TOEBM at BEA 2026 Shared Task 1: Improving Lexical Difficulty Prediction with Context-Aligned Contrastive Learning and Ridge Ensembling

Wicaksono Leksono Muhamad[†,1], **Joanito Agili Lopo**[†,1], **Tsamarah Rana Nugraha**[♡,1,2], **Ahmad Cahyono Adi**[♡,1,3], **Muhammad Oriza Nurfajri**[3]
[1]Mantera Studio [2]The University of Manchester [3]Universitas Gadjah Mada
{wcksnlxn,amalopo99,ahmadseverine83}@gmail.com
tsamarah.nugraha@postgrad.manchester.ac.uk
oriza_nurfajri@mail.ugm.ac.id

Abstract

Lexical difficulty prediction is a fundamental problem in language learning and readability assessment, requiring models to estimate word difficulty across different first-language (L1) backgrounds. However, existing approaches rely on regression-only training with scalar supervision, which does not explicitly structure the representation space, limiting their ability to capture cross-lingual alignment and ordinal difficulty. To mitigate these issues, we propose *Context-Aligned Contrastive Regression*, which integrates Ridge regression ensemble with two complementary objectives, i.e., Cross-View Context and Ordinal Soft Contrastive Learning. Experiments on three L1 datasets show that (i) contrastive objectives improve cross-lingual representation alignment while preserving language-specific nuances, (ii) the learned representations capture the ordinal structure of lexical difficulty, and (iii) the ensemble effectively mitigates systematic biases of individual models, leading to more stable performance across difficulty levels.[1]

1 Introduction

Vocabulary is a central component of English as a Foreign Language proficiency, supporting the development of reading, listening, writing, and speaking skills (Alshumrani, 2024). Since learners' proficiency levels influence how vocabulary is acquired (Bao and Peng, 2024), estimating lexical difficulty is an important step in the development of level-appropriate learning materials and valid assessment instruments (Goyibova et al., 2025)

Prior work has explored related tasks such as Complex Word Identification, Lexical Difficulty Prediction, and Lexical Simplification (Paetzold and Specia, 2016b; Yimam et al., 2018; Shardlow et al., 2021). However, they were not tailored for English language learners and ignored how learners' first language (L1) can make English vocabulary easier or harder. This limitation is critical, as vocabulary knowledge is inherently multi-layered and involves several interrelated components that must be acquired for effective language use (Schmitt, 2010).

Among the interrelated components, L1 interference plays a central role across linguistic levels, including grammar, syntax, phonology, and vocabulary, leading to systematic differences in how learners from different L1 backgrounds perceive and process words (Alisoy, 2024). Consequently, lexical difficulty is not an intrinsic property of a word, but a relational phenomenon that varies across L1 backgrounds, motivating the need for L1-aware modeling (Skidmore et al., 2025).

However, learner-specific factors alone are insufficient. The ability to distinguish similar-sounding words does not guarantee successful acquisition or comprehension (Pajak et al., 2016). Therefore, lexical difficulty cannot be reliably inferred from form-level properties alone. Instead, difficulty is shaped by how words are encountered and interpreted in context. Psycholinguistic evidence shows that contextual cues guide meaning interpretation and comprehension (Garten et al., 2019). Effective lexical difficulty modeling should therefore capture both learner-specific and contextual dimensions.

To address these challenges, we propose *Context-Aligned Contrastive Regression*, a multi-objective framework for L1-aware lexical difficulty prediction. Our approach (i) integrates direct regression with Cross-View Context Contrastive Learning to align representations across lexical views, (ii) incorporates Ordinal Soft Contrastive Learning to preserve graded difficulty structure, and (iii) leverages complementary encoder models through Ridge ensembling. Together, these components enable more

[†] Main contributors.
[♡] Major contributors.
[1] https://github.com/airlanggawicaksono/BEA2026TOEBM

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 1124–1136

robust, aligned, and interpretable difficulty estimation.

2 Background

Recent studies show that transformer-based models can predict lexical difficulty from contextualized and multilingual representations (Shardlow et al., 2021, 2024; Skidmore et al., 2025). These approaches are motivated by the growing importance of lexical complexity prediction in applications such as text simplification, readability assessment, and language learning (Shardlow, 2022; Rotaru, 2021). In particular, shared tasks such as SemEval-2021 have demonstrated that fine-tuned transformer models can achieve strong performance by leveraging contextual information (Shardlow et al., 2021; Rotaru, 2021).

However, most approaches rely on regression-only training, where supervision is given only through a scalar difficulty score. This can improve prediction accuracy, but it does not directly structure the representation space. As a result, items with similar difficulty may not be close in the latent space, while items with different difficulty levels may not be clearly separated. This limitation becomes more pronounced in multilingual settings, where models must capture both cross-lingual alignment and language-specific variation (Skidmore et al., 2025).

Contrastive learning addresses this limitation by shaping representations through relations between examples (Chen et al., 2020; Khosla et al., 2020). This fits lexical difficulty prediction, where difficulty depends on the target word, context, and learner-specific variation. Recent contrastive regression methods define similarity by proximity in continuous target values rather than discrete labels (Zha et al., 2023; Keramati et al., 2024), making them suitable for ordinal difficulty scores. We therefore formulate L1-aware lexical difficulty prediction as a multi-objective problem that combines direct regression with auxiliary contrastive supervision.

3 Context-Aligned Contrastive Regression

Building on recent work in lexical difficulty prediction and L1-aware modeling (Shardlow et al., 2021, 2024; Skidmore et al., 2025), we propose *Context-Aligned Contrastive Regression*, which integrates direct regression with representation-level contrastive regularization. Given an input instance enriched with L1-aware contextual information, such as translated context and English target information, the model encodes multiple contextual views into a shared representation space. The resulting representation is used for both difficulty prediction and contrastive learning.

Specifically, we optimize the model using Regression loss (§3.1) for direct difficulty prediction, a Cross-View Context Contrastive loss (§3.2) for learning view-invariant contextual representations, and an Ordinal Soft Contrastive loss (§3.3) for encoding the continuous ordering of difficulty scores in the latent space. By combining this, the overall training objective does not merely fit the target score, but also provides a representation space that is both contextually stable and smoothly aligned with the ordinal structure of the task. The general explanation of our system is presented in the Figure 1.

