval-2021 shared task, Shardlow et al. (2021) showed that fine-tuned contextual models are strong predictors of lexical complexity, while traditional lexical features, especially frequency-based ones, can still provide complementary signal.

For L1-aware vocabulary difficulty, however, linguistic and psycholinguistic factors remain important. Prior work shows that word length, phonological similarity, morphological complexity, and related form-based properties affect vocabulary learning beyond frequency alone (Laufer and Goldstein, 2004). This suggests that difficulty is not only a property of the target word, but also of its relationship to the learner's linguistic background.

Our work follows this perspective by combining contextual transformer representations with explicit phonological, morphological, semantic, and cross-lingual features. This hybrid setting allows the model to capture both usage-dependent difficulty and L1-specific signals under the limited-data conditions of the shared task.

4 Exploratory Data Analysis

Before modelling, we examine the GLMM target distribution and its structure across L1 groups under the Closed Track setting. This analysis informs two design choices: using linguistically motivated features, and modelling each L1 separately while still exploiting shared difficulty patterns.

Fig. 1 shows that GLMM scores are approximately symmetric and centred near zero for Spanish, German, and Mandarin Chinese, suggesting no major distributional imbalance across L1 groups. The distributions are similar overall, but differ slightly in dispersion (CN $\sigma\approx1.26$, ES $\sigma\approx1.43$, DE $\sigma\approx1.38$), with Mandarin more concentrated and Mandarin/German showing more pronounced negative outliers.

We also find strong correlations for identical items across L1 groups (ES–DE: $r\approx0.85$; ES–CN: $r\approx0.78$; DE–CN: $r\approx0.80$), indicating a shared global difficulty component. However, the remaining L1-specific variation motivates separate per-L1 models and feature engineering that captures cross-linguistic effects. Together, these observations support the hybrid modelling strategy developed in the following sections.

5 Feature Engineering

Motivated by the EDA findings, we design 33 linguistically informed features to capture both shared and L1-specific sources of vocabulary difficulty. The features are derived from WordNet (Miller, 1995), the CMU Pronouncing Dictionary (Carnegie Mellon University, 2014), and PHOIBLE (Moran and McCloy, 2019). Table 3 summarizes the resulting feature groups.

The feature set is organized into four complementary groups:

- **Phonological features** capture English pronunciation complexity, including consonant

L1	Target	POS	Clue	L1 translation	Context	GLMM
ES	*span*	noun	`s__`	lapso	El eclipse solar fue visible durante un breve …	−3.264099
ES	*radically*	adverb	`r_____`	fundamentalmente	Los métodos nuevos son fundamentalmente difere…	−1.871782
DE	*span*	noun	`s__`	Zeitraum	Über einen Zeitraum von fünf Monaten wurden Ze…	−2.769408
DE	*radically*	adverb	`r_____`	fundamental, völlig	Nach unserem Gespräch hat sich seine Einstellu…	−1.567126
CN	*span*	noun	`s__`	持续时间段	智能手机使人的注意力集中时间越来越短。	−2.741183
CN	*radically*	adverb	`r_____`	彻底地，激进地，根本地	我们应当彻底改变只顾经济发展的思路，以保护日渐脆弱的环境。	−1.337120

Table 1: Representative Extended KVL items across Spanish, German, and Mandarin Chinese. Higher GLMM values indicate easier items.

