AIDA at BEA 2026 Shared Task 1: A Two-Stage Framework for L1-Aware Vocabulary Difficulty Prediction with Representation Diversity and Residual Calibration

Seok Hyeon Cho, JunHyeok Choi, Sangeun Ji, and Sung Won Han*
AIDA Lab, Department of Industrial Management Engineering, Korea University
{seokces315, urisem, jse1769, swhan}@korea.ac.kr

Abstract

We study vocabulary difficulty prediction for second language (L2) learners, a key component for adaptive language learning and assessment. Existing approaches often treat difficulty as an intrinsic property of words or contexts, overlooking representation-dependent variation and learner-specific factors such as L1 transfer.

We participate in the BEA 2026 Shared Task Closed Track using the Spanish (L1) subset of the KVL dataset. We propose a two-stage framework that decouples representation learning from learner-aware calibration. Stage 1 constructs diverse representations using multiple pretrained encoders with varied pooling and prediction strategies, capturing complementary aspects of lexical and contextual complexity. Stage 2 models systematic residual errors with psycholinguistic and cross-lingual features, enabling explicit correction of prediction biases.

Experiments show that our method outperforms strong baselines, improving RMSE (1.257 $\rightarrow$ 0.976) and correlation (0.765 $\rightarrow$ 0.857). These results highlight the importance of jointly modeling representation diversity and learner-specific effects. Our system ranked 3rd in the official BEA 2026 Shared Task Closed Track.

1 Introduction

Vocabulary is central to language proficiency, shaping what learners can understand and produce. Accurately estimating vocabulary difficulty is essential for level-appropriate content, adaptive assessment, and personalized learning, yet traditional approaches still rely on costly expert judgment and large-scale pretesting.

Recent work leverages pretrained language models and regression methods, but typically treats difficulty as an intrinsic property of words or contexts, overlooking learner-specific variation. In particular, prior approaches do not explicitly model how difficulty varies with the learner's first language (L1), despite strong evidence that cross-lingual effects—such as cognates and false friends—systematically influence acquisition.

We revisit vocabulary difficulty prediction from a representation perspective, arguing that difficulty is inherently representation-dependent. Different encoder architectures and aggregation strategies capture complementary aspects of linguistic complexity, making single-model approaches often insufficient in practice.

To address this, we propose a two-stage framework that decouples representation learning from learner-aware calibration via structured residual modeling. Stage 1 constructs representations using multiple pretrained backbones with diverse pooling and prediction heads, aggregating out-of-fold predictions to capture complementary signals. Stage 2 models systematic residuals using psycholinguistic and cross-lingual features, capturing structured biases such as first language (L1)–second language (L2) alignment and exposure differences. This design improves stability and generalization while preserving reliability.

We evaluate our approach on the BEA 2026 Shared Task (KVL dataset), achieving top-tier performance with consistent improvements in RMSE and correlation. Our code and trained models are available online.[1]

Our contributions are:

- We show that vocabulary difficulty is inherently representation-dependent, motivating the use of diverse representations.
- We propose a two-stage framework with residual calibration for L1-aware prediction.
- We demonstrate that combining representation diversity with residual calibration improves L1-aware vocabulary difficulty prediction.

*Corresponding author.

[1] https://github.com/seokces315/ACL_BEA2026_2SF

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 1029–1040

2 Related Work

2.1 Lexical Complexity Prediction

Vocabulary difficulty prediction is closely related to lexical complexity prediction (LCP), which extends complex word identification from classification to regression. Early approaches relied on surface features such as frequency and length (Kuperman et al., 2012), while recent work adopts transformer-based contextual representations (Gooding et al., 2020; Shardlow et al., 2021).

Recent studies demonstrate strong performance with pretrained multilingual encoders in cross-lingual settings (Bani Yaseen et al., 2021; Kelious et al., 2024). However, these approaches typically treat complexity as an intrinsic property of words or contexts, largely ignoring learner-specific variation. In particular, they do not explicitly model how difficulty varies with the learner's linguistic background, which limits their applicability to L2 settings. Our work addresses this limitation by combining learner-aware representations with feature-based calibration for difficulty prediction.

2.2 L1 Transfer and Cross-lingual Effects

L1 transfer plays a central role in second language acquisition, affecting both facilitation and interference (Ellis, 1997). Cognates can ease processing through cross-lingual similarity, whereas false friends introduce systematic errors due to form–meaning mismatches (Van Assche et al., 2009; Schepens et al., 2012).

While cross-lingual effects are widely studied, their integration into lexical difficulty prediction remains limited. Existing approaches often rely on surface similarity (e.g., edit distance), which fails to capture semantic divergence or exposure differences across languages. As a result, they are prone to systematic biases, particularly in cases such as false friends. In contrast, our work goes beyond surface similarity by leveraging cross-lingual and linguistic signals to capture such divergences.

3 Task and Dataset

3.1 Task Definition

The BEA 2026 Shared Task addresses vocabulary difficulty prediction for English learners with diverse L1 backgrounds. The task is formulated as a regression problem, where the goal is to predict a continuous difficulty score for a given English word conditioned on the learner's L1.

Each instance includes an English word and its associated L1 information (e.g., translation or context), and the target is a psychometrically calibrated difficulty score. We use the Knowledge-based Vocabulary Lists (KVL) dataset (Schmitt et al., 2021, 2024), a multilingual resource annotated with difficulty scores.