3.1 Regression Objective

The regression objective is used to supervise the final prediction directly. Let $h_i \in \mathbb{R}^d$ denote the shared encoder representation of item i, and let $\hat{y}_i$ denote the scalar prediction generated by the regression head. The model is optimized with the mean squared error:

$$\mathcal{L}_{\text{reg}} = \frac{1}{B} \sum_{i=1}^{B} (\hat{y}_i - y_i)^2, \tag{1}$$

where y_i is the label associated with item i and B is the batch size.

While this objective provides direct supervision for score prediction, it does not explicitly enforce representation consistency across alternative contextual realizations of the same lexical item, nor does it preserve proximity structure among items with similar difficulty levels. To address these limitations, we introduce two auxiliary contrastive objectives that regularize the representation space with respect to contextual alignment and ordinal difficulty structure.

3.2 Cross-View Context Contrastive Objective

Lexical Difficulty is largely determined by intrinsic properties of the target word rather than by superficial variation in context (Paetzold and Specia, 2016a; Gooding and Kochmar, 2018; Shardlow et al., 2021). However, contextualized encoders can mix lexical information with context-specific cues,

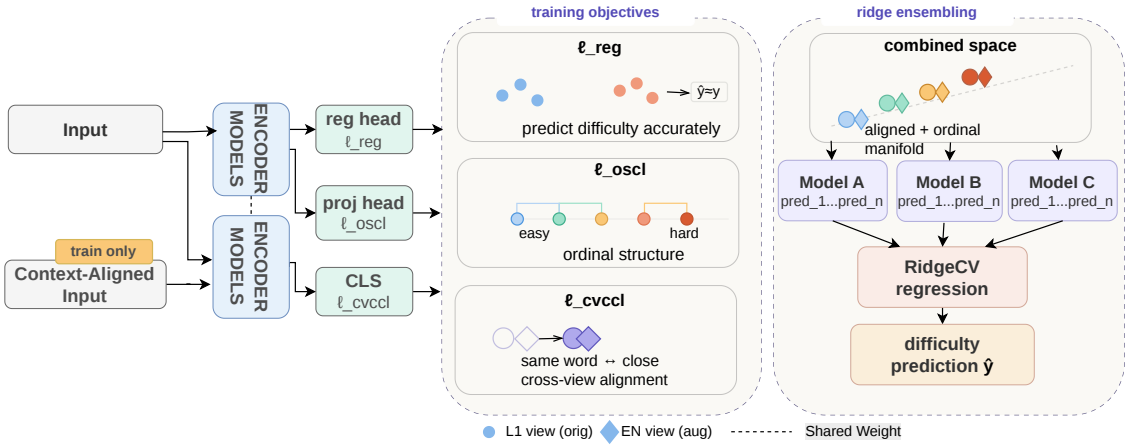


Figure 1: The proposed method combines regression and contrastive auxiliary objective, including cross-view alignment and ordinal-aware representation learning, to learn representations that are aligned across views and difficulty levels. Outputs from multiple encoder models are ensembled using ridge regression to produce the final prediction.

which may produce inconsistent representations for the same lexical item across different contexts.

To reduce this effect, we use a cross-view contrastive objective for lexical difficulty prediction. Unlike standard contrastive learning methods that rely on stochastic augmentations (van den Oord et al., 2019; Chen et al., 2020), our method uses task-specific paired views derived from the L1-aware input and the context-aligned representation described in Section 4.2. This encourages the encoder to learn representations that remain stable across contextual variation while still distinguishing different lexical items (Khosla et al., 2020).

$$z_i^{\text{tgt}} = H(en\_tgt_i) \tag{2}$$

$$z_i^{\text{full}} = H(full\_input_i) \tag{3}$$

Given a mini-batch of size B, each instance is encoded through two matched views, en_tgt (Eq. 2) and $full\_input$ (Eq. 3), yielding $2B$ representations. Since lexical difficulty prediction is formulated as a regression task, positive pairs are defined by cross-view correspondence rather than shared class labels. Thus, each anchor representation z_i is paired with its matched representation from the alternative view. The objective is formalized as

$$\mathcal{L}_{\text{CVCCL}} = -\frac{1}{2B}\sum_{i=1}^{2B}\log\frac{\exp\left(z_i^\top z_{i^+}/\tau\right)}{\sum_{k=1}^{2B}\mathbf{1}_{[k\neq i]}\exp\left(z_i^\top z_k/\tau\right)} \tag{4}$$

3.3 Ordinal Soft Contrastive Objective

While the cross-view objective promotes consistency across views, it does not explicitly capture the ordinal structure of lexical difficulty. Since difficulty is represented as a continuous psychometric estimate rather than a discrete class label (Shardlow et al., 2021), items with nearby scores should be closer in representation space than items with distant scores. Standard contrastive objectives rely on instance discrimination or discrete class supervision (Chen et al., 2020; Khosla et al., 2020), while recent regression-oriented contrastive methods address this limitation by organizing representations according to target distance or order (Zha et al., 2023; Keramati et al., 2024; Xue et al., 2024).

Following this motivation, we introduce an ordinal soft contrastive objective that replaces binary pair assignments with continuous pairwise weights derived from score proximity. This allows the model to preserve the graded structure of lexical difficulty in the embedding space. Let u_i denote the representation of item i, and let y_i be its lexical difficulty score. For a batch of size B, we define the affinity between items i and j as

$$w_{ij} = \exp\left(-\frac{(y_i - y_j)^2}{2\sigma^2}\right), \qquad w_{ii} = 0, \tag{5}$$

where σ controls how strongly the objective responds to differences in lexical difficulty. Pairs with similar scores receive larger weights, while pairs with distant scores receive smaller weights.

We then define the similarity distribution for anchor u_i over the remaining items in the batch as

$$p_{ij} = \frac{\exp\left(\mathrm{sim}(u_i, u_j)/\tau\right)}{\sum_{k \neq i} \exp\left(\mathrm{sim}(u_i, u_k)/\tau\right)}, \qquad j \neq i, \tag{6}$$

where $\mathrm{sim}(u_i, u_j)$ denotes cosine similarity and τ is a temperature parameter. The ordinal soft contrastive loss is then defined as

$$\mathcal{L}_{\mathrm{OSCL}} = \frac{1}{B} \sum_{i=1}^{B} \left(- \frac{\sum_{j \neq i} w_{ij} \log p_{ij}}{\sum_{j \neq i} w_{ij}} \right). \tag{7}$$

This objective complements the cross-view contrastive loss by shaping the representation space according to graded difficulty similarity. Items with nearby difficulty scores are encouraged to lie closer together, while items with larger score differences exert weaker attractive force. As a result, the learned embedding space reflects the continuous and ordinal nature of lexical difficulty more faithfully.