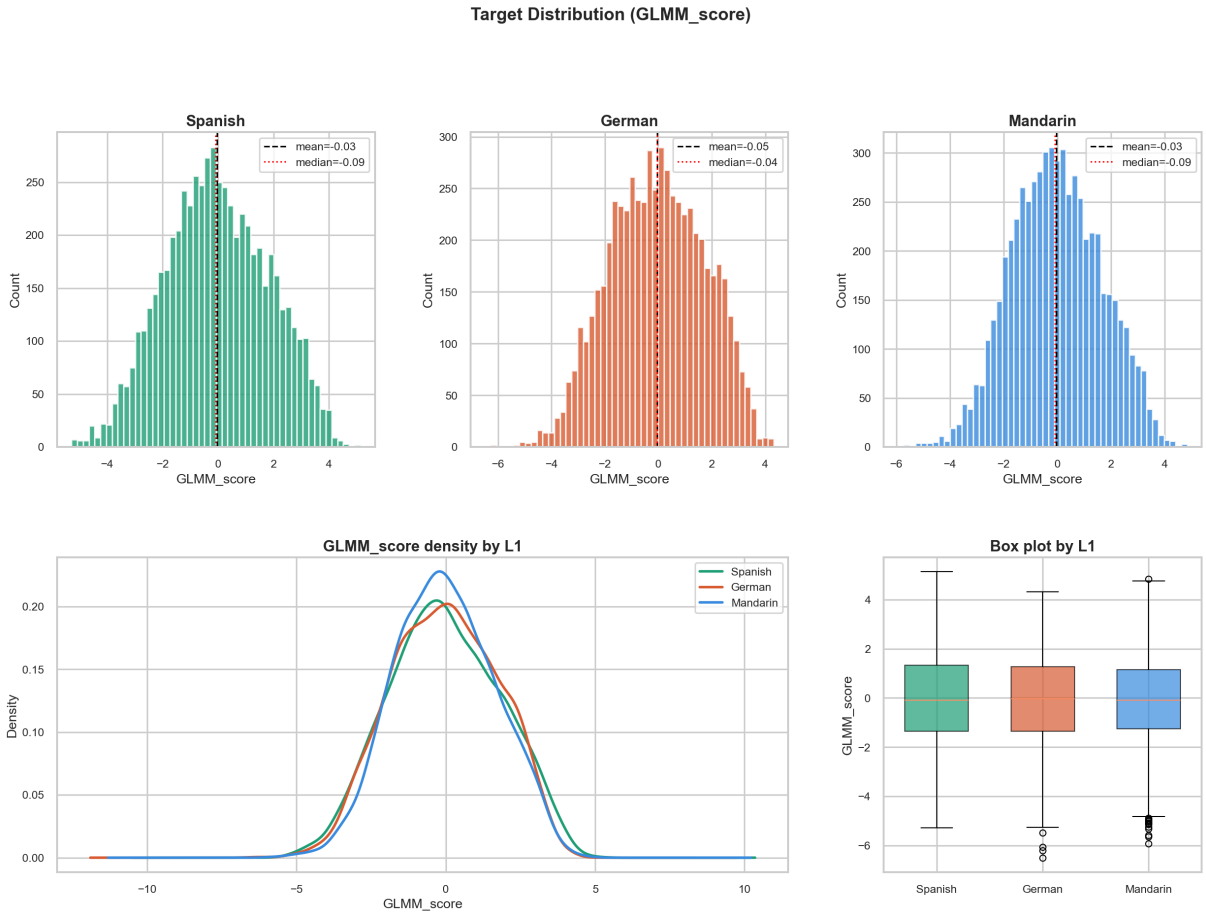

Figure 1: GLMM difficulty score distributions by L1 (training set).

L1	Train	Dev	Test
Spanish (ES)	6,091	677	748
German (DE)	6,091	677	748
Mandarin (CN)	6,091	677	748
Total	18,273	2,031	2,244

Table 2: Extended KVL dataset split sizes per L1.

clusters, word-final clusters, rhoticity, silent letters, and spelling-to-phoneme mismatch. The feature `r_count` counts /r/ phonemes in the CMU pronunciation, not occurrences of the orthographic character *r*.

- **L1-specific phonological features** encode whether the English target contains sounds or graphemic patterns that may be absent or difficult for a given learner L1. For example, `difficult_phonemes` counts English phonemes not present in the learner's L1 inventory according to PHOIBLE.
- **Semantic and morphological features** represent lexical properties from WordNet, including polysemy, homonymy, hypernym depth,

Category	#	Features	Source
Phonological	7	n_consonant_clusters, max_cluster_length, final_cluster_size, r_count, has_r, silent_letters, spelling_phoneme_ratio	CMU Dict
L1-specific phonological	5	difficult_phonemes, has_unfamiliar, has_diphthong, primary_stress_pos, has_l1_flag	PHOIBLE
Semantic & morphological	8	polysemy_pos, polysemy_all, is_homonym, homonym_pos_count, min_sense_depth, morphological_complexity, word_family_size, word_frequency	WordNet
Cross-lingual & contextual	13	cosine_dist_l1_en, clue_ratio, source_word_length, source_word_count, context_char_length, shared_prefix_len, pos_adjective, pos_adverb, pos_determiner, pos_misc, pos_noun, pos_number, pos_preposition, pos_verb	Text / WordNet
Total	**33**	25 numerical + 8 POS one-hot	–

Table 3: Engineered feature groups and sources. cosine_dist_l1_en is a character bigram Jaccard distance between the L1 form and the English target.

morphological complexity, word family size, and lemma frequency. These features approximate how semantically broad, frequent, or morphologically complex a target word is.