3.2 Task Setup

We follow the Closed Track setting, where models are trained only on the provided data for each L1, without using external training data or combining different L1 datasets. Pre-trained Transformer models and standard NLP tools are allowed, while LLMs are prohibited. Public linguistic resources (e.g., WordNet) may be used for feature extraction.

In this work, we focus on the Spanish (L1) subset of the dataset. All models are trained and evaluated exclusively on Spanish learner data, in accordance with the Closed Track constraints.

3.3 Dataset

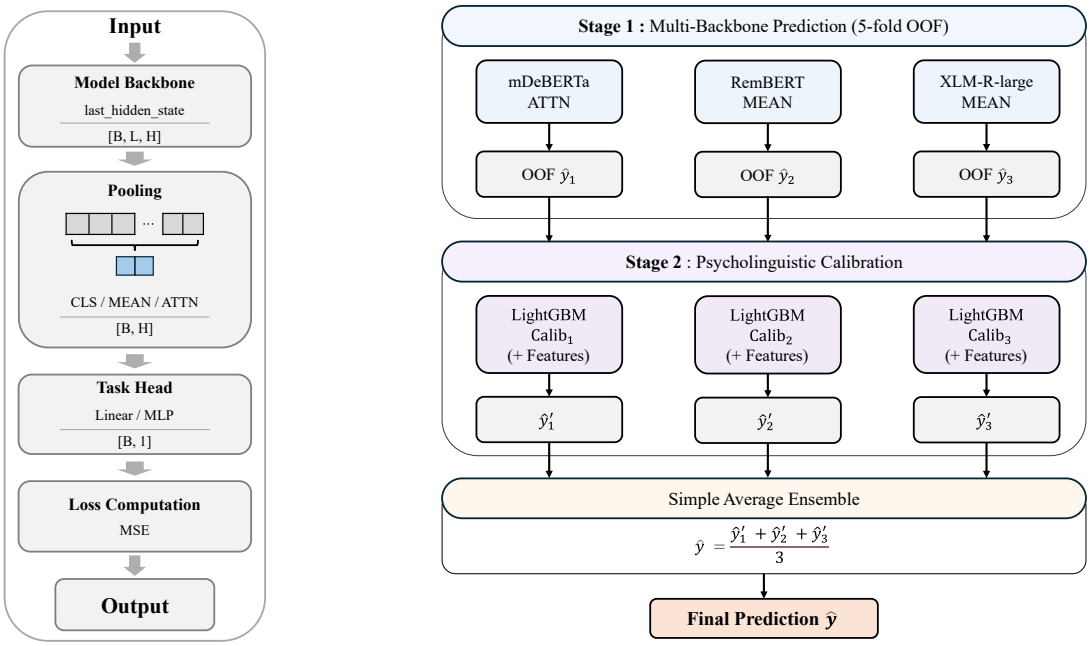
We use an extended version of the British Council's Knowledge-based Vocabulary Lists (KVL) dataset (Skidmore et al., 2025) to model vocabulary difficulty for English learners with diverse L1 backgrounds. We use only the Spanish subset in our experiments.

The dataset is constructed from large-scale assessments of productive vocabulary knowledge, where learners recall English target words from L1 prompts in a translation-based format (Laufer and Goldstein, 2004). Each item consists of an L1 context sentence, an L1 cue word, and a partially masked English target (e.g., h____ for "house").

Difficulty scores are estimated from approximately 3.3 million learner responses using random-item-random-person (RPRI) Rasch models within a generalized linear mixed model (GLMM) framework. The resulting GLMM score (logit scale) is used as the prediction target, where higher values indicate greater difficulty for learners.

The dataset is provided per L1 and split into train/dev/test sets (6,091 / 677 / 748 items). Each instance includes the target word, part-of-speech, a partial spelling clue, the L1 word and context, and a difficulty score. All data are released under CC BY-NC 4.0, with official baselines and evaluation scripts.[2]

[2] https://github.com/britishcouncil/bea2026st



(a) Stage 1: Single-backbone pipeline

(b) Two-stage framework

Figure 1: Overview of the proposed two-stage framework. (a) Stage 1 uses a single-backbone pipeline (encoder, pooling, and head) with multiple PLMs to produce diverse out-of-fold (OOF) predictions. (b) Stage 2 refines these via residual calibration with psycholinguistic features, followed by simple averaging ensemble.

4 Method

4.1 Stage 1: Multi-Backbone Representations

In the first stage, we construct a diverse set of sequence representation models based on multiple pretrained language model (PLM) backbones and pooling strategies. The goal of this stage is to obtain robust and complementary predictions by leveraging representation diversity across different model architectures and aggregation mechanisms. Since the task requires both lexical cues and contextual understanding, we adopt heterogeneous backbones to capture multi-level signals. This is particularly important for our task, where difficulty prediction requires capturing both surface-level patterns and deeper semantic features that individual models fail to consistently handle.

PLM Backbones. We utilize a diverse set of multilingual encoder-based pretrained language models (PLMs), including XLM-R (Conneau et al., 2020), RemBERT (Chung et al., 2021), mDeBERTa (He et al., 2023), mmBERT (Marone et al., 2025), and mBERT (Devlin et al., 2019), and select a subset based on validation performance and prediction diversity. We observe that their outputs vary for the same input, indicating complementary behavior.

Pooling Strategies. Given an input sequence, each backbone model produces token-level contextual embeddings. Since downstream prediction requires a fixed-length representation, we apply pooling over the final hidden states to construct a sequence-level representation. Specifically, we consider three pooling strategies: the [CLS] token representation, mean pooling, and attention pooling (Lin et al., 2017). We explore various combinations of pooling methods and prediction heads, and select a subset based on validation performance. This design enables the model to capture both global context and salient token-level signals.