3.4 Ridge-Based Ensemble

After training the model with the proposed multi-objective learning framework, we further enhance prediction performance by leveraging an ensemble of models with different encoder backbones. Each model is trained independently using the same objective, resulting in diverse yet complementary predictions.

Let $f_k(x)$ denote the prediction of the k-th model for an input x. We construct a meta-representation by stacking predictions from K models:

$$\mathbf{z}(x) = [f_1(x), f_2(x), \ldots, f_K(x)]. \tag{8}$$

We then apply ridge regression to combine these predictions, allowing the model to learn adaptive weights over individual predictors while maintaining robustness through regularization. This approach provides a more flexible alternative to simple averaging and leads to more stable and accurate predictions.

4 Experimental Setup

4.1 Dataset

We perform lexical difficulty prediction as a regression task using the multilingual L1-aware dataset introduced by Skidmore et al. (2025). The data are organized into three first-language groups, namely German, Spanish, and Mandarin Chinese. Total of each L1 group is presented in Table 1. For the translated training and development splits, we additionally include *en_context*, as described in Section 4.2.

Furthermore, the target variable, *GLMM_score*, represents the estimated difficulty of each vocabulary item. It is computed from large-scale learner response data using a generalized linear mixed model (GLMM) framework, which models both item-level and learner-level variability (Schmitt et al., 2024). Lower scores indicate greater lexical difficulty.

Table 1: Data distribution across L1 groups and splits.

L1 group	Train	Dev	Test
German	6,091	677	748
Spanish	6,091	677	748
Mandarin Chinese	6,091	677	748
Total	18,273	2,031	2,244

4.2 L1-Aware Input Representation

Skidmore et al. (2025) reported that their best-performing setup used a multilingual model trained jointly on all L1 subsets of the KVL together with an L1-aware input representation. Following this formulation, we represent each instance as a single sequence that combines the L1 source word (w), its L1 context (ctx), the English clue ($clue$), and the target English word (tgt). These components are separated by a special token before being passed to a pretrained encoder, as shown in the example input text below:

```
casa [SEP] Vivo en una casa grande
que tiene tres dormitorios. [SEP]
h____ [SEP] house
```

In addition, we construct a context-aligned view by combining the English target word from the source language input (en_tgt) with the translated English context (en_ctx), which is obtained by translating the source-language context ctx.[2] This augmented view is then used to form contrastive pairs for cvccl objective (§3.2).

[2]We use `google/mt5-large` to translate the source-language context.

4.3 Model Architecture

Our model follows a shared-encoder architecture with task-specific heads for regression and contrastive learning. Given an input sequence, the encoder produces a contextualized representation from the [CLS] token, which is then used for both prediction and representation-level objectives.

Projection Head For contrastive learning, we apply a projection head that maps the encoder representation into a lower-dimensional space. The projected representation is used for contrastive objectives. Separating the projection space from the original representation allows the model to preserve task-relevant features for regression while learning more structured representations for contrastive objectives.

Models Setting We use three multilingual encoder models as base learners in the ensemble: XLM-RoBERTa (Conneau et al., 2020), multilingual DeBERTaV3 (He et al., 2023), and mmBERT (Marone et al., 2025). At inference time, predictions from all base learners are combined using a ridge regression meta-model, which learns to aggregate their outputs into a final prediction. The hyperparameters and training settings are reported in Appendix D.

5 Results & Discussion

In this section, we present our main result (see Table 2 and Appendix A for detailed results) along with its observations and discussion. We also provide qualitative examples of analysis across languages, describing the top-5 improved and failure cases in the dataset.

5.1 Main Result

Performance Across Languages Table 2 shows that the ensemble consistently outperforms individual base models across the three L1 groups. These gains suggest that combining input design, auxiliary objectives, and model ensembling provides more robust predictions than relying on a single encoder. Spanish achieves the best performance with the en_ctx setting, while German and Chinese benefit more from en_tgt-based representations. Rather than indicating instability, this pattern suggests that lexical difficulty is shaped by different linguistic cues across L1 backgrounds. This makes the variation between input settings analytically meaningful and motivates a deeper examination of model behavior and representation structure. [3]

Effect of Input and Objective Design A closer comparison of input and objective variants shows that more complex configurations do not always improve performance. Although $full\_input$ (§4.2) and $en\_ctx + en\_tgt$ perform strongly, they do not consistently outperform focused single-view representations, such as *en_tgt (CVCCL + OSCL)* or *en_context (CVCCL + OSCL)*. This suggests that carefully selected inputs can be more effective than combining all available features, likely by reducing noise and improving representation stability.

At the objective level, CVCCL + OSCL remains consistently competitive across languages. It does not always achieve the best score, but it stays among the top configurations in each setting. This indicates that contrastive alignment and ordinal supervision provide complementary learning signals, while their effectiveness depends on the structure of the input information.

BEA 2026 Submission We submitted our best model to the BEA 2026 Shared Task Closed Track. Full leaderboard comparisons are provided in Appendix A. Our submission, **TOEBM**, achieved top-15 performance across all L1 groups, ranking 14th for Spanish, 11th for German, and 7th for Mandarin Chinese, as shown in Tables 5, 6, and 7, respectively. This pattern suggests that the proposed approach remains reasonably robust across different L1 backgrounds, even when the relative difficulty cues vary by language. Although its RMSE is about ± 0.125 higher than the top-ranked Glite Team on average, the model maintains competitive Pearson correlation scores across languages, with 0.832 for Spanish and German, and 0.853 for Mandarin Chinese. These results indicate that, while the model is not yet fully optimized for minimizing prediction error, it captures the relative ordering of lexical difficulty effectively.