- **Cross-lingual and contextual features** describe the relation between the English target and the L1-side information. This includes cosine_dist_l1_en, implemented as character bigram Jaccard distance between the L1 translation and the English target, as well as clue ratio, source-word length, context length, shared prefixes, and POS indicators. The feature clue_ratio is defined as the proportion of orthographic characters of the English target that are revealed by the letter-based clue, i.e., the ratio between the number of revealed characters (non-underscore positions) and the total length of the target word. Higher values correspond to more informative clues and, empirically, to easier items (see Table 4).

To reduce redundancy, we compute pairwise Pearson correlations among candidate features on the training set. For any pair with $|r| > 0.85$, we remove the feature with the lower absolute correlation with the GLMM target. This process yields the final 33-feature set with no remaining pair above the collinearity threshold.

Table 4 reports the strongest feature–target correlations. clue_ratio and word_frequency are the most predictive global features, while the per-L1 values show that feature relevance varies across learner backgrounds. Figure 2 extends this analysis to all 33 features, highlighting both globally predictive signals and L1-specific variation.

Feature	Global	ES	DE	CN
clue_ratio	−0.381	−0.350	−0.353	−0.449
word_frequency	0.314	0.237	0.233	0.488
source_word_length	−0.253	−0.217	−0.311	−0.338
primary_stress_pos	−0.247	−0.216	−0.258	−0.271
cosine_dist_l1_en	−0.246	−0.271	−0.352	0.060

Table 4: Top-5 features by absolute Pearson r with the GLMM difficulty score, computed globally and per L1 on the training set.

6 System Architecture

Our official Closed-track prediction files were generated with XLM-RoBERTa-large Solo and Hybrid models. We then explored whether a smaller multilingual encoder, combined with structured linguistic features and prediction-level ensembling, could improve performance. The refined system is built around mDeBERTa-v3-base and combines four learners through a cross-validated Ridge stacking ensemble (Wolpert, 1992): (i) a Solo transformer, (ii) a Hybrid transformer, (iii) an Optuna-tuned XGBoost regressor, and (iv) a Ridge regressor. We treat this ensemble as the primary system and report XLM-R Large as the official-submission reference.

6.1 Training Protocol

All systems follow the official train/development/test splits (Table 2) and are trained independently for each L1 group (ES, DE, CN), following the L1-specific variation observed in Section 4. Model development follows two phases. First, models are trained on the official training set and selected on the development set using RMSE, with Pearson's r as a secondary metric. This phase is also used for hyperparameter search and, in the stacking ensemble, for selecting the Ridge meta-learner's regularisation strength α using 5-fold out-of-fold RMSE. Second, the selected configuration is retrained on train+dev

