Jinnie's Lab at BEA 2026 Shared Task 1: Precalibration of Vocabulary Item Difficulty with Multilingual Transformers and Multi-Task Learning

Zhe Li
University of Florida
Gainesville, FL
zheli@ufl.edu

Pauline Aguinalde
University of Florida
Gainesville, FL
aaguinalde@ufl.edu

Jinnie Shin
University of Florida
Gainesville, FL
jinnie.shin@coe.ufl.edu

Abstract

This paper describes our submission to the BEA 2026 shared task 1 on vocabulary item difficulty prediction in multilingual settings. We investigated whether transformer-based representations learned directly from item content can support the prediction of vocabulary item difficulty across different L1 groups. Our approach adopted a multilingual BERT-based architecture, specifically the mmBERT, with representation augmentation at both the layer and token levels, followed by a multi-task cascade learning that incorporates part-of-speech information as an auxiliary structural signal. Results showed that multi-task mmBERT consistently outperforms the shared-task XLM-RoBERTa baseline across languages, while gains from more complex aggregation are not uniform. The findings showed that strong multilingual representations provide a competitive foundation for vocabulary item difficulty prediction, while the benefits of additional architectural complexity depend on the language and training setting.

1 Introduction

Item difficulty prediction aims to estimate item parameters from item features, with the long-term goal of reducing reliance on resource-intensive calibration procedures while preserving the quality of measurement (Wauters et al., 2012). Within the educational measurement community, this task is not only a question of predictive accuracy, but also of whether model-based estimates can serve as valid and interpretable proxies for empirically calibrated item parameters (e.g., Ulitzsch et al., 2026).

Recent advances in machine learning and natural language processing (NLP), particularly transformer-based models and large language models (LLMs), have substantially expanded the methodological toolkit available for this task, most often through supervised deep learning approaches. Broadly, existing approaches to item difficulty prediction can be organized along three dimensions: (1) the use of representations learned directly from item text, (2) the incorporation of features generated by LLM-based or other generative frameworks, and (3) the inclusion of response-based information (e.g., student correctness) as part of the prediction process. In this study, we focus specifically on model classes that leverage representations learned directly from item content, treating these representations as the primary source of information for predicting item difficulty.

Transformer-based models are particularly well-suited for this setting due to their ability to generate contextualized, high-dimensional representations of text that capture semantic and syntactic relationships. Their use in item difficulty prediction implicitly assumes that the linguistic signal contained in the item, such as the target word, its context, and surrounding cues, encodes sufficient information to infer the cognitive demand required for a correct response (e.g., Peters et al., 2025). While this assumption has shown promise in domains where item stimuli are relatively self-contained (e.g., reading comprehension), its applicability to more narrowly defined constructs, such as vocabulary knowledge, raises distinct methodological and measurement considerations.

In vocabulary assessment, item difficulty is closely tied to lexical complexity, which reflects not only properties of the target word but also the interaction between the word, its context, and the examinee's linguistic background. Earlier approaches to vocabulary item difficulty prediction relied heavily on hand-crafted features, including word length, frequency, part-of-speech (POS) tags, concreteness, and age-of-acquisition (Yang and Suyong, 2018; Settles et al., 2020). While these features offer interpretability, they are limited in their ability to capture contextual meaning and semantic variability across usages.

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 1077–1090

More recent work has shifted toward leveraging contextual embeddings from pre-trained language models, particularly encoder-only architectures such as BERT (Devlin et al., 2019), to model lexical complexity in context (e.g., Zaharia et al., 2021; Kelious et al., 2024). These approaches often combined contextual representations with psycholinguistic features to improve performance. However, from a measurement perspective, it remains unclear to what extent these representations capture construct-relevant variances, as opposed to incidental properties of language use. This challenge becomes more pronounced in cross-lingual settings, where item difficulty may vary as a function of examinees' first-language (L1) backgrounds and the linguistic relationship between source and target languages. For example, Skidmore et al. (2025) demonstrated that multilingual transformer models (e.g., XLM-RoBERTa Conneau et al., 2020) can be fine-tuned to predict vocabulary item difficulty across multiple languages, suggesting the potential of cross-lingual transfer for item calibration. At the same time, these approaches face challenges related to polysemy, low-frequency cognates, and subword tokenization, all of which may introduce construct-irrelevant variance into predictions.

In this paper, we participate in the BEA 2026 shared task on vocabulary item difficulty prediction and investigate how representation learning and multi-task modeling choices affect predictive performance across languages. Specifically, we explore a multilingual BERT-based architecture (mmBERT; Marone et al., 2025) with layer-wise and token-wise representation aggregation, alongside a multi-task cascade that incorporates part-of-speech (POS) information as an auxiliary structural signal.

Our goal is twofold: (1) to evaluate the effectiveness of these design choices within the shared-task setting, and (2) to examine, at a high level, how different representation strategies relate to the modeling of lexical item difficulty. Through this analysis, we aim to provide both a competitive system and a set of practical insights into what types of representations are most effective for vocabulary item difficulty prediction in multilingual contexts.

2 Related Work

2.1 Transformer-based Item Difficulty Prediction

Within educational measurement, item difficulty is a fundamental parameter that underpins score interpretation, test assembly, and the evaluation of measurement quality. In both Classical Test Theory (CTT) and Item Response Theory (IRT), difficulty is not directly observed but inferred from examinee response data through calibration procedures. Although these approaches provide a rigorous and defensible basis for parameter estimation and validity arguments, they rely on sufficiently large and representative samples, which can limit the efficiency and scalability of item development. Consequently, a growing body of research has examined the prediction of item difficulty prior to operational calibration using features derived from item content (e.g., Peters et al., 2025). These efforts are best understood not as substitutes for psychometric calibration, but as methods for producing provisional estimates that can inform item screening, test construction, and adaptive content generation in the absence of response data.