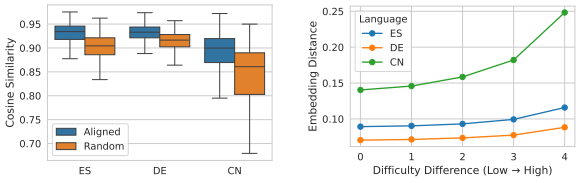
5.2 Representation Analysis

To better understand the impact of the contrastive objectives, we analyze the learned representations beyond the final prediction layer. In particular, we examine whether the contrastive losses improve cross-lingual alignment (§3.2) and preserve the ordinal structure of lexical difficulty (§3.3) across

[3] The Ensemble Baseline combines the XLM-R, mDeBERTa-v3, and mmBERT baseline models using ridge regression.

Model	ES			DE			CN		
	RMSE ↓	MSE ↓	ρ ↑	RMSE ↓	MSE ↓	ρ ↑	RMSE ↓	MSE ↓	ρ ↑
Individual Model									
XLM-R (base)	1.257	1.580	0.765	1.258	1.583	0.773	1.140	1.300	0.753
mDeBERTa-v3 (base)	1.151	1.326	0.836	1.160	1.344	0.833	1.027	1.054	0.855
mmBERT (base)	1.079	1.164	0.825	1.010	1.021	0.826	0.911	0.830	0.843
Ensemble Variants									
ensemble baseline	1.051	1.104	0.835	0.998	0.996	0.831	0.893	0.798	0.848
full_input (CVCCL + OSCL)	1.045	1.092	0.835	1.016	1.032	0.828	0.907	0.823	0.843
en_ctx + en_tgt (CVCCL + OSCL)	1.052	1.107	0.831	1.004	1.008	0.829	0.892	0.796	0.846
en_ctx (CVCCL + OSCL)	**1.041**	**1.084**	**0.837**	1.010	1.020	0.827	0.890	0.792	0.852
en_tgt (CVCCL + OSCL)	1.063	1.130	0.826	**0.997**	**0.994**	**0.832**	**0.880**	**0.774**	**0.853**
en_tgt (CVCCL)	1.052	1.107	0.831	0.998	0.996	0.831	0.891	0.794	0.847
en_tgt (OSCL)	1.077	1.160	0.825	1.001	1.002	0.832	0.866	0.750	0.858

Table 2: Performance comparison with RMSE, MSE, and Spearman's ρ across languages and models in the BEA shared task test set. Individual model denotes fine-tuned model with regression head on top of the architecture and *full_input* or L1-aware input representation §4.2 as model input. Lower RMSE/MSE and higher ρ indicate better performance.



(a) Cross-lingual alignment (b) Ordinal structure

Figure 2: Representation analysis across Spanish, German, and Chinese using mmBERT-base. (a) Cosine similarity between aligned cross-lingual lexical pairs and randomly paired inputs. (b) Relationship between embedding distance and absolute lexical difficulty score differences, larger score gaps correspond to greater representational separation.

multiple languages.

Cross-lingual Representation Alignment We first examine whether the contrastive objective leads to improved cross-lingual alignment in the learned representations. To this end, we measure cosine similarity between representations of semantically equivalent inputs (L1 and English) and compare them against randomly paired inputs.[4] As shown in Figure 2a, aligned pairs consistently exhibit higher similarity than random pairs across all three languages (ES, DE, and CN), indicating that the model learns a shared semantic space across languages.

Notably, the degree of separation varies across languages, with German showing the clearest gap, while Chinese exhibits greater variance and overlap. The presence of overlap between the two distributions suggests that the alignment is not fully separable. This indicates that while the contrastive objective contributes to reducing cross-lingual representation gaps, it does not enforce strict alignment. We hypothesize that this moderate alignment is beneficial, as it preserves language-specific nuances that may still be relevant for lexical difficulty prediction.

Ordinal Structure in Representation We next analyze whether the learned representations capture the ordinal nature of lexical difficulty. Specifically, we examine the relationship between embedding distances and differences in GLMM scores[5]. As shown in Figure 2b, we observe a clear and consistent increasing trend across all three languages, where pairs of instances with larger difficulty differences exhibit greater distances in the embedding space.

We further quantify this trend using Spearman's rank correlation. We observe a consistent positive monotonic relationship across all three languages (Spearman's ρ = 0.22 for ES, 0.20 for DE, and 0.39 for CN), indicating that embedding distances tend to increase with larger differences in lexical difficulty. While the correlation is moderate in magnitude, its consistency across languages suggests that the ordinal structure is reliably encoded in the learned representations, although the relationship is not strictly linear.

[4]The details of measurement technique is presented in Appendix F.

[5]The details of measurement technique is presented in Appendix E

Condition	Low	High	Rel. ↓ (%)
Context Length	0.841	0.834	0.8
Lexical Diversity	0.847	0.837	1.2
Target Word Length	0.843	0.831	1.4
Orthographic Complexity	0.855	0.813	4.9

Table 3: Mean absolute error (MAE) of the Ridge ensemble across input characteristics on the Spanish dataset. Relative improvement (%) measures the reduction from low to high condition. Lower values indicate better performance.

Input Representation Behavior Beyond representation structure, we further investigate how input characteristics influence prediction behavior. Specifically, we analyze prediction error with respect to several surface-level input properties, including context length, lexical diversity (measured via type-token ratio), target word length, and orthographic complexity. We observe a generally consistent pattern in which richer contextual and lexical signals are associated with lower prediction error. In particular, longer and lexically more diverse contexts, as well as more orthographically informative target words, tend to improve prediction quality, suggesting that contextual and lexical richness provide useful signals for lexical difficulty modeling.

However, the strength of this effect varies across languages. As shown in Table 3, for Spanish, the improvements are more pronounced, especially in lexically diverse contexts and orthographically complex words, which exhibit the largest relative gain. This indicates that Spanish benefits more from contextualized representations, explaining why the en_ctx configuration outperforms target-only inputs in this setting. In contrast, for other languages, the gains from context are less dominant, suggesting that target-side information can be sufficient depending on the dataset characteristics. This finding is consistent with the representation-level analysis, where contextual signals contribute to richer and more structured embeddings, which in turn improve downstream prediction.

5.3 Ensemble Investigation

We analyze the contribution of each base model to the meta-input of the Ridge regression ensemble. This analysis aims to assess how well individual models approximate the GLMM scores and whether they provide complementary prediction patterns that can be effectively leveraged by the ensemble.