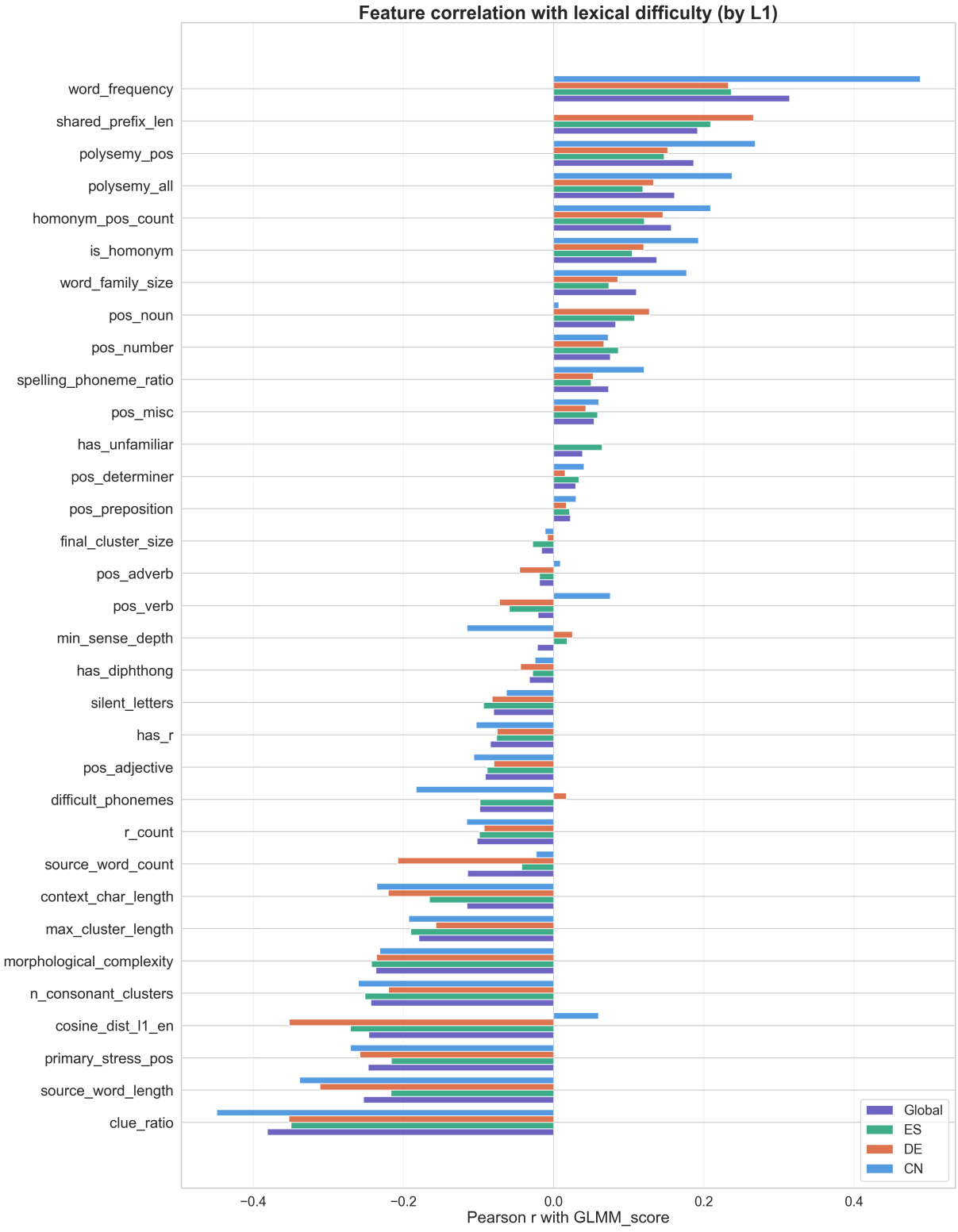

Figure 2: Pearson r between each engineered feature and the GLMM score (training set), computed globally and per L1.

and applied to the test set.

For the Solo and Hybrid transformer variants, we use the same per-L1 grid: learning rate $\{5\times10^{-6}, 10^{-5}, 2\times10^{-5}, 3\times10^{-5}\}$, weight decay $\{0.01, 0.05, 0.10\}$, dropout $\{0.10, 0.15\}$, and epochs $\{3, 5, 7, 8, 10\}$. This produced 42 `mDeBERTa-v3-base` runs and 33 `XLM-R Large` runs. The selected configurations are shown in Table 5. Feature-only XGBoost and Ridge models were tuned with Optuna as described in Sec-

tion 6.4. All transformer experiments were run on rented Google Colab GPU compute, mainly NVIDIA A100 40GB when available, requiring approximately 200 GPU hours.

Encoder	L1	Model	lr	wd	drop	ep
mDeBERTa-base	ES	Solo	3×10^{-5}	0.05	0.10	5
	ES	Hybrid	1×10^{-5}	0.10	0.15	5
	DE	Solo	1×10^{-5}	0.05	0.10	7
	DE	Hybrid	1×10^{-5}	0.10	0.15	5
	CN	Solo	1×10^{-5}	0.10	0.15	10
	CN	Hybrid	2×10^{-5}	0.05	0.10	3
XLM-R Large	ES	Solo	2×10^{-5}	0.01	0.10	8
	ES	Hybrid	5×10^{-6}	0.10	0.15	5
	DE	Solo	1×10^{-5}	0.05	0.10	10
	DE	Hybrid	1×10^{-5}	0.01	0.10	5
	CN	Solo	2×10^{-5}	0.01	0.10	8
	CN	Hybrid	1×10^{-5}	0.05	0.10	5

Table 5: Best hyperparameter configuration per encoder, L1, and model type (Solo vs. Hybrid), selected on the development set. `lr`: learning rate; `wd`: weight decay; `drop`: dropout rate; `ep`: training epochs.