Recent advances in transformer-based models and large language models (LLMs) have expanded the methodological approaches available for this purpose. One class of methods uses pretrained language models to derive semantic representations of item text, which are then used to predict relative or absolute difficulty. For example, Loginova et al. (2021) demonstrated that contextual embeddings can support pairwise difficulty comparisons without relying on calibrated parameters or observed response vectors, although such predictions remain model-based approximations rather than psychometric estimates grounded in examinee behavior. Related work (e.g., AlKhuzaey et al., 2024; McCarthy et al., 2021) similarly treats text as the primary source of information for difficulty prediction.

A second line of work attempts to approximate difficulty through simulated response processes, often by treating LLMs as proxy examinees. Maeda (2025) generate synthetic response data from trained models to estimate item difficulty in settings where field testing is limited. While promising, these approaches introduce additional assumptions regarding the correspondence between model-generated responses and human performance, raising important questions about construct representation, population invariance, and the interpretability of resulting parameters.

2.2 Vocabulary Item Difficulty Prediction

A specific area within item difficulty prediction concerns vocabulary assessment, where the item difficulty is closely tied to the linguistic properties

of target words. In these contexts, predicting difficulty prior to calibration often relies on features that capture how learners process lexical content.

A closely related line of work is lexical complexity prediction, which reflects the perceived difficulty of a word for a given population (North et al., 2023). This construct has been widely studied in NLP and applied linguistics, particularly in tasks such as lexical simplification (Shardlow et al., 2022). In vocabulary learning contexts, complexity is typically modeled using morphological, semantic, and contextual features, sometimes incorporating learner characteristics such as L1. For example, Alfter and Volodina (2018) predicted comprehension levels for Swedish learners using linguistically informed features and machine learning methods. More recent work has leveraged contextualized representations from transformer-based models (e.g., BERT, RoBERTa), showing improved performance over traditional approaches such as SVM and Random Forest (Zaharia et al., 2021; Yaseen et al., 2021).

From a measurement perspective, lexical complexity and item difficulty are conceptually distinct. Lexical complexity operates at the word level and reflects processing demands, whereas item difficulty is a latent parameter defined through examinee responses. Nonetheless, lexical complexity provides a construct-relevant source of information for vocabulary item difficulty prediction. Word-level features may explain meaningful variation in item difficulty, particularly when aligned with population characteristics such as learners' L1, which can systematically influence lexical processing (Graham et al., 2010; Derakhshan and Karimi, 2015; Schneider et al., 2026).

2.3 Multilingual BERT regression

Predicting item difficulty that accounts for lexical complexity can be approached as a regression problem. As such, transformer-based regression approaches have been adopted to predict lexical complexity. Ortiz-Zambrano et al. (2025)'s study highlighted the utility of incorporating linguistic features with BERT, XLMR, and RoBERTa towards improving lexical complexity prediction across English and Spanish contexts. Variations of transformer-based models employing regression models have been introduced in a lexical complexity prediction task in 2021 (North et al., 2023). DeepBlueAI, one of the top performing models, incorporated a final estimator with a simple linear regression model (Pan et al., 2021).

Similarly, BERT-based regression models have been employed to predict item difficulty (Yaneva et al., 2024). Benedetto et al. (2023) analyzed quantitative approaches to item difficulty and found that BERT consistently outperforms baseline machine learning models, while hybrid models that incorporate linguistic features further improve prediction performance. Further studies have adopted regression-based approaches for difficulty prediction within various domains, including medicine (Ram and Kesanam, 2024), algebra (Aradelli, 2019), and reading (Kapounová, 2025). Foundational to this year's BEA shared task, Skidmore et al. (2025) examined item difficulty in a vocabulary learning setting. Thus, integrating insights from research on lexical complexity and item difficulty prediction are promising towards modeling difficulty in multilingual learning contexts.

2.4 Research Gap

Despite progress in lexical item complexity prediction, estimating the continuous difficulty of vocabulary test items in cross-lingual contexts remains difficult. Traditional methods emphasize hand-crafted linguistic features but fail to capture deep semantic dependencies (Yang and Suyong, 2018; Settles et al., 2020), while standard multilingual transformers offer cross-lingual transfer yet often miss nuanced difficulty variations confounded by specific target word senses (Skidmore et al., 2025). Although some work proposes integrating psycholinguistic features with deep embeddings, most automated approaches merely concatenate deep semantic representations with hand-crafted features prior to final prediction (Bulut et al., 2024; Benedetto et al., 2023; Ortiz-Zambrano et al., 2025). Furthermore, advanced frameworks such as multi-task learning, which leverage auxiliary linguistic tasks to improve primary predictions (Sanh et al., 2018; Liu et al., 2019; Galal et al., 2024), remain underexplored. To address this gap, we introduce an mmBERT-based multi-task architecture with layer-wise and token-wise aggregation that explicitly models cross-lingual semantic relationships and syntactic biases through auxiliary POS tagging, enabling more accurate and context-aware difficulty prediction.