Base Models Prediction Behaviour Figure 3 illustrates the relationship between predicted scores and ground truth GLMM scores across all three L1 groups. The models exhibit systematic differences in how lexical difficulty is estimated. XLM-RoBERTa consistently produces a relatively flat trend, particularly in the Chinese dataset, indicating a tendency to compress predictions toward the mean and underestimate more difficult items. In contrast, mDeBERTa-v3 shows a steeper slope, often overshooting the ground-truth line, suggesting systematic overestimation for complex lexical items. Meanwhile, mmBERT demonstrates a more moderate slope that more closely follows the optimal diagonal trend, indicating a more balanced estimation behaviour.

Statistical Patterns These contrasting trends reveal that the base models capture different aspects of the lexical difficulty space. While XLM-RoBERTa tends to under-predict and mDeBERTa-v3 tends to over-predict, mmBERT produces more balanced predictions with errors closer to zero and generally competitive variance. This complementary behavior is further reflected in the trade-off between correlation and error metrics, where mDeBERTa-v3 often achieves higher Pearson correlation while mmBERT attains lower RMSE and MAE in several cases.

When considered alongside the statistical results in Table 8 (Appendix), these findings suggest that the models are not redundant but instead provide diverse and complementary signals. Such diversity is crucial for the effectiveness of the ensemble, as it enables the meta-learner to reconcile systematic biases and produce more accurate final predictions.

Difficulty-wise Error Analysis To further evaluate ensemble effectiveness across the lexical difficulty spectrum, we analyze model errors over five GLMM-based difficulty bins averaged across all three L1 groups. Figure 4 presents the resulting mean absolute error (MAE) patterns.

We observe that XLM-RoBERTa exhibits a substantial increase in error from harder to easier lexical items, indicating that its predictions become progressively less reliable for comparatively easier words and suggesting strong score compression effects. Similarly, mmBERT demonstrates moderate but consistent error growth toward easier bins, although its overall performance remains more balanced than XLM-RoBERTa. In contrast, mDeBERTa-v3 maintains relatively stable perfor-

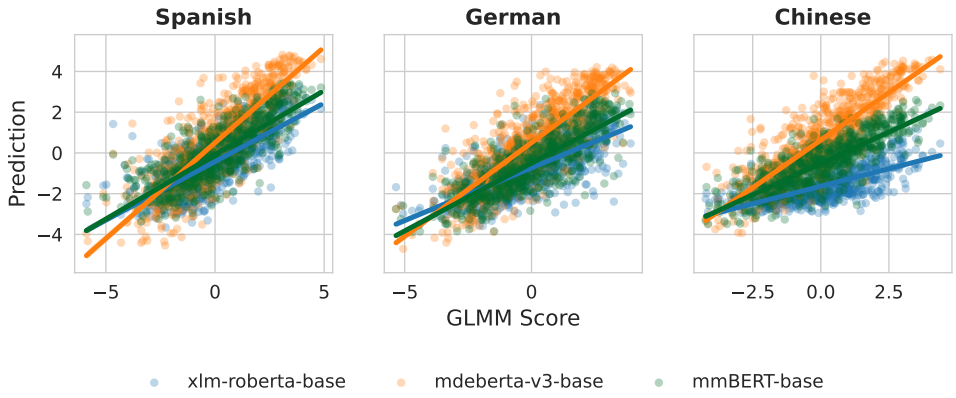


Figure 3: Relationship between predicted scores and ground-truth GLMM scores across Spanish (ES), German (DE), and Chinese (CN). Each subplot shows regression trends for individual base models, highlighting systematic differences in prediction behavior, including underestimation, overestimation, and balanced estimation patterns.

mance across the full spectrum, with only minor variation between difficulty levels.

Importantly, the Ridge ensemble consistently achieves the lowest or near-lowest MAE across nearly all bins while maintaining the flattest error profile overall. This suggests that the meta-learner effectively reconciles the complementary weaknesses of individual base models, substantially reducing systematic prediction bias and producing more stable lexical difficulty estimation across both harder and easier lexical items.

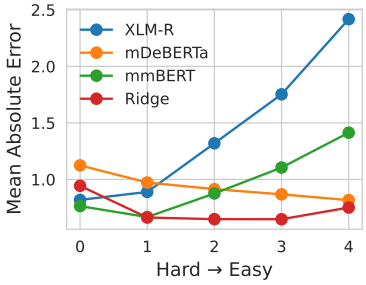


Figure 4: Mean absolute error (MAE) across lexical difficulty bins, averaged over all three L1 groups. Difficulty bins are derived from GLMM score quantiles and ordered from harder (lower GLMM scores) to easier (higher GLMM scores) lexical items.

5.4 Improved vs Failures Cases

To complement the quantitative findings in Section §5.1, we further examine representative cases to understand how the ensemble model corrects prediction errors from individual base models.

Error Correction Table 4 presents selected examples across Spanish, Chinese, and German. We observe that the ensemble consistently corrects large underestimation errors from individual models. For instance, words such as *cruise*, *run*, and *character* in Spanish, as well as *schoolmate* and *dating* in Chinese, are substantially underestimated by mmBERT, yet the Ridge ensemble successfully adjusts their predictions toward the ground-truth GLMM scores. Similar correction patterns are observed in German, suggesting that the ensemble effectively leverages complementary signals across models to reduce systematic bias.

Lang	Word	GLMM	mmBERT	Ridge
ES	cruise	1.21	-0.50	1.23
ES	run	2.20	0.31	1.99
ES	character	1.08	-1.11	0.55
ES	dining	-4.67	-0.76	0.19
ES	baking	-3.78	-0.20	0.87
CN	schoolmate	1.32	-2.55	-0.09
CN	dating	0.64	-2.31	-0.001
CN	goldfish	1.51	-1.44	0.63
CN	double	-2.58	0.11	0.82
CN	stake	-4.19	-1.67	-1.13
DE	cruise	1.21	-1.48	0.60
DE	handbook	0.95	-1.36	0.31
DE	workbook	2.10	-2.07	-0.44
DE	anger	-3.39	0.10	0.57
DE	umbrella	3.24	-1.13	-0.20

Table 4: Qualitative examples across Spanish (ES), Chinese (CN), and German (DE), comparing mmBERT-base and Ridge ensemble predictions. The first three rows per language show cases where the ensemble corrects large underestimation errors from base models, while the last two rows highlight failure cases.