Selected Model Variants. We retain three representative models, each with a distinct pooling and head configuration, and use them as base predictors in the subsequent ensemble stage (Figure 1).

Training Protocol. To ensure robust and unbiased predictions, we adopt 5-fold cross-validation (CV) for each model variant. Models are trained on different splits, and out-of-fold (OOF) predictions are generated for all training instances. These OOF predictions ensure each instance is represented by predictions from models not trained on it, preventing data leakage and reducing overfitting bias in Stage 2. The resulting OOF outputs from all base models serve as inputs to the ensemble stage.

4.2 Stage 2: Psycholinguistic Calibration

While Stage 1 captures contextual and semantic representations, it underrepresents linguistically grounded signals (e.g., frequency, morphology, and L1-driven orthographic patterns) related to word difficulty. To address this limitation, Stage 2 leverages handcrafted features and trains a LightGBM regressor (Ke et al., 2017) to predict the residual $(y - \hat{y}_1)$, where $\hat{y}_1$ is the Stage 1 prediction. This residual formulation refines predictions without overriding accurate estimates, yielding more stable improvements and lower variance across folds.

Rather than directly predicting the target value, this residual formulation focuses on correcting errors from Stage 1, allowing the model to refine predictions without overriding already accurate estimates. This leads to more stable performance improvements and empirically reduced variance across folds compared to direct regression.

Feature Engineering. We derive features from four linguistically motivated groups, with cross-lingual features capturing L1-positive transfer by approximating how Spanish speakers transform English words orthographically. Specifically, we apply rule-based modifications to the English target, such as *-tion* $\rightarrow$ *-cion* and *ph* $\rightarrow$ *f*, including suffix, phonetic, and orthographic adjustments. We then compute the reduction in Levenshtein distance to the Spanish equivalent before and after transformation, capturing the degree of morphological predictability under L1 influence. We further incorporate a cognate similarity score based on positional character n-gram overlap to capture systematic cross-lingual similarity.

Psycholinguistic features primarily include the log-frequency gap between L2 words and their L1 equivalents and concreteness scores, reflecting cognitive processing difficulty. Lexical and morphological features provide complementary signals, including word frequency, length, and morpheme structure. Full details of cross-lingual transformations are provided in Appendix A.1.

Optimization Strategy. To identify an effective feature configuration, we perform a group ablation study over 119 valid combinations of feature subgroups, evaluated using out-of-fold (OOF) RMSE. We consistently observe that cross-lingual features are included in all competitive configurations, highlighting their central role.

Based on this analysis, we select a subset combining lexical, morphological, cross-lingual, and psycholinguistic features. Although this configuration does not achieve the best performance under default settings, it shows stronger gains after tuning and becomes optimal. This suggests that configurations with psycholinguistic features are more responsive to tuning.

We perform hyperparameter optimization with Optuna, tuning key parameters such as tree complexity and regularization. This achieves the best post-tuning performance across candidate groups.

We also explore an end-to-end alternative that directly integrates handcrafted features into Stage 1, but this approach does not yield consistent or substantial improvements over the standalone neural model. This suggests a mismatch between PLM representations and handcrafted feature distributions, making direct integration less effective. In contrast, the proposed two-stage design separates representation learning from calibration, allowing Stage 2 to model non-linear feature interactions and correct errors more effectively.

Post-processing. We apply a lightweight post-processing pipeline with an OOF-gated selection strategy. Candidate transformations (e.g., clipped blending, isotonic regression, and tail calibration) are sequentially and systematically evaluated and retained only if they strictly reduce OOF RMSE, ensuring consistent gains without overfitting. In practice, only isotonic regression is selected, yielding additional improvements.

Full implementation details, ablation results, and optimization are provided in the Appendix.

5 Experimental Setup

5.1 Data Usage

We use the official train/dev/test splits with model selection on the development set. Each instance is formatted as a natural language prompt combining the L1 context, source word, English target, auxiliary clues, and part-of-speech, as shown in Table 9 in Appendix A.4, and fed into the PLM.

5.2 Evaluation Metrics

We evaluate performance using RMSE (Root Mean Squared Error) between predicted and gold GLMM scores. RMSE is the primary metric for model comparison. We also report the Pearson correlation coefficient to measure the linear relationship between predictions and ground-truth scores.

5.3 Baselines and Ablation Models

We compare our method with external baselines and internal ablation models.

As baselines, we consider (1) the official baseline provided by the shared task organizers and (2) a feature-based Ridge regression model using handcrafted features, including cross-lingual exposure, contextual average word length, frequency-based statistics (e.g., Zipf score, frequency rank, L1/L2 frequency), and concreteness.

We also include ablation models corresponding to the first stage of our framework. Stage 1 (Single) denotes single-backbone models with different pooling and head configurations, each evaluated independently; performance is reported as the average. Stage 1 (Ensemble) aggregates predictions from the top three encoder backbones, each paired with its best pooling and head configuration.

These variants analyze the contribution of model diversity and the effectiveness of the two-stage design. Final results are presented in Section 6.1.

6 Results

6.1 Main Results

We evaluated the proposed approach against the official baseline and progressively stronger model variants to assess the contribution of each component. Table 1 summarizes the overall performance.