3 Methods

In this section, we present the system architecture, including the data preprocessing, BERT-based rep-

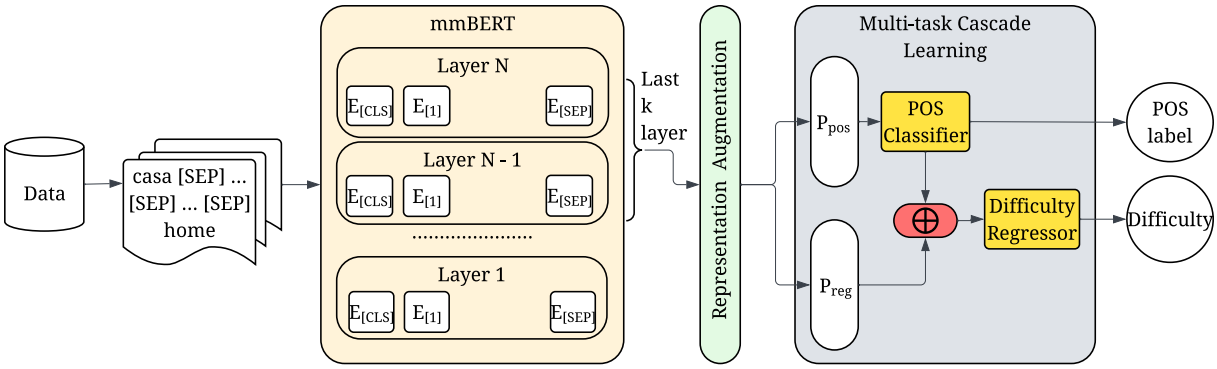

Figure 1: Overview of the proposed system architecture.

resentation generation, the layer-wise and token-wise representation augmentation, and multi-task downstream prediction layer. An overview of the system is presented in Figure 1. First, the item text is tokenized by concatenating distinct linguistic components into a single multilingual sequence. Second, this sequence is passed through a BERT family model, specifically the mmBERT architecture, to extract comprehensive cross-lingual semantic representations. To maximize the utility of the deep semantic information, the system applies both layer-wise fusion, such as scalar mixing or mean/-max token pooling, and token-wise aggregation via self-attention to output an augmented item representation. Third, the augmented representation is utilized in a hierarchical multi-task architecture. It initially computes auxiliary part-of-speech tagging logits for the English target word. Then, these auxiliary signals are detached and injected into the primary task by concatenating directly with the primary semantic representation to predict the final continuous lexical difficulty score.

To encode the test item, we followed the approach by Skidmore et al. (2025) to organize the input text into a single multilingual sequence. The sequence follows the presentation order of the vocabulary test items, concatenating the source language word, the source language context, the English target clue, and the target English word using the [SEP] tokens. This structure allows the encoder to capture cross-lingual relationships and contextual constraints.

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

Target labels In this supervised framework, each task utilizes a label to guide learning. The primary label is a continuous difficulty score derived from generalized linear mixed model (GLMM) estimates to quantify the lexical challenge of the English target word. Concurrently, the system predicts an auxiliary label representing the POS tag of the target word. Because English target words are lemmas agnostic to specific word senses, and certain senses are more difficult for learners, the POS tag provides a structural inductive bias. This is directly injected into the final difficulty predictor through a multi-task cascading architecture (Goldberg, 2022).

Representation Generation Given BERT-based regression model's strong performance and widespread success on both lexical complexity prediction and item difficulty prediction tasks (Ortiz-Zambrano et al., 2025; Ram and Kesanam, 2024; Kelious et al., 2024; Skidmore et al., 2025; Kapounová, 2025), our system uses the encoder-only model to serve as the representation generator. Specifically, the system utilizes mmBERT (Marone et al., 2025), the latest multilingual encoder that excels across diverse multilingual NLP tasks compared to its predecessors (e.g. XLM-RoBERTa).

3.1 Representation augmentation

After obtaining the token-level representation of the item sequence using the mmBERT encoder, we aggregate hidden states across both intermediate layers and sequence tokens, inspired by the finding that task-relevant information in transformers is distributed within its deep and wide structure (Rogers et al., 2021; Galal et al., 2024; Behrendt et al., 2025; Ciernik et al., 2026). For layer-wise aggregation, we experimented with three approaches separately: scalar mixing, mean pooling, and max pooling. The last several layers are selected to prevent contam-

ination of irrelevant information encoded in the lower layer. Once we get an aggregated layer representation, we utilize either the self-attention mechanism or token-wise mean pooling to expand the sentence head into a sequence-wide representation.

Given a BERT-based encoder with T number of intermediate layers, we write $\mathbf{H} \in \mathbb{R}^{k \times l \times d}$ for the tensor of item sequence's hidden states of multiple layers, where k is the number of chosen layers, l is the sequence length, and d is the hidden size. Let $h_i \in \mathbb{R}^{l \times d}$ denote the sequence representation at the i-th layer.

Scalar Mixing Learn softmax weights s_i to dynamically compute a weighted sum of hidden states h_i across k layers of the network, where γ is a trainable scaling factor.

$$\mathbf{H}_{mix} = \gamma \sum_{i=T-k+1}^{T} \text{softmax}(s_i) \cdot h_i$$

For each layer, a layer normalization (Ba et al., 2016) is used to respect the fact that their activations might distribute differently to address certain aspects of the task. Specifically, we adopted PyTorch's `layer_norm()` on each layer $h_i = \text{LN}(h_i)$. After the contextually mixed representation $\mathbf{H}_{mix}$ is generated, we apply the attention mask to zero-out padding tokens in each sequence that would corrupt the loss signals.

Mean/Max layer pooling Apply a deterministic pooling operation across the depth dimension to extract the element-wise average or maximum values. Specifically, we aggregate the hidden states from the k chosen layers of the network:

$$\mathbf{H}_{mean} = \frac{1}{k} \sum_{i=T-k+1}^{T} h_i, \quad \mathbf{H}_{max} = \max_{T-k+1 \le i \le T} h_i$$

We set $k = 4$ to aggregate the deep semantic information captured in the final four encoder layers.