6.2 Solo Transformer

We evaluate two multilingual encoders: `XLM-RoBERTa-large` (Conneau et al., 2020), used in the official submission, and `mDeBERTa-v3-base` (He et al., 2023), used in the refined system. Both models are used in their publicly released pre-trained form, without intermediate task-adaptive pre-training, and are fine-tuned under the same protocol. Hyperparameters are selected independently per L1 (Table 5).

Each encoder is paired with a regression head over the first-token representation of the last hidden layer:

$$\mathbf{h} = \mathbf{H}_{[0]}, \qquad \hat{y} = \mathbf{W}\,\mathrm{Dropout}(\mathbf{h}), \qquad (1)$$

where $\mathbf{H} \in \mathbb{R}^{|T|\times d}$ denotes the sequence of last-layer hidden states. The input concatenates the L1-SOURCE-WORD, L1-CONTEXT, EN-TARGET-CLUE, and EN-TARGET-WORD, tokenised with the encoder's native SentencePiece tokeniser using a maximum length of 192. This format allows the model to jointly attend to the L1 context and the English target information.

6.3 Hybrid Transformer

The Hybrid model augments the transformer representation $\mathbf{h}$ with a 33-dimensional linguistic feature vector $\mathbf{f}$, using median imputation for missing values. The concatenated representation is passed through a two-layer MLP:

$$\hat{y} = \mathbf{W}_2\,\mathrm{Dropout}\left(\phi\left(\mathbf{W}_1\,\mathrm{Dropout}\left([\mathbf{h};\,\mathbf{f}]\right)\right)\right), \qquad (2)$$

with $\mathbf{W}_1 \in \mathbb{R}^{256\times(d+33)}$ and $\mathbf{W}_2 \in \mathbb{R}^{1\times256}$. The hidden size is $d = 768$ for `mDeBERTa-v3-base` and $d = 1024$ for `XLM-R Large`. We use GELU for `mDeBERTa-v3-base` and ReLU for `XLM-R Large`.

Both encoders are fine-tuned with AdamW, mixed-precision training (`fp16`), a linear-warmup and linear-decay schedule with a 10% warmup ratio, gradient clipping at 1.0, and an MSE objective. Learning rate, weight decay, dropout, and number of epochs are tuned per L1.

6.4 Feature-Only Regressors

We also train **XGBoost** (Chen and Guestrin, 2016) and **Ridge** regression models directly on the 33 engineered linguistic features. Both are trained independently per L1, using median imputation; Ridge additionally uses standardisation. Hyperparameters are selected with Optuna (Akiba et al., 2019), using 150 trials per L1 for XGBoost and 100 trials per L1 for Ridge. These models serve both as ensemble components and as feature-only baselines.

6.5 Ridge Stacking Ensemble

The refined system combines the four `mDeBERTa-v3-base` components: Solo ($\hat{y}_S$), Hybrid ($\hat{y}_H$), XGBoost ($\hat{y}_X$), and Ridge ($\hat{y}_R$). For each L1, a Ridge meta-learner combines their predictions:

$$\hat{y}_{\text{ens}} = \mathbf{w}^\top[\hat{y}_S,\,\hat{y}_H,\,\hat{y}_X,\,\hat{y}_R], \qquad (3)$$

with non-negative weights ($w_i \geq 0$). The regularisation strength α is selected from 40 logarithmically spaced values in $[10^{-2},\,10^{3}]$ using 5-fold out-of-fold RMSE on the development set. The meta-learner is then refit on the full development set and applied to the test set. The `XLM-R Large` reference systems are not included in this ensemble and are reported separately. The full training protocol, including the two-phase train+dev retraining scheme and the grid search ranges shared with the Solo and Hybrid models, is described in Section 6.1.