Model	Dev RMSE ↓	Test	
		RMSE ↓	r ↑
Baselines			
Official Baseline	–	1.257	0.765
Feature-based (Ridge)	1.426	1.429	0.652
Ablations			
Stage 1 (Single)	1.040	1.053	0.838
Stage 1 (Ensemble)	0.975	0.989	0.851
Two-Stage (Ours)	**0.958**	**0.976**	**0.857**

Table 1: Performance comparison across baselines and ablation models. RMSE is reported for the development set, and both RMSE and Pearson correlation (r) are reported for the test set.

Our method clearly outperformed all baselines. Compared to the official baseline, the proposed two-stage model significantly reduced RMSE (1.257 → 0.976) while improving Pearson correlation (0.765 → 0.857), indicating better alignment with ground-truth difficulty scores.

A large performance gap between the baselines (both official and feature-based) and Stage 1 (Single) highlighted the importance of task-specific tuning of pretrained encoders with appropriate pooling strategies and prediction heads.

We observed consistent gains across modeling stages. Moving from Stage 1 (Single) to Ensemble improved performance, highlighting representation diversity, while the two-stage model further enhanced accuracy and stability through refinement.

Overall, the results suggested that the primary gains stemmed from effective representation learning and diversity, while additional improvements were achieved through explicit refinement in the second stage. Consistent with these findings, our approach ranked 3rd on the official shared task leaderboard.

6.2 Representation Analysis and Ensembling

To analyze representation design, we evaluated multiple encoder families (BERT, mDeBERTa, RemBERT, and XLM-R) across pooling and prediction head configurations. For each encoder, we report both the average RMSE across all configurations (as a robustness proxy) and the best RMSE.

As shown in Table 2, XLM-R (large) achieved the best performance (1.029 best, 1.038 avg), followed by RemBERT and mDeBERTa-v3.

Encoder	Dev RMSE ↓	
	Avg	Best
XLM-R (large)	**1.038**	**1.029**
RemBERT	1.045	1.036
mDeBERTa-v3	1.057	1.047
mmBERT	1.072	1.062
XLM-R (base)	1.193	1.185
mBERT (cased)	1.267	1.254

Table 2: Encoder comparison across pooling and head configurations. We report mean RMSE (Avg; reflecting robustness across configurations) and best RMSE for each encoder, highlighting encoder-dependent performance variation.

Despite the general advantage of larger encoders, representation design introduced substantial variation: within a single encoder, performance differences across configurations reached 0.02–0.04 RMSE, comparable to or exceeding inter-encoder gaps (e.g., 0.007 between XLM-R (large) and RemBERT). This indicated that encoder choice alone was insufficient, and pooling/head design played an equally critical role.

Table 3 presents representative top configurations for each encoder, suggesting encoder-specific

preferences. For instance, XLM-R (large) worked best with mean pooling and a linear head, whereas mDeBERTa-v3 favored attention-based pooling.

Encoder	Pooling + Head (Dev RMSE ↓)		
	MEAN+Lin	ATTN+Lin	ATTN+MLP
XLM-R (large)	**1.029**	1.031	1.041
RemBERT	**1.036**	1.048	1.042
mDeBERTa-v3	1.050	**1.047**	1.063

Table 3: Representative top-performing pooling and head configurations for each encoder. The results indicate encoder-dependent preferences over design choices, highlighting systematic differences in preferred strategies across models.

Moreover, configuration rankings varied across encoders, suggesting that models captured different aspects of the task in practice. A single unified design would therefore underutilize encoder-specific strengths. Instead, we adopted a model-wise optimal strategy, pairing each encoder with its best-performing configuration.

Our final ensemble integrated the top three encoders—XLM-R, RemBERT, and mDeBERTa-v3—each paired with its optimal pooling and head configuration. This design leveraged diverse representations across models, consistently leading to improved overall performance.

Overall, our results highlighted that performance gains stemmed not only from stronger pretrained backbones, but also from effectively leveraging representation diversity inherently induced by pooling and head design choices.

6.3 Calibration and Residual Correction

We constructed the Stage 1 ensemble by averaging three models—XLM-R (large), RemBERT, and mDeBERTa-v3—each with its optimal pooling and prediction head. Although effective, the ensemble still exhibited structured residual errors. To address this, we applied calibration in the prediction space (Guo et al., 2017). Table 4 summarizes the results.

The uncalibrated ensemble achieved 1.028 OOF RMSE and 0.995 Dev RMSE. Calibration consistently improved performance. Under Avg-then-Calib, RMSE decreased to 0.999 (LightGBM), indicating that residual errors remained learnable after ensembling.

Calib-then-Avg yielded larger gains, outperforming Avg-then-Calib across all calibrators. The best performance was achieved by LightGBM, reaching 0.976 OOF RMSE and 0.958 Dev RMSE.

Strategy	Calibrator	RMSE ↓	
		OOF	Dev
No calibration	–	1.028	0.995
Avg-then-Calib	Ridge	1.017	0.984
	XGBoost	1.011	0.974
	LightGBM	0.999	0.977
Calib-then-Avg	Ridge	1.013	0.978
	XGBoost	0.995	0.968
	LightGBM	**0.976**	**0.958**

Table 4: Comparison of calibration strategies for the Stage 1 ensemble. Avg-then-Calib denotes averaging followed by calibration, while Calib-then-Avg denotes calibrating before averaging. OOF RMSE is computed on out-of-fold predictions.

These results suggested that calibrating individual models before ensembling was more effective, correcting model-specific residual patterns prior to aggregation. Additionally, tree-based calibrators (LightGBM, XGBoost) outperformed Ridge, highlighting the importance of modeling nonlinear residuals. Based on this, we adopted Calib-then-Avg with LightGBM as the final approach.