Self-attention To integrate sequence-wide contextual information into the global representation, we extract the [CLS] token $q_{cls} \in \mathbb{R}^{1 \times d}$ from the layer-aggregated sequence $\mathbf{H}^{(1)} \in \mathbb{R}^{l \times d}$ to serve as the single query, while the entire sequence $\mathbf{H}^{(1)}$ serves as the keys and values. We write self-attention aggregated representation $\mathbf{h}_{attn} \in \mathbb{R}^d$ as:

$$\begin{aligned} \mathbf{h}_{attn} &= \text{Attention}(q_{cls}, \mathbf{H}^{(1)}, \mathbf{H}^{(1)}) \\ &= \text{Weight} \cdot \mathbf{H}^{(1)} \cdot W_V \end{aligned}$$

where the attention weights are defined as

$$\text{Weight} = \text{softmax}\left(\frac{(q_{cls} \cdot W_Q)(\mathbf{H}^{(1)} \cdot W_K)^\top}{\sqrt{d}} \right)$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d \times d}$ are trainable projection matrices. Here, Weight $\in \mathbb{R}^{1 \times l}$ represents the attention weights between the [CLS] token and each token in the item sequence. This mechanism enables the global representation to selectively attend to relevant segments of the source, context, and clue components, producing a contextualized representation that informs the final prediction.

Mean token pooling Same as the depth-wise mean pooling, it applies a deterministic token-wise mean pooling operation that aggregates the sequence into a unified representation. It computes the element-wise average of the hidden states across all existing tokens in the sequence. A binary attention mask $m \in \{0, 1\}^l$ is applied to filter out padding tokens, where $m_j = 1$ for valid tokens and 0 for padded tokens. The mean-pooled sequence representation $\mathbf{h}_{mean} \in \mathbb{R}^d$ is given as:

$$\mathbf{h}_{mean} = \frac{1}{\sum_{j=1}^{l} m_j} \sum_{j=1}^{l} m_j \cdot h_j$$

3.2 Multi-task Cascade Learning

To capture the interdependency between syntax and lexical complexity, the system employs a multi-task learning (MTL) strategy utilizing hard parameter sharing. Instead of concatenating the one-hot POS labels with continuous deep representation, which introduces numerical incompatibility, our model is jointly supervised by the primary difficulty prediction task and the auxiliary POS tagging task. In this way, the auxiliary task training loss introduces a strong inductive bias for regularizing the training of shared hidden layers (Liu et al., 2019).

Furthermore, rather than deploying two isolated task heads on top of the shared hidden layers, our system utilizes a cascade architecture for hierarchical feature injection (Sanh et al., 2018; Aziz et al.). The model first computes a C-dimensional real-valued POS logit vector $\hat{y}_{pos}$ from the auxiliary classification head, where C represents the number of POS classes. To prevent the primary continuous regression loss from introducing noisy interference into the auxiliary syntactic head, these logits are detached from the primary gradient computation. The detached syntactic signals are then concatenated with the augmented deep representation for

the regressor, p_{reg}, to form the combined vector for the final difficulty prediction:

$$X_{combined} = [p_{reg} \oplus \text{detach}(\hat{y}_{pos})]$$

Note that while both tasks share the same backbone hidden layers, their additional augmentation modules are optimized independently. Consequently, the augmented deep representation utilized for POS classification, p_{pos}, remains distinct from the regression representation p_{reg}. This structural separation allows each head to specialize in its respective objective while ensuring the learned syntactic signals act as a fixed structural hint to guide the primary difficulty estimation.

Dynamic loss For the primary difficulty regression task, we calculate the mean squared error loss L_{reg}, while applying cross-entropy loss L_{pos} for the auxiliary POS tagging task. Because these two tasks operate on different scales, a naive summation would cause one objective to dominate the gradient updates. To jointly optimize the network without the need of extensive manual search for task weights, we employ dynamic precision weighting based on homoscedastic task uncertainty (Kendall et al., 2018). The total loss adaptively scales each task's contribution:

$$\mathcal{L}_{total} = \frac{1}{2\sigma_0^2}\mathcal{L}_{reg} + \log \sigma_0 + \frac{1}{\sigma_1^2}\mathcal{L}_{pos} + \log \sigma_1$$

where σ_0 and σ_1 represent the task-dependent observation noise. In practice, we utilize the log-variance of each task to avoid zero division error.

4 Experiments

4.1 Dataset

The dataset for the shared task is taken from the "Extended KVL Dataset for NLP" released by Skidmore et al. (2025) and is organized into two independent tracks with specific data constraints.

Closed Task Contains 6,091 training, 677 development, and 748 test items for each of three source languages (German, Spanish, and Mandarin). Each item is associated with the POS label of the target English word and its GLMM score. Systems in this track are restricted to using L1-specific training data and publicly available pre-trained transformer models and their embeddings.

Open Task Permits the integration of all available data, allowing the combination of the three L1 subsets into a unified training set of 18,273 items. Systems may train on this joint training set and evaluate on the L1-specific development set and test set. No algorithmic or architectural restrictions are imposed. Because the proposed system relies exclusively on the publicly available mmBERT encoder for representation generation, it complies with the constraints of both tracks.

4.2 Setup

We fine-tuned all models utilizing the mmBERT-base architecture, which contains roughly 307M parameters. Inputs are processed using the Gemma 2 tokenizer, which is a native implementation employed by mmBERT to support multilingual pretraining (Marone et al., 2025). The augmented classification token from the sequence representation serves as the input for both the classification and regression tasks. Specifically, this 768-dimensional hidden representation is passed into a two-layer MLP prediction head consisting of an initial linear layer, a GELU activation, layer normalization, dropout, and a final linear layer to generate the output logits. Our baseline employs the ordinary single-task MLP regressor which uses only the representation of the first token of the last hidden layer on top of mmBERT and XLM-RoBERTa base.