Failure Cases Despite these substantial improvements, certain lexical items remain challenging even after ensembling. In several cases, such as *dining* and *baking* in Spanish, *double* and *stake* in

Chinese, and *anger* and *umbrella* in German, the predicted difficulty remains far from the ground-truth values. These errors often involve strong overestimation or residual bias, indicating that certain lexical items remain challenging due to semantic ambiguity or context-dependent interpretations. This suggests that while the ensemble mitigates systematic biases, it cannot fully resolve cases where all base models fail to capture the underlying difficulty signal.

6 Conclusion

Our method consistently outperforms individual multilingual encoders across Spanish, German, and Chinese learner groups, effectively correcting systematic under- and overestimation biases while maintaining stable performance across all difficulty levels. Representation analysis confirms that the contrastive objectives enhance cross-lingual alignment and preserve the ordinal structure of lexical difficulty in the embedding space. Collectively, these results demonstrate that combining direct regression with representation-level regularization significantly improves both the accuracy and structural reliability of L1-aware lexical difficulty prediction.

Acknowledgments

The authors gratefully acknowledge Mantera Studio for providing the computational resources and APIs utilized in this research. Additionally, this work was partially supported by the 2026 Publication Funding from the Department of Computer Science and Electronics at Universitas Gadjah Mada.

Limitations

Although the proposed method improves performance across the evaluated L1 groups, several limitations remain. First, the experiments are limited to Spanish, German, and Mandarin Chinese, so the findings may not fully generalize to other L1 backgrounds. Second, the use of contrastive objectives introduces additional training objectives and weighting hyperparameters, which may increase sensitivity to objective balancing across languages. Third, the ridge ensemble improves predictive performance by combining multiple encoders, but it also increases computational cost because several models must be trained and evaluated. Future work should examine more robust weighting strategies, reduce the computational overhead of ensembling, and evaluate the method on a wider range of L1 backgrounds.

Ethical Considerations

We acknowledge that our research utilized AI tools for writing, rewriting, and generating code. Although these tools offer significant advantages in terms of efficiency and productivity, their use raises important ethical considerations. We recognize the potential for bias and errors inherent in AI-generated content and have taken steps to mitigate these risks through rigorous human review and validation. Furthermore, we are mindful of the potential impact on the broader software development community, particularly regarding job displacement and the need for upskilling. We believe that responsible AI integration should prioritize transparency, accountability, and the empowerment of human developers, ensuring that these tools augment rather than replace human expertise. This research aims to contribute to the ongoing dialogue on ethical AI development and usage, advocating for a future where AI tools are harnessed responsibly to enhance human creativity and innovation in the field of software engineering.

References

Hasan Alisoy. 2024. Exploring language acquisition: The role of native language interference in esl learners. *Journal of Azerbaijan Language and Education Studies*, 1(1):50–66.

H. A. Alshumrani. 2024. Unveiling vocabulary teaching and learning beliefs of teachers and learners in an EFL context. *Asian Journal of Second and Foreign Language Education*, 9(1):20.

Z. Bao and C. Peng. 2024. The effects of EFL wordlist and proficiency on vocabulary knowledge. *Frontiers in Psychology*, 15:1289106.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

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.

Justin Garten, Blair Kennedy, Kenji Sagae, and Morteza Dehghani. 2019. Measuring the importance of context when modeling language comprehension. *Behavior Research Methods*, 51(2):480–492.

Sian Gooding and Ekaterina Kochmar. 2018. CAMB at CWI shared task 2018: Complex word identification with ensemble-based voting. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 184–194, New Orleans, Louisiana. Association for Computational Linguistics.

Nigora Goyibova, N. Muslimov, Barnokhon Samatova, and 1 others. 2025. Differentiation approach in education: Tailoring instruction for diverse learner needs. *International Journal of Educational Research*, 125:102501. Accessed via ScienceDirect.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Mahsa Keramati, Lili Meng, and R. David Evans. 2024. Conr: Contrastive regularizer for deep imbalanced regression. In *The Twelfth International Conference on Learning Representations*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmbert: A modern multilingual encoder with annealed language learning. *Preprint*, arXiv:2509.06888.

Gustavo Paetzold and Lucia Specia. 2016a. Inferring psycholinguistic properties of words. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 435–440, San Diego, California. Association for Computational Linguistics.

Gustavo Paetzold and Lucia Specia. 2016b. SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.

Bozena Pajak, Sarah C. Creel, and Roger Levy. 2016. Difficulty in learning similar-sounding words: A developmental stage or a general property of learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(9):1377–1399.

Armand Rotaru. 2021. ANDI at SemEval-2021 task 1: Predicting complexity in context using distributional models, behavioural norms, and lexical resources. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 655–660, Online. Association for Computational Linguistics.

N. Schmitt, K. Dunn, B. O'Sullivan, L. Anthony, and B. Kremmel. 2024. *Knowledge-based Vocabulary Lists*. British Council Monographs on Modern Language Testing. University of Toronto Press.

Norbert Schmitt. 2010. *Researching Vocabulary: A Vocabulary Research Manual*. Palgrave Macmillan, Basingstoke.

Matthew Shardlow. 2022. Agree to disagree: Exploring subjectivity in lexical complexity. In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, pages 9–16, Marseille, France. European Language Resources Association.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, and 3 others. 2024. The BEA 2024 shared task on the multilingual lexical simplification pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589, Mexico City, Mexico. Association for Computational Linguistics.

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.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. *Preprint*, arXiv:1807.03748.

Tengfei Xue, Fan Zhang, Leo R. Zekelman, Chaoyi Zhang, Yuqian Chen, Suheyla Cetin-Karayumak, Steve Pieper, William M. Wells, Yogesh Rathi, Nikos Makris, Weidong Cai, and Lauren J. O'Donnell. 2024. Tractoscr: a novel supervised contrastive regression framework for prediction of neurocognitive measures using multi-site harmonized diffusion mri tractography. *Frontiers in Neuroscience*, 18:1411797.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana. Association for Computational Linguistics.

Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. 2023. Rank-n-contrast: Learning continuous representations for regression. In *Thirty-seventh Conference on Neural Information Processing Systems*.

A BEA 2026 Leaderboard Comparison

Table 5, Table 6, and Table 7 present selected leaderboard comparisons for the Spanish, German, and Chinese subtasks, respectively. Rather than listing the complete leaderboard, we report the highest-ranked systems and nearby reference systems to contextualize our submission. Our team is registered as **TOEBM**. In the closed-track results, TOEBM ranked 14th for Spanish, 11th for German, and 7th for Chinese, with RMSE scores of 1.063, 0.997, and 0.880, respectively. The leaderboard scores differ slightly from those reported in the main experimental section because the present paper uses a unified prompt setting across all subtasks, while the official leaderboard reflects the submitted prediction files. Nevertheless, the relative ranking of our submission remains unchanged.

#	Team Name	Prediction Type	RMSE	Pearson
1	Glite	predictions_run_3	0.903	0.877
...	...	...	...	...
7	Sakura	predictions_closed_max	0.983	0.854
11	NLP-Explorers	predictions_run_3	1.041	0.838
...	...	...	...	...
14	TOEBM	predictions_run_3	1.063	0.826

Table 5: Leaderboard comparison of Closed Track Spanish Language

#	Team Name	Prediction Type	RMSE	Pearson
1	Glite	predictions_run_3	0.885	0.871
...	...	...	...	...
4	Uogal	8enc_dmeta_elasticnet...csv	0.903	0.869
7	NLP-Explorers	predictions_run_3	0.992	0.845
...	...	...	...	...
11	TOEBM	predictions_run_3	0.997	0.832

Table 6: Leaderboard comparison of Closed Track German Language

#	Team Name	Prediction Type	RMSE	Pearson
1	Glite	predictions_run_3	0.776	0.889
...	...	...	...	...
4	Sakura	predictions_closed_max	0.816	0.874
5	uogal	8enc_dmeta_elasticnet...csv	0.820	0.879
...	...	...	...	...
7	TOEBM	predictions_run_3	0.880	0.853

Table 7: Leaderboard comparison of Closed Track Chinese Language

B Base Model Statistic

C Representation analysis

To analyze the learned representation space, we distinguish between the Ridge ensemble used for prediction and the fused embedding used for visualization. The Ridge ensemble prediction is defined as

$$\hat{y}(x) = b + \sum_{n=1}^{N} w_n f_n(x), \tag{9}$$

where $f_n(x)$ denotes the scalar prediction produced by the n-th base model, w_n is the learned Ridge coefficient for that model, and b is the intercept. This formulation is used only for final score prediction. It is not suitable for direct t-SNE visualization because it produces a one-dimensional output.

For representation analysis, we instead construct a fused embedding by concatenating the weighted embedding vectors from each base model:

$$z(x) = \left[w_1 E_1(x);\ w_2 E_2(x);\ \cdots;\ w_N E_N(x) \right], \tag{10}$$

where $E_n(x) \in \mathbb{R}^{d_n}$ is the embedding produced by the n-th encoder. In this study, $E_n(x)$ is obtained from the CVCCL+OSCL model using the en_tgt view as the encoder input. The Ridge coefficients are then used as model-level weights to scale each embedding before concatenation.

We apply t-SNE to the fused representation $z(x)$ to visualize the structure of the embedding space. The resulting points are grouped into five difficulty bins based on the minimum and maximum GLMM scores. This allows us to inspect whether examples with similar lexical difficulty are located close to one another in the representation space, while the Ridge ensemble prediction $\hat{y}(x)$ is used separately to evaluate predictive performance. Figure 5 shows that the separation across difficulty bins is generally strong, although the quality of the structure varies across models and languages. Since GLMM is a continuous score and the bins are created only for visualization, perfect discrete clusters are not

Metrics	Español			Deutsch			Chinese		
	XLM-R	mDeBERTa	mmBERT	XLM-R	mDeBERTa	mmBERT	XLM-R	mDeBERTa	mmBERT
RMSE	1.326	1.266	**1.113**	1.441	1.219	**1.154**	2.070	1.135	**1.091**
MAE	1.066	0.993	**0.861**	1.145	0.959	**0.908**	1.793	0.891	**0.880**
Pearson	0.746	**0.836**	0.808	0.719	**0.832**	0.813	0.689	**0.858**	0.824
Mean Error	-0.432	0.497	**-0.098**	-0.722	0.507	**-0.491**	-1.644	0.628	**-0.520**
Std Error	1.253	1.165	**1.109**	1.247	1.108	**1.045**	1.258	**0.945**	0.960

Table 8: Performance and error analysis of base models across three L1 groups. Best values per language are highlighted in bold.