6.4 Performance on Tail Distributions

Modeling extreme cases was a key challenge in continuous difficulty estimation. To analyze this, we evaluated performance across difficulty tiers, treating Easy and Hard as tails and Mid as center.

Table 5 reports RMSE for Stage 1 (S1) and Stage 2 (S2), along with relative changes, further stratified by cognate status. Difficulty groups are defined as Easy ($y > 2.5$), Mid ($-2.0 < y \leq 2.5$), and Hard ($y \leq -2.0$), and N denotes the number of samples in each subset. All rows summarize group-level performance.

Difficulty	Cognate	N	Dev RMSE ↓		Δ (%)
			S1	S2	
Easy	**All**	621	1.019	1.047	**-2.75**
	Cog.	170	0.985	0.945	+4.06
	Non-Cog.	451	1.031	1.083	-5.04
Mid	**All**	4537	0.881	0.847	**+3.86**
	Cog.	626	0.920	0.872	+5.22
	Non-Cog.	3911	0.874	0.843	+3.55
Hard	**All**	933	1.450	1.369	**+5.59**
	Cog.	50	1.833	1.794	+2.13
	Non-Cog.	883	1.425	1.341	+5.89

Table 5: Performance across difficulty groups and lexical similarity subsets. S1 and S2 denote Stage 1 and Stage 2 models, respectively. Δ (%) indicates RMSE reduction from S1 to S2 (positive = better).

Results showed a clear asymmetry across regions. In the Mid tier, S2 consistently improved performance (All: $+3.86\%$, Cognates: $+5.22\%$, Non-cognates: $+3.55\%$), indicating effective residual correction in stable regions where feature signals were well-aligned.

In contrast, performance degraded in the Easy tail (All: -2.75%), mainly due to non-cognates (-5.04%), despite cognate gains ($+4.06\%$). This suggested over-correction in low-variance regions where predictions were already reliable, likely due to over-reliance on similarity-based features.

At the Hard tail, S2 yielded the largest improvements (All: $+5.59\%$, Cognates: $+2.13\%$, Non-cognates: $+5.89\%$), indicating that residual modeling effectively reduced errors in high-uncertainty cases, including challenging *false-friend* scenarios where surface similarity was misleading.

Overall, calibration yielded the greatest improvements in high-uncertainty regions, where predictions were less reliable, but might lead to overcorrection in low-variance regions, where predictions were already stable, thereby underscoring the need for difficulty-aware calibration.

7 Discussion

7.1 Two-Stage Modeling

A central finding was that the proposed framework constituted a structural decomposition between representation and calibration. Rather than attempting to jointly model all factors in a single space, the two-stage design separated robust difficulty representation from systematic error correction.

A key insight underlying Stage 1 was that difficulty representation was fundamentally model-dependent. Different encoder architectures and aggregation choices captured complementary aspects of linguistic complexity, indicating that no single configuration was universally optimal. As a result, relying on a fixed architecture risked overlooking important signals. Instead, Stage 1 constructed a robust representation layer by integrating strong configurations, enabling the model to capture diverse facets of language-intrinsic difficulty while reducing sensitivity to individual design choices.

This design was motivated by the observation that optimal configurations varied across encoders, highlighting the importance of encoder-specific optimization for stable representation learning.

Stage 2 operated in the residual space, performing learner-dependent and psycholinguistic correction. By incorporating language-specific signals (e.g., cognate effects and cross-lingual similarity patterns) grounded in the Spanish domain, it addressed systematic discrepancies—such as mismatches between surface similarity and actual difficulty—that were not captured by semantic representations alone.

Crucially, calibration in this framework functioned as a structured bias correction mechanism. By decoupling representation from correction, the model preserved the underlying difficulty signal while explicitly modeling residual errors, which led to more stable and interpretable predictions, particularly in challenging cases where single-stage models tended to conflate linguistic similarity with true difficulty.

7.2 Feature Importance Analysis

The tree-based calibration model enabled interpretability via feature importance measured by split frequency. To support reproducibility and clarify the feature design, a complete inventory of handcrafted features and their sources is provided in Table 11.

As shown in Table 6, top features fell into three categories—cross-lingual, lexical, and psycholinguistic. We utilized handcrafted features based on lexical frequency information from Wordfreq (Speer, 2022), psycholinguistic properties such as concreteness (Brysbaert et al., 2014), word complexity, and cross-lingual exposure derived from English–Spanish frequency differences.

Feature	Category	Split Count
Crossling_Exposure	Cross-lingual	180
Ctx_Avg_Word_Len	Psycholinguistic	176
Global_Zipf_Score	Lexical	171
Freq_Rank_Norm	Lexical	160
Freq_L1_L2_Gap	Psycholinguistic	157
L1_Zipf_Score	Psycholinguistic	143
Concreteness	Psycholinguistic	136

Table 6: Top features ranked by importance based on split frequency in the tree-based model. Feature names are abbreviated; detailed descriptions are provided in the text. *Split Count* denotes the number of times a feature is selected for node partitioning across all trees.

Psycholinguistic and exposure-related features dominated: four of the top seven were psycholinguistic, with others reflecting frequency or cross-lingual exposure. Importantly, *Crossling_Exposure* ranked highest, followed by *Global_Zipf_Score* and *Freq_L1_L2_Gap*.