All models were trained for 10 epochs using the AdamW optimizer with a fixed learning rate of 2e-5, a weight decay of 0.1, and 100 warmup steps, with a batch size of 64 and a dropout rate of 0.1 throughout training. The best model checkpoint was selected and saved based on the lowest RMSE on the development set, with Pearson's correlation coefficient (ρ) reported as a supplementary evaluation metric. The system was trained and validated on the provided training and development sets, and the top three performing configurations were subsequently selected for hyperparameter optimization before generating final test set predictions for each L1 as the submission (See Table 6 in the Appendix).

5 Results

This section presents the evaluation results of our mmBERT-based multi-task cascade models (MTL) for the shared task. We present four specific augmentation configurations: self-attention token aggregation (MTL+SelfAttn), mean token pooling (MTL+MeanToken), scalar mixing layer aggregation followed by self-attention token aggregation (MTL+ScalarMix+SelfAttn), and mean layer pooling followed by mean token pooling (MTL+MeanLayer+MeanToken). They were eval-

Table 1: Model performance overview in the open track.

Model	ES		DE		CN		L1 avg	
	RMSE	ρ	RMSE	ρ	RMSE	ρ	RMSE	ρ
XLM-RoBERTa base	1.198	0.783	1.166	0.786	1.034	0.804	1.133	0.791
mmBERT base	1.078	0.820	1.014	0.826	0.910	0.842	1.000	0.829
MTL	1.081	0.819	1.011	0.826	0.910	0.842	1.000	0.829
MTL+SelfAttn	1.068	**0.824**	1.002	**0.830**	**0.906**	**0.843**	**0.992**	**0.832**
MTL+MeanToken	**1.066**	**0.824**	**1.001**	**0.830**	0.910	0.841	0.993	0.832
MTL+ScalarMix+SelfAttn	1.153	0.792	1.067	0.808	0.968	0.819	1.063	0.806
MTL+MeanLayer+MeanToken	1.087	0.817	1.018	0.824	0.910	0.842	1.005	0.828

Table 2: Model performance overview in the closed track.

Model	ES		DE		CN		L1 avg	
	RMSE	ρ	RMSE	ρ	RMSE	ρ	RMSE	ρ
XLM-RoBERTa base	1.257	0.765	1.258	0.773	1.140	0.753	1.218	0.764
mmBERT base	1.138	0.797	1.069	0.804	0.950	0.825	1.033	0.812
MTL	1.126	0.802	**1.058**	**0.809**	**0.936**	**0.832**	**1.020**	**0.817**
MTL+SelfAttn	**1.120**	**0.805**	1.059	0.810	0.939	0.830	1.034	0.816
MTL+MeanToken	1.112	0.801	1.050	0.812	0.942	0.829	1.035	0.814
MTL+ScalarMix+SelfAttn	1.325	0.715	1.213	0.740	0.977	0.815	1.172	0.756
MTL+MeanLayer+MeanToken	1.117	0.805	1.061	0.810	0.943	0.830	1.028	0.816

uated on the German (DE), Spanish (ES), and Mandarin (CN) test sets across both the closed and open tracks, and compared against the reference baselines (Section 4.2). Results are summarized in Tables 1 and 2 for the open and closed tracks, respectively. Detailed comparisons across all configurations are provided in Table 4 and Table 5 in the Appendix for both tracks.

Our top configurations ranked 13th/24 for Spanish and German and 14th/26 for Mandarin in the open track, and 16th/56 for Spanish and 13th/53 for both German and Mandarin in the closed track. As expected, performance in the open track consistently exceeded that of the closed track across all languages, reflecting the advantage of access to larger and combined training data.

Across both tracks, models built on the mmBERT encoder consistently outperformed the XLM-RoBERTa baseline, indicating that improved multilingual representations provide a strong foundation for item difficulty prediction. In particular, the base mmBERT model already yielded substantial gains, with further improvements observed when incorporating selected representation aggregation strategies. Among these aggregations, methods such as self-attention token aggregation and mean layer pooling plus mean token pooling yielded consistent improvements, with the former excelling in the open track and the latter in the closed track. The mean token pooling produced comparable results. The only exception was the most complex mechanism (self-attention-based aggregation, scalar mixing). It led to performance degradation in almost all languages, particularly in the closed track.

The impact of multi-task learning with auxiliary POS prediction consistently improved the results across languages and tracks, and performance in the closed track was boosted more prominently than the open track. It suggests that syntactic information provided useful signals that contribute to the lexical difficulty prediction, with the degree of improvement depending on the data conditions. Lastly, performance patterns varied across languages. Mandarin generally exhibited larger gains from representation learning and aggregation strategies, whereas results for German and Spanish were more stable but showed smaller improvements.

6 Discussion

6.1 Effect of the mmBERT Encoder

Table 1 and 2 show that fine-tuning mmBERT encoder yielded substantially better performance than the standard XLM-RoBERTa encoder baseline across almost all languages and data conditions. In the open track, replacing the XLM-RoBERTa backbone with mmBERT reduced the average RMSE

from 1.133 to 1.042, and increased the average ρ from 0.791 to 0.815. The performance gain largely persisted in the restricted data condition of the closed track, where mmBERT reduced the overall average RMSE from 1.218 to 1.033 and increased average ρ from 0.764 to 0.812.

As the multilingual adaptation of the ModernBERT (Warner et al., 2025), mmBERT is pretrained on the 3T tokens and 1,833 languages with innovative training recipes. Its superior performance over XLM-RoBERTa may be attributed to a combination of modernized architectural enhancements, a more diverse and larger pretraining corpus, and highly optimized data scaling (Marone et al., 2025). Our results showed that mmBERT provides a better alternative to encoder-only models in cross-lingual vocabulary item difficulty prediction.

6.2 Multitask performance

Table 1 shows that in the open track, which provides an expanded joint training set, the multitask approach produced similar performance regarding the difficulty prediction objective. In the restricted data conditions (Table 2), the MTL showed some marginal improvements, with average RMSE decreasing from 1.033 to 1.020, and average ρ increasing from 0.812 to 0.817.