7 Results

Table 6 reports our official Closed Track submissions. All submitted runs used `XLM-R Large` and

Lang.	Run	Pos.	RMSE	Pear.	ΔRMSE
Spanish – Baseline RMSE = 1.257 | r = 0.765					
ES	Solo (tuned)	33/59	**1.182**	0.820	**+0.075**
ES	Hybrid (tuned)	34/59	1.186	**0.825**	+0.071
ES	Solo (base)	35/59	1.190	0.814	+0.067
German – Baseline RMSE = 1.258 | r = 0.773					
DE	Hybrid (tuned)	26/56	**1.117**	**0.830**	**+0.141**
DE	Solo (base)	35/56	1.140	0.829	+0.118
DE	Solo (tuned)	40/56	1.177	0.822	+0.081
Chinese – Baseline RMSE = 1.140 | r = 0.753					
CN	Hybrid (tuned)	26/51	**1.006**	**0.854**	**+0.134**
CN	Solo (tuned)	28/51	1.008	0.850	+0.132
CN	Solo (base)	29/51	1.013	0.842	+0.127

Table 6: Official Closed Track test results. Positive ΔRMSE indicates improvement over the baseline; bold marks the best value per L1 group.

improved over the official baseline for the three L1 groups. The largest gains were obtained for German and Chinese, where the Hybrid model achieved the best official RMSE. For Spanish, the Solo model was slightly stronger.

Table 7 compares the official systems with the post-submission refinements and includes development-set RMSE. The `XLM-R Large` Hybrid model was the strongest single model on development, but the `mDeBERTa` stacking ensemble achieved the lowest RMSE on both development and test. On the test set, the ensemble obtained 1.037 for Spanish, 0.997 for German, and 0.913 for Chinese, reducing the mean RMSE from 1.103 for the best official `XLM-R Large` system to 0.982.

Feature-only models did not outperform the baseline, despite the correlations reported in Table 4. This suggests that features such as `clue_ratio`, `word_frequency`, `source_word_length`, and `primary_stress_pos` capture useful but incomplete signal. Transformer models provided stronger contextual representations, while prediction-level stacking benefited from combining transformer and feature-based signals.

Direct feature fusion helped `XLM-R Large` but not `mDeBERTa`. The `XLM-R Large` Hybrid improved over Solo in mean test RMSE (1.103 vs. 1.122), whereas the `mDeBERTa` Hybrid was weaker than `mDeBERTa` Solo (1.150 vs. 1.110). One possible explanation is that the 33-dimensional feature vector is a relatively small addition to the 1024-dimensional `XLM-R Large` representation, but a more influential input to the smaller `mDeBERTa` head. The stacking ensemble avoids this issue by combining feature-based and transformer-based predictions only at the output level.

Figure 3 visualizes the same pattern across Spanish, German, and Mandarin Chinese. The final `mDeBERTa` Ensemble is consistently the lowest-error system, supporting the conclusion that prediction-level ensembling is more robust than relying only on direct feature fusion or a single model family.

8 Conclusions

We presented a system for the BEA 2026 Shared Task on vocabulary difficulty prediction, structured around an official Closed Track submission and a post-submission refinement. The official submissions used `XLM-R Large` Solo and Hybrid models and improved over the baseline for Spanish, German, and Mandarin Chinese. A post-submission refinement based on `mDeBERTa-v3-base`, feature-only regressors, and Ridge stacking further reduced error, achieving the best RMSE among our systems for all L1s: 1.037 (ES), 0.997 (DE), and 0.913 (CN), with a mean RMSE of 0.982.

The results show that engineered features contributed useful linguistic signal, although they were most effective when combined with transformer-based predictions rather than used as standalone models or through direct Hybrid concatenation. Although the final ensemble was not part of the official submission, its estimated ranks suggest a substantial improvement over our official systems, reaching approximately 8/60 for Spanish, 11–12/57 for German, and 15/52 for Chinese. Overall, L1-aware vocabulary difficulty prediction benefits most from combining complementary model signals at the prediction level rather than relying on a single model family or feature fusion strategy alone.

Limitations

This study is limited to the Extended KVL dataset and three L1 backgrounds (Spanish, German, and Mandarin Chinese), so generalization to other learner groups and vocabulary benchmarks remains future work. The hyperparameter search was constrained by available GPU time, and the ensemble was not evaluated under fully nested cross-validation. Finally, our analysis focuses mainly on RMSE and Pearson's r; future work should include finer-grained error analysis by word type, POS, clue structure, and L1-specific properties.