These results suggested that difficulty for Spanish L1 learners was driven less by intrinsic word properties and more by exposure and alignment with prior linguistic experience. The importance of *Freq_L1_L2_Gap* highlighted cross-lingual frequency mismatch. Finally, *Concreteness* supported prior findings that concrete words were easier to acquire (Brysbaert et al., 2014; Paivio, 1991).

7.3 Error Analysis

To analyze model limitations, we examined representative failure cases from two complementary perspectives: systematic errors due to false friends and qualitative examples of model behavior.

A primary source of error arose from *false friends*, where words were orthographically similar to their Spanish counterparts but semantically divergent. In such cases, the model overestimated easiness due to surface-form similarity (e.g., *embarrassed* vs. *embarazada*). This suggested that while cross-lingual similarity was effectively captured, semantic alignment was insufficiently modeled, particularly in semantically divergent cases.

Table 7 shows representative examples of Stage 2 calibration. GLMM denotes ground-truth difficulty, and S1/S2 indicate predictions from Stage 1 and Stage 2, respectively. *Cognates* (↑) were consistently adjusted toward easier difficulty, *Non-cognate* instances remained appropriately difficult without substantial change, and *Failure* highlighted errors arising from missed cognate detection.

English	Spanish	GLMM	S1	S2
Cognates (↑)				
domination	dominación	+1.40	−0.36	**+0.75**
consecutive	consecutivo	+1.19	−0.62	**+0.44**
discipline	disciplina	+2.06	+0.51	**+1.56**
compulsive	compulsivo	+1.79	−0.25	**+0.75**
selective	selectivo	+2.13	−0.07	**+0.92**
Non-cognate				
skillet	sartén	−3.02	−1.08	−2.56
Failure				
synonym	sinónimo	−1.72	+0.83	−0.26

Table 7: Stage 2 calibration examples by word type. GLMM: ground truth; S1/S2: Stage 1/2 predictions.

For transparent cognates (e.g., *domination* / *dominación*), Stage 1 underestimated easiness, whereas Stage 2 corrected predictions using additional features. Non-cognates (e.g., *skillet*) remained appropriately difficult, indicating that calibration was not applied indiscriminately.

However, failure cases (e.g., *synonym* / *sinónimo*) revealed limitations of the cognate detection mechanism. Despite semantic equivalence, orthographic mismatch led to low similarity scores and inaccurate predictions.

Overall, while the model effectively leveraged cross-lingual similarity, it remained vulnerable to semantic divergence and surface-form limitations.

8 Conclusion

In this work, we addressed the problem of vocabulary difficulty prediction for L2 learners by focusing on representation diversity and learner-specific factors. We proposed a two-stage framework that separates representation learning from psycholinguistic calibration, enabling the model to capture both language-intrinsic difficulty and systematic cross-lingual effects.

Experimental results demonstrated that our approach consistently outperformed strong baselines, achieving substantial improvements in both RMSE and correlation. The primary gains stemmed from combining diverse pretrained encoders with optimized pooling and prediction head configurations, allowing the model to capture complementary lexical, contextual, and semantic signals. The calibration stage further improved predictions by modeling structured residual errors using psycholinguistic and cross-lingual features, particularly in high-uncertainty regions where semantic representations alone were insufficient.

Our analysis further suggested that vocabulary difficulty should not be viewed solely as an intrinsic property of words, but as a learner-dependent phenomenon shaped by prior linguistic experience and L1–L2 alignment. More specifically, features related to cross-lingual exposure and frequency mismatch played a central role in explaining learner difficulty for Spanish L1 speakers.

Overall, the proposed framework demonstrates that separating representation learning from residual correction provides a stable and effective approach for L1-aware vocabulary difficulty prediction. Beyond the shared task setting, we believe this framework offers a practical foundation for adaptive language learning and personalized educational applications.

Future work will explore semantics-aware cross-lingual representations and more adaptive calibration strategies to better handle semantically challenging cases across diverse learner settings.

9 Limitations

Despite the strong overall performance, our framework still exhibited several limitations. First, the model remained vulnerable to cases involving semantic divergence, such as false friends, where reliance on surface-level similarity occasionally produced systematic prediction errors. This suggests that the current representation space does not fully capture deeper semantic correspondence across languages. In addition, the calibration stage sometimes introduced over-correction in low-variance regions, indicating that more adaptive calibration strategies may be necessary for stable refinement.

We also note that the exploration of Stage 1 ensemble configurations was not fully exhaustive due to computational limitations and time constraints. Finally, although the proposed framework demonstrated strong effectiveness for Spanish L1 learners, its generalizability to broader learner populations and additional L1 settings remains to be validated.

References

Tuqa Bani Yaseen, Qusai Ismail, Sarah Al-Omari, Eslam Al-Sobh, and Malak Abdullah. 2021. JUST-BLUE at SemEval-2021 task 1: Predicting lexical complexity using BERT and RoBERTa pre-trained language models. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 661–666, Online. Association for Computational Linguistics.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46(3):904–911.

Joanne F. Carlisle. 2000. Awareness of the structure and meaning of morphologically complex words: Impact on reading. *Reading and Writing*, 12(3–4):169–190.

Hyung Won Chung, Thibault Févry Lee, Yingjie Huang, Andrew Dai, and Saliou Adafre. 2021. Rethinking embedding coupling in pre-trained language models. In *Proceedings of the International Conference on Learning Representations*.

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, Online. Association for Computational Linguistics.

David Crystal. 2003. *The Cambridge Encyclopedia of the English Language*, 2 edition. Cambridge University Press, Cambridge.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rod Ellis. 1997. *Second Language Acquisition*. Oxford University Press, Oxford.