A further benefit of the MTL extension that we observed involves improved robustness in difficulty prediction. As shown in Figure 2, the MTL model demonstrated more consistent performance, particularly under the restricted data condition in the closed track. The incorporation of POS-tagging signals may help the model disambiguate target English words with variable semantic roles, thereby supporting more stable predictions of the word-level item difficulty.

Table 3: Fixed-effects regression results on ΔRMSE.

Predictor	Coef.	SE	t	p	95% CI
Intercept	0.452	0.382	1.183	0.238	[-0.300, 1.204]
𝕀(Closed-DE)	0.001	0.004	0.204	0.838	[-0.008, 0.009]
𝕀(Closed-ES)	-0.000	0.006	-0.009	0.993	[-0.012, 0.012]
𝕀(Open-CN)	0.014	0.004	3.121	0.002	[0.005, 0.022]
𝕀(Open-DE)	0.012	0.004	2.756	0.006	[0.003, 0.020]
𝕀(Open-ES)	0.017	0.004	4.025	< 0.001	[0.009, 0.026]
POS Accuracy	-0.478	0.392	-1.219	0.223	[-1.248, 0.293]

Despite the robustness against semantic noises, accurate POS tagging did not show statistically significant correlations with difficulty prediction improvement. We operationalized the performance improvement as the RMSE difference on the test set between each pair of the MTL model and the base model (i.e. $\Delta\text{RMSE} = \text{RMSE}_{\text{MTL}} - \text{RMSE}_{\text{Base}}$) across tracks and L1s.

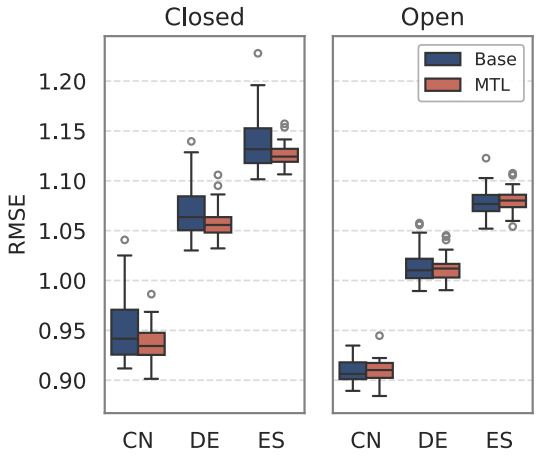

Figure 2: Test RMSE performance between mmBERT base model and its MTL variant across tracks and L1s

The regression plots in Figure 3 did not reveal a consistent linear relationship between POS prediction accuracy and ΔRMSE across tracks and L1 conditions. Although slight negative trends appeared for some language–track combinations (e.g., CN and DE), the overall patterns were weak and inconsistent, with substantial variability around the fitted regression lines. To further isolate the effect of POS accuracy on RMSE differences, we conducted a fixed-effects linear regression controlling for track and L1 conditions. The results (Table 3) likewise did not support a significant association between POS prediction accuracy and RMSE improvement. One possible explanation is that word-level POS tagging provides only limited linguistic information, making it insufficient to substantially enhance the difficulty prediction objective.

In general, MTL learns a more robust and generalized representation on related tasks by ignoring data-dependent noise and reduces overfitting risks (Ruder, 2017; Goldberg, 2022), which is especially beneficial with restricted data in the closed track. However, the benefit to the primary difficulty prediction objective appears to be limited as the MTL model did not yield significantly better RMSE results than the baseline model. This may be because the auxiliary POS tagging task is relatively simple and may not provide sufficiently rich linguistic information to substantially improve word-level difficulty prediction. It suggested that POS tagging is a related task that can be jointly trained

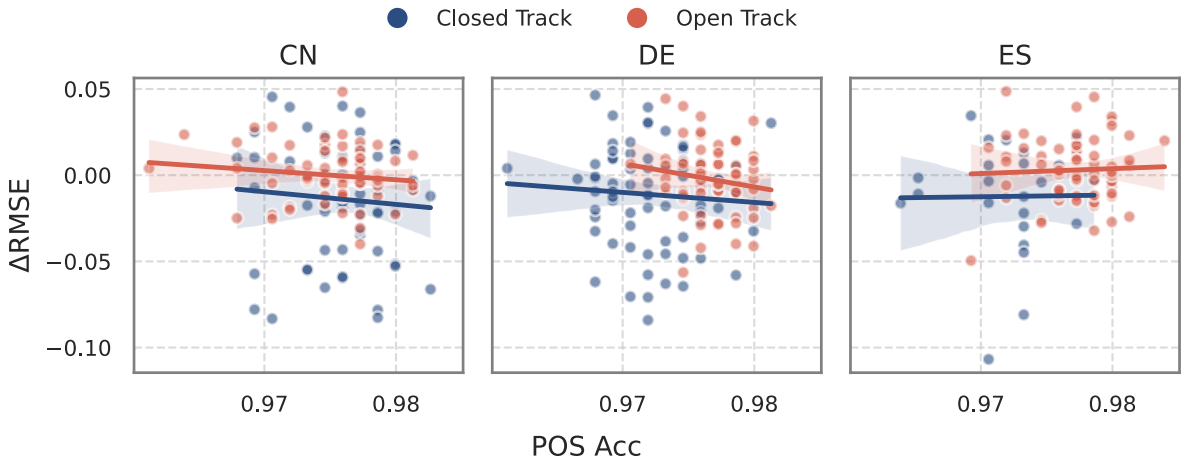

Figure 3: The regression plot between RMSE improvement and POS Acc across tracks and L1s. The x-axis refers to the POS prediction accuracy by MTL model, while the y-axis refers to the difference between RMSEs scored by the MTL model and the base model on the same L1 test set. Color stands for which track has been used as the training set.

with lexical difficulty prediction without introducing conflicting signals. This aligns with the finding that different word senses of a polysemy contribute differently to the difficulty of the vocabulary item (Skidmore et al., 2025).