System	Submission	Dev RMSE			Test RMSE			Mean	Rank (ES/DE/CN)
		ES	DE	CN	ES	DE	CN		
XLM-R Base Baseline	–	1.357	1.328	1.175	1.257	1.258	1.140	1.218	–
XGBoost (Optuna)	Post-sub.	1.471	1.427	1.285	1.461	1.351	1.279	1.364	–
Ridge (Optuna)	Post-sub.	1.540	1.477	1.354	1.505	1.407	1.293	1.402	–
`XLM-R Large` Solo	Official	1.139	1.126	1.022	1.182	1.177	1.008	1.122	33/40/28
`XLM-R Large` Hybrid	Official	1.118	1.101	1.019	1.186	1.117	1.006	1.103	34/26/26
`mDeBERTa` Solo	Post-sub.	1.153	1.172	1.057	1.152	1.141	1.037	1.110	–
`mDeBERTa` Hybrid	Post-sub.	1.181	1.173	1.054	1.180	1.157	1.112	1.150	–
`mDeBERTa` Ensemble	Post-sub.	**1.068**	**1.065**	**0.981**	**1.037**	**0.997**	**0.913**	**0.982**	≈8/11–12/15

Table 7: Development and test RMSE for all evaluated systems. Bold marks the best score in each column; ensemble ranks are estimated, not official.

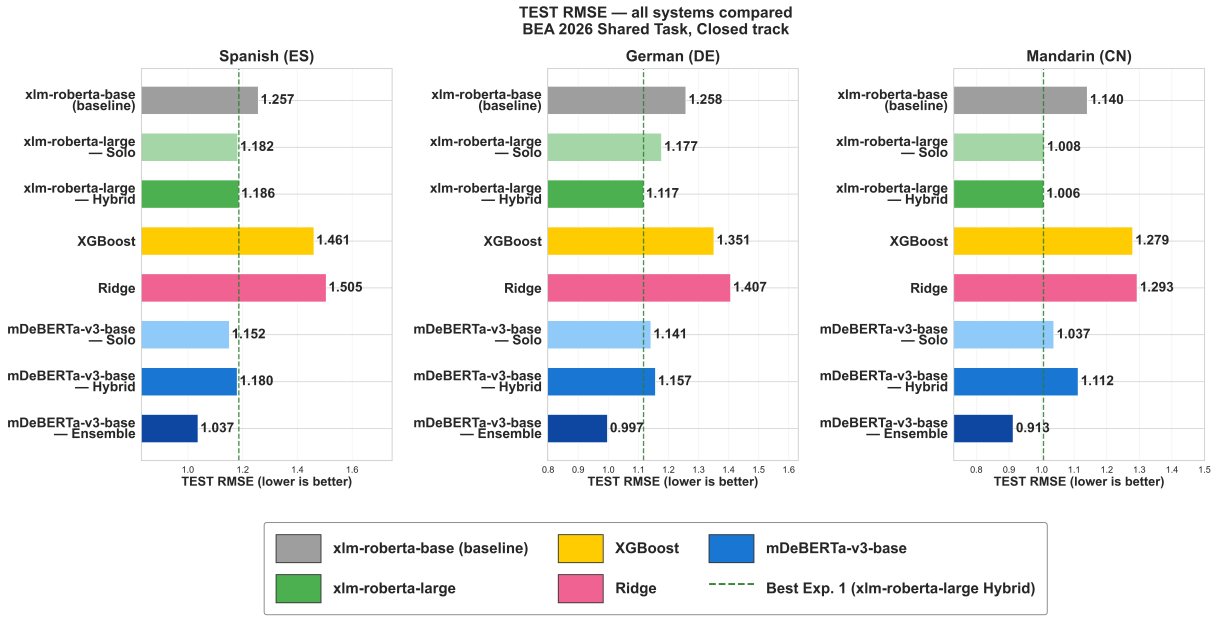

Figure 3: Test RMSE per system and L1 group. Lower is better; the dashed line marks the official `XLM-R Large` Hybrid reference.

A Code and Reproducibility

Code, notebooks, feature pipelines, model configurations, figures, prediction files, and output tables are available at: https://github.com/AdrianPinedaSanchez/bea2026_shared_task_DataAsgardians

References

Takuya Akiba, Shotaro Sano, Toshihiro Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2623–2631. ACM.

R. Harald Baayen, Douglas J. Davidson, and Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412.

Carnegie Mellon University. 2014. The CMU pronouncing dictionary, v0.7b.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.

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. Association for Computational Linguistics.

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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *Proceedings of the 11th International Conference on Learning Representations*.

Batia Laufer and Zahava Goldstein. 2004. Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, 54(3):469–523.

George A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Steven Moran and Daniel McCloy. 2019. PHOIBLE 2.0.

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. SemEval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16. Association for Computational Linguistics.

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.

David H. Wolpert. 1992. Stacked generalization. *Neural Networks*, 5(2):241–259.