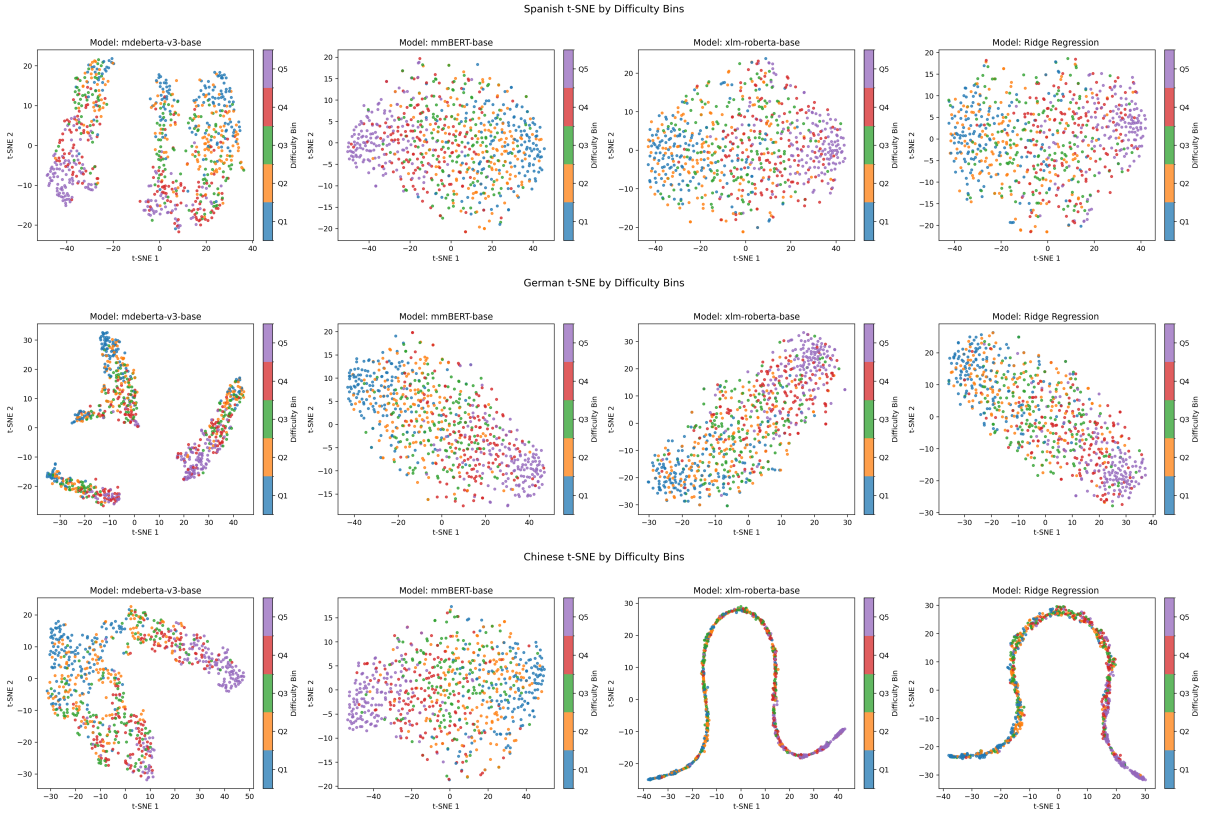


Figure 5: t-SNE visualization of fused model representations for Spanish, German, and Chinese. Each point represents one sample, and colors indicate five GLMM difficulty bins constructed from the minimum and maximum GLMM scores. The visualization is based on the weighted concatenation of encoder embeddings, while $\hat{y}(x)$ is used separately for predictive evaluation.

expected. Even so, a clear ordinal organization is visible in many panels, where neighboring regions are dominated by adjacent difficulty bins and the progression from lower to higher difficulty appears smooth.

Among the individual encoders, `mdeberta-v3-base` often produces well-separated local clusters, especially in Spanish and German, indicating that the learned space captures meaningful difficulty structure. `mmBERT-base` also shows some ordering, but the bins are more mixed and the boundaries between difficulty levels are less distinct. In contrast, `xlm-roberta-base` provides the clearest structure for Mandarin Chinese. In the Chinese panel, the bins are arranged along a highly organized manifold with a very strong gradual transition from easier to harder examples. This suggests that `xlm-roberta-base` captures the ordinal nature of lexical difficulty particularly well for Mandarin Chinese.

The Ridge-weighted fused representation remains competitive and in some cases preserves a smooth global arrangement of the bins, showing that the ensemble combines complementary information from the base encoders. However, the clearest single-model structure for Mandarin Chinese is produced by `xlm-roberta-base`. Overall, these visualizations support the claim that the

CVCCL+OSCL representations encode lexical difficulty in a meaningful way, with strong separation between bins and especially clear ordinal structure in the Mandarin Chinese setting.

D Training Settings

All models are trained using the Adam optimizer for 5 epochs with a learning rate of `2e-5`, a batch size of `8`, weight decay of `0.01`, and a warmup ratio of `0.06`. The best model is selected based on the Mean Squared Error (MSE) on the validation set. The maximum input length is set to 128 tokens, while the maximum length for `en_context` is limited to 32 tokens. Furthermore, we set $\lambda_{\text{cv}} = 0.1$ and $\lambda_{\text{ord}} = 0.2$. The temperature parameter τ is set to 0.01 and 0.02 for the `cvccl` and `oscl` losses, respectively.

For the Ridge regression meta-model, we consider a set of candidate regularization parameters $\alpha \in \{0.01, 0.1, 1.0, 10.0, 100.0\}$ and select the best value via cross-validation. Finally, We use Root Mean Squared Error (RMSE) as the primary metric for system ranking and report Pearson correlation for completeness.

E Ordinal Structure Analysis Details

To quantitatively assess whether the learned representations preserve the ordinal structure of lexical difficulty, we analyze the relationship between embedding distances and differences in ground-truth difficulty scores. Given a set of instances $\{(x_i, y_i)\}_{i=1}^{N}$, where y_i denotes the GLMM difficulty score, we first obtain normalized sentence embeddings $\mathbf{h}_i \in \mathbb{R}^d$. For a sampled subset of size N', we compute all pairwise cosine distances:

$$d_{ij} = 1 - \frac{\mathbf{h}_i \cdot \mathbf{h}_j}{\|\mathbf{h}_i\|\|\mathbf{h}_j\|}$$

and corresponding absolute difficulty differences:

$$\Delta_{ij} = |y_i - y_j|.$$

We then group instance pairs into K bins based on quantiles of Δ_{ij} and compute the mean embedding distance within each bin. This results in a function $f(k)$ that maps increasing difficulty differences to average embedding distances. A monotonically increasing trend in $f(k)$ indicates that the representation space is structured according to difficulty.

F Cross-lingual Alignment Analysis Details

To evaluate whether the learned representations capture cross-lingual semantic alignment, we measure the cosine similarity between representations of semantically corresponding inputs across languages. For each instance x_i, we construct two views: an L1-based input and its corresponding English-based input. We then obtain normalized embeddings $\mathbf{h}_i^{\text{L1}}$ and $\mathbf{h}_i^{\text{EN}}$ using the encoder.

We define the similarity of aligned pairs as:

$$s_i^{\text{aligned}} = \mathbf{h}_i^{\text{L1}} \cdot \mathbf{h}_i^{\text{EN}},$$

and construct a baseline by randomly permuting the English representations to obtain mismatched pairs:

$$s_i^{\text{random}} = \mathbf{h}_i^{\text{L1}} \cdot \mathbf{h}_{\pi(i)}^{\text{EN}},$$

where $\pi(i)$ is a random permutation. By comparing the distributions of s^{aligned} and s^{random}, we assess whether the model places semantically corresponding inputs closer in the embedding space than unrelated ones.