6.3 Mixed Outcomes by Representation Augmentation

Systematic comparison of configurations revealed mixed outcomes by hidden states aggregation (see Table 4 and Table 5). Prior studies demonstrated that allowing the classification token to attend to the full sequence via an additional multi-head attention module consistently enhances the standard BERT baseline, with strong benefits observed in low-resource scenarios (Behrendt et al., 2025). Consistently, our results indicated that token-wise self-attention aggregation alone provided a slight performance boost, while it severely conflicted with layer-wise aggregation methods. A possible explanation is that naively aggregating intermediate layers through simple deterministic operations, such as mean/max pooling, failed to adapt to task-specific requirements and introduced severe optimization instability during training (Ciernik et al., 2026).

To mitigate this instability, scalar mixing computes a weighted average of hidden layers by dynamically determining the importance of each layer during the training. In theory, it would allow optimal exploitation of deeper information and surpass naive deterministic pooling methods (Gombert et al., 2024). However, our results showed a contradicting image, with the scalar mixing being the worst approach. An investigation into the module showed that all trained scalar weights s collapsed into a delta distribution with a single layer sticking out and all else muted. Also, the scaling factor γ either exploded or vanished. This indicated that the training with scalar mixing virtually failed to learn meaningful weights as expected.

Moreover, the simple mean token pooling approach only trailed the self-attention approach by a small margin. Most importantly, our result indicated that it worked smoothly with other deterministic layer aggregation approaches, especially with limited data. This suggested mean token pooling as a robust and efficient alternative.

7 Conclusions

In this paper, we examined the precalibration of vocabulary item difficulty in multilingual settings using transformer-based representations and multi-task learning. Across both open and closed tracks, our findings showed that mmBERT provided a stronger foundation than the shared-task XLM-RoBERTa baseline. Moreover, the auxiliary POS-based multi-task learning further improved the performance. At the same time, the benefits of additional architectural complexity, including layer-wise and token-wise aggregation, were not uniform across language and data conditions.

8 Limitations

As discussed in section 6.3, while some representation aggregation strategies offered promising results, their improvements over the baseline were context-dependent and not uniform. This indicates that, within the multi-task learning regime for this specific task, there may not be a universally superior representation aggregation strategy. Consequently, any future research seeking to utilize internal encoder states may benefit from extensive and systematic model selection to identify the optimal configuration. Also, the reason behind our failed training with scalar mixing needs to be further researched. Given that the performance gains are marginal and unstable, the standard last-layer [CLS] approach still remained a highly competitive and computationally efficient default.

Second, our study largely relied on a multilingual encoder that provides a language-agnostic foundation for cross-lingual transfer. However, we did not evaluate language-specific monolingual encoders. Previous studies indicated that specialized monolingual models may capture distinct, language-dependent morphological and syntactic patterns more effectively (Ortiz-Zambrano et al., 2025; Skidmore et al., 2025). The absence of experiments using language-specific encoders, such as Chinese ModernBERT (Zhao et al., 2025) or ModernGBERT (Wunderle et al., 2025), limits our understanding of whether specialized pre-training could yield superior lexical difficulty estimations. This limitation is particularly relevant for the closed track, where models were trained and evaluated within each language space.

References

David Alfter and Elena Volodina. 2018. Towards single word lexical complexity prediction. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 79–88. Association for Computational Linguistics.

Samah AlKhuzaey, Floriana Grasso, Terry R Payne, and Valentina Tamma. 2024. Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, 34(3):862–914.

Gokul Aradelli. 2019. Transformers for question difficulty estimation from text.

Abdul Aziz, Md. Akram Hossain, Abu Nowshed Chy, Md. Zia Ullah, and Masaki Aono. Leveraging contextual representations with BiLSTM-based regressor for lexical complexity prediction. 5:100039.

Jimmy Ba, J. Kiros, and Geoffrey E. Hinton. 2016. Layer Normalization.

Maike Behrendt, Stefan Sylvius Wagner, and Stefan Harmeling. 2025. MaxPoolBERT: Enhancing BERT Classification via Layer- and Token-Wise Aggregation. *arXiv preprint*.

Luca Benedetto, Paolo Cremonesi, Andrew Caines, Paula Buttery, Andrea Cappelli, Andrea Giussani, and Roberto Turrin. 2023. A survey on recent approaches to question difficulty estimation from text. *ACM Computing Surveys*, 55(9):1–37.

Okan Bulut, Guher Gorgun, and Bin Tan. 2024. Item difficulty and response time prediction with large language models: An empirical analysis of usmle items. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 522–527.

Laure Ciernik, Marco Morik, Lukas Thede, Luca Eyring, Shinichi Nakajima, Zeynep Akata, and Lukas Muttenthaler. 2026. Beyond the final layer: Attentive multilayer fusion for vision transformers. *arXiv preprint*.

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. *arXiv preprint*.

Ali Derakhshan and Ehsan Karimi. 2015. The interference of first language and second language acquisition. *Theory and Practice in Language Studies*, 5(10):2112–2117.

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: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Omar Galal, Ahmed H. Abdel-Gawad, and Mona Farouk. 2024. Rethinking of BERT sentence embedding for text classification. *Neural Computing and Applications*, 36:20245–20258.

Yoav Goldberg. 2022. Cascaded, multi-task and semi-supervised learning. In *Neural Network Methods for Natural Language Processing*, pages 235–250. Springer.

Sebastian Gombert, Lukas Menzel, Daniele Di Mitri, and Hendrik Drachsler. 2024. Predicting item difficulty and item response time with scalar-mixed transformer encoder models and rational network regression heads. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 483–492.