Sian Gooding, Shiva Taslimipoor, and Ekaterina Kochmar. 2020. Incorporating multiword expressions in phrase complexity estimation. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 14–19, Marseille, France. European Language Resources Association.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning*. PMLR.

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 International Conference on Learning Representations*.

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. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, volume 30.

Abdelhak Kelious, Mathieu Constant, and Christophe Coeur. 2024. Investigating strategies for lexical complexity prediction in a multilingual setting using generative language models and supervised approaches. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 96–114, Rennes, France. LiU Electronic Press.

Michael J. Kieffer and Nonie K. Lesaux. 2012. Development of morphological awareness and vocabulary knowledge in spanish-speaking language minority learners: A parallel process latent growth curve model. *Applied Psycholinguistics*, 33(1):23–54.

Donald E. Knuth. 1973. *The Art of Computer Programming, Volume 3: Sorting and Searching*. Addison-Wesley, Reading, Massachusetts.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, 44(4):978–990.

Batia Laufer and Zahava Goldstein. 2004. Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, 54(3):399–436.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Zhouhan Lin, Minwei Feng, Cicero Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *Proceedings of the International Conference on Learning Representations*.

M. Marone, O. Weller, W. Fleshman, E. Yang, D. Lawrie, and Benjamin Van Durme. 2025. mmBERT: A modern multilingual encoder with annealed language learning. *arXiv preprint arXiv:2509.06888*.

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

Allan Paivio. 1991. Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology*, 45(3):255–287.

Job Schepens, Ton Dijkstra, and Frank Grootjen. 2012. Distributions of cognates in europe as based on levenshtein distance. *Bilingualism: Language and Cognition*, 15(1):157–166.

Norbert Schmitt, Karen Dunn, Barry O'Sullivan, Laurence Anthony, and Benjamin Kremmel. 2021. Introducing knowledge-based vocabulary lists (kvl). *TESOL Journal*, 12(4):e622.

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

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, Online. 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.

Robyn Speer. 2022. wordfreq: Frequency data for natural language processing.

Eva Van Assche, Wouter Duyck, and Robert J. Hartsuiker. 2009. The cognitive mechanisms of bilingual reading: The case of cognate facilitation. *Journal of Memory and Language*, 60(1):92–107.

William E. Winkler. 1990. String comparator metrics and enhanced decision rules in the Fellegi–Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pages 354–359.

A Appendix

A.1 Cross-lingual Feature Details

A.1.1 Hispanized Transformation Rules

To approximate the orthographic transformation process of Spanish learners, we apply a sequence of rule-based modifications to English target words. The rules are grouped as follows:

Suffix Mapping. We map common English suffixes to their Spanish counterparts:

- *-tion* → *-cion*
- *-ty* → *-dad*
- *-ly* → *-mente*
- *-ous* → *-oso*
- *-ence / -ance* → *-encia / -ancia*

Phonetic Mapping. We simulate phonological adaptation:

- *ph* → *f*
- *th* → *z* (Peninsular) or *s* (Latin American seseo)

Prothetic *e*. We prepend *e* to words beginning with an *s*+consonant cluster:

- *special* → *especial*

Consonant Simplification. We reduce double consonants except *ll* and *rr*.

A.1.2 Distance-Based Feature

Let d_{orig} denote the Levenshtein distance between the original English word and its Spanish translation, and d_{trans} denote the distance after applying the transformation rules. We define the cross-lingual predictability feature as:

$$\Delta d = d_{\text{orig}} - d_{\text{trans}} \tag{1}$$

A larger Δd indicates that the transformed form is closer to the Spanish equivalent, reflecting higher morphological predictability under L1 influence.

A.1.3 Cognate Similarity

We compute a positional character n-gram overlap score to capture cognate similarity. The score is defined as a weighted sum of matching character sequences:

- First-character match: +0.3

- First three-character match: +0.4
- Last three-character match: +0.3

Words with a score above 0.4 are considered cognates. We additionally include a binary feature indicating systematic cognate correspondence.

A.2 Feature Group Ablation

To analyze the contribution of each feature group, we perform a group ablation study over combinations of feature subgroups and evaluate them using out-of-fold (OOF) RMSE. Table 8 reports the top-performing configurations.

Feature Groups	OOF RMSE ↓
lex + cross + derived	1.0506
lex + morph + cross + derived	1.0508
lex + morph + cross	1.0510
lex + cross	1.0511
lex + morph + cross + psycho	**1.0513**

Table 8: Top-5 feature group combinations from the ablation study. Cross-lingual features consistently appear in all competitive configurations. The selected setting (lex + morph + cross + psycho) is used for Stage 2.

We consistently observe that cross-lingual features are present across all high-performing configurations, confirming their central role in modeling word difficulty.

A.3 Optimization and Postprocessing Details

A.3.1 Hyperparameter Optimization

We conduct hyperparameter optimization using Optuna to tune a set of key LightGBM parameters such as n_estimators, learning_rate, num_leaves, min_child_samples, subsample, colsample_bytree, and L1/L2 regularization.

The selected configuration achieves the best post-tuning performance across all candidate feature groups, with an OOF RMSE of 1.040.

A.3.2 Post-processing Details

We apply a sequence of candidate post-processing steps to further calibrate model predictions. Each step is evaluated using out-of-fold (OOF) RMSE and is only retained if it strictly improves performance.

Candidate Steps. The post-processing pipeline includes clipped blending, isotonic regression, and tail calibration.

Selection Outcome In practice, only isotonic regression satisfies the OOF gating criterion, resulting in an additional reduction of approximately 0.010 in OOF RMSE.

A.4 Prompt Construction

We construct the input as a structured natural language prompt by mapping dataset fields to semantically meaningful components. The *Spanish context* provides the L1 sentence containing the target word, while the *Spanish word* denotes the aligned lexical item. The *English translation* specifies the target whose difficulty is to be predicted. The *Meaning clue* provides auxiliary hints, with part-of-speech information appended in parentheses. The prediction target is the GLMM score, which reflects the perceived difficulty of the English word for L1 speakers. Table 9 shows the mapping from each component to its textual realization.

Component	Example
Spanish context	`El eclipse solar fue visible durante un breve lapso de tiempo.`
Spanish word	lapso
English translation	span
Meaning clue	`s___ (noun)`

Table 9: Example of the structured prompt, showing the mapping from each component to its textual realization.

B Training Hyperparameters

Setting	Value
Shared settings	
Optimizer	AdamW
Learning rate	2×10^{-5}
Batch size	16
Gradient accumulation steps	2
Warmup ratio	0.1
Weight decay	0.01
Max sequence length	256
Precision	FP16
LightGBM calibrator (Optuna-tuned)	
n_estimators	400
learning_rate	0.03
num_leaves	31
min_child_samples	20
subsample	0.8
colsample_bytree	0.8
reg_alpha (L1)	0.1
reg_lambda (L2)	1.0

Table 10: Training hyperparameters for all components of the proposed pipeline.

C Feature Inventory for Stage 2 Calibration

To support reproducibility and clarify the feature design, Table 11 provides a complete inventory of handcrafted features used in the Stage 2 calibration model, including feature descriptions, data sources, and references for externally derived resources.

Feature	Description	Source
Lexical Features (9)		
Global_Zipf_Score	English Zipf frequency	Speer (2022)
Is_Low_Zipf	Low-Zipf flag (< 3.5)	Speer (2022)
Freq_Rank_Norm	Rank-normalised dataset frequency	Dataset
Dataset_Freq_Raw	Raw target frequency	Dataset
Char_Count	Character count	Dataset
Vowel_Ratio	Vowel-to-character ratio	Orthographic
Consonant_Cluster_Max	Maximum consonant cluster length	Orthographic
Has_Double_Letter	Repeated-letter flag	Orthographic
Letter_Rarity	Proportion of rare letters	Orthographic
Morphological Features (8)		
Syllable_Count	Estimated syllable count	Crystal (2003)
Morpheme_Count_Est	Estimated morpheme count	Carlisle (2000)
Suffix_Count	Number of suffix matches	Carlisle (2000)
Prefix_Count	Number of prefix matches	Kieffer and Lesaux (2012)
Morph_Complexity	Prefix/suffix complexity flag	This work
Is_Abstract	Abstract suffix flag	This work
POS_Encoded	Encoded POS category	Dataset
POS_Frequency	POS frequency in training data	Dataset
Cross-lingual Features (18)		
Norm_Lev_Dist	Normalised Levenshtein distance	Levenshtein (1966)
Lev_Dist_Raw	Raw Levenshtein distance	Levenshtein (1966)
Hispanized_Lev_Dist	Hispanized edit distance	This work
Hispanized_Lev_Gain	Hispanization distance reduction	This work
JaroWinkler_Similarity	Jaro–Winkler similarity	Winkler (1990)
Bigram_Overlap	Bigram overlap ratio	String algorithms
Phonetic_Similarity	Soundex-based similarity	Knuth (1973)
LCS_Ratio	Normalised LCS score	Knuth (1973)
LCS_Length	Longest common substring length	Knuth (1973)
Cognate_Score	Heuristic cognate similarity	This work
Systematic_Cognate	EN–ES suffix correspondence	This work
L1_Char_Count	L1 character count	Dataset
Char_Count_Diff	EN–L1 character difference	Dataset
Char_Count_Ratio	EN–L1 character ratio	Dataset
Internal_Polysemy	Distinct L1 cues per EN target	Dataset
L1_Polysemy	Distinct EN targets per L1 cue	Dataset
WordNet_Sense_Count	Number of WordNet synsets	Miller (1995)
Crossling_Exposure	EN Zipf frequency in Spanish corpora	Speer (2022)
Psycholinguistic Features (11)		
Concreteness	Mean concreteness score	Brysbaert et al. (2014)
Is_Highly_Concrete	High-concreteness flag	Brysbaert et al. (2014)
Is_Highly_Abstract	High-abstractness flag	Brysbaert et al. (2014)
WordNet_Max_Depth	Maximum WordNet depth	Miller (1995)
WordNet_Avg_Path_Len	Average WordNet path length	Miller (1995)
L1_Zipf_Score	L1 Zipf frequency	Speer (2022)
Freq_L1_L2_Gap	EN–ES Zipf frequency gap	Speer (2022)
Log_Freq_Band	Discretised Zipf frequency band	Speer (2022)
Ctx_Avg_Word_Len	Average context word length	Dataset
Ctx_TTR	Context type-token ratio	Dataset
Clue_Coverage	Revealed-letter ratio in clue	Dataset
Residual Interaction Feature		
BERT_x_Lev_Dist	Stage 1 prediction $\times$ edit distance	This work

Table 11: Inventory of the 47 handcrafted features used for Stage 2 calibration. "Dataset" denotes features computed directly from the shared-task training data, "Orthographic" refers to deterministic character-level features, and "This work" indicates heuristic features introduced in this work.