Charles R. Graham, James McGhee, Brenton Millard, Matthew T. Prior, Yasuko Watanabe, and Sun Lee. 2010. The role of lexical choice in elicited imitation item difficulty. In *Selected Proceedings of the 2008 Second Language Research Forum*, pages 57–72, Somerville, MA. Cascadilla Proceedings Project.

Ivana Kapounová. 2025. Predicting item difficulty by applying machine learning algorithms using item text features.

Abdelhak Kelious, Matthieu 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.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. *arXiv preprint*.

Yifan Liu, Bohan Zhuang, Chunhua Shen, Hao Chen, and Wei Yin. 2019. Auxiliary Learning for Deep Multi-task Learning. *arXiv preprint*.

Ekaterina Loginova, Luca Benedetto, Damien Benoit, and Paolo Cremonesi. 2021. Towards the application of calibrated transformers to the unsupervised estimation of question difficulty from text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 846–855. INCOMA Ltd.

Hirofumi Maeda. 2025. Field-testing multiple-choice questions with ai examinees: English grammar items. *Educational and Psychological Measurement*, 85(2):221–244.

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

Arianna D. McCarthy, Kevin P. Yancey, Gregory T. LaFlair, Jesse Egbert, Mengxuan Liao, and Burr Settles. 2021. Jump-starting item parameters for adaptive language tests. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 883–899. Association for Computational Linguistics.

Kathryn North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical complexity prediction: An overview. *ACM Computing Surveys*, 55(9):1–42.

Juan A. Ortiz-Zambrano, Carlos H. Espín-Riofrío, and Arturo Montejo-Ráez. 2025. Deep encodings vs. linguistic features in lexical complexity prediction. *Neural Computing and Applications*, 37(3):1171–1187.

Cheng Pan, Bing Song, Shuo Wang, and Zhen Luo. 2021. Deepblueai at semeval-2021 task 1: Lexical complexity prediction with a deep ensemble approach. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 578–584. Association for Computational Linguistics.

Sydney Peters, Nan Zhang, Hong Jiao, Ming Li, Tianyi Zhou, and Robert Lissitz. 2025. Text-based approaches to item difficulty modeling in large-scale assessments: A systematic review. *arXiv preprint arXiv:2509.23486*.

G. V. R. Ram and A. Kesanam. 2024. Leveraging physical and semantic features of text item for difficulty and response time prediction of usmle questions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 534–541. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Victor Sanh, Thomas Wolf, and Sebastian Ruder. 2018. A Hierarchical Multi-task Approach for Learning Embeddings from Semantic Tasks. *arXiv preprint*.

Christoph Schneider, Jing Chen, and James Heneger. 2026. Text complexity versus task complexity: Item difficulty modeling for reading items. *Practical Assessment, Research, and Evaluation*, 31(1).

Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine learning–driven language assessment. *Transactions of the Association for computational Linguistics*, 8:247–263.

Matthew Shardlow, Richard Evans, and Marcos Zampieri. 2022. Predicting lexical complexity in english texts: The complex 2.0 dataset. *Language Resources and Evaluation*, 56(4):1153–1194.

Lucy Skidmore, Mariano Felice, and Karen J Dunn. 2025. Transformer architectures for vocabulary test item difficulty prediction.

Esther Ulitzsch, Dmitry Belov, Oliver Luedtke, and Alexander Robitzsch. 2026. Using item parameter predictions for reducing calibration sample requirements—a case study based on a high-stakes admission test. *Journal of Educational Measurement*, 63(1):e12426.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, and 1 others. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast,

memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547.

Kelly Wauters, Piet Desmet, and Wim Van Den Noortgate. 2012. Item difficulty estimation: An auspicious collaboration between data and judgment. *Computers & Education*, 58(4):1183–1193.

Julia Wunderle, Anton Ehrmanntraut, Jan Pfister, Fotis Jannidis, and Andreas Hotho. 2025. New Encoders for German Trained from Scratch: Comparing ModernGBERT with Converted LLM2Vec Models. *arXiv preprint*.

Victoria Yaneva, Kathryn North, Peter Baldwin, Le An Ha, Sara Rezayi, Yufan Zhou, and Brian Clauser. 2024. Findings from the first shared task on automated prediction of difficulty and response time for multiple-choice questions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 470–482. Association for Computational Linguistics.

Hua Yang and EUM Suyong. 2018. Feature analysis on english word difficulty by gaussian mixture model. In *2018 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 191–194. IEEE.

Talal B. Yaseen, Qusai Ismail, Sara Al-Omari, Eyad Al-Sobh, and Mohammad 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. Association for Computational Linguistics.

George-Eduard Zaharia, Dumitru-Clementin Cercel, and Mihai Dascalu. 2021. Upb at semeval-2021 task 1: Combining deep learning and hand-crafted features for lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 609–616.

Zeyu Zhao, Ningtao Wang, Xing Fu, and Yu Cheng. 2025. Chinese ModernBERT with Whole-Word Masking. *arXiv preprint*.

A Appendix

Table 4: Comparison of multi-task mmBERT models with variant augmentations in the open track.

Model	RMSE	ρ
baseline	0.910	0.842
SelfAttn	**0.906**	**0.843**
ScalarMix+SelfAttn	0.968	0.819
MeanLayer+SelfAttn	0.938	0.833
MaxLayer+SelfAttn	0.918	0.840
MeanToken	0.910	0.841
ScalarMix+MeanToken	0.980	0.813
MeanLayer+MeanToken	0.910	0.842
MaxLayer+MeanToken	0.918	0.839

(a) Open Track: Mandarin

Model	RMSE	ρ
baseline	1.011	0.827
SelfAttn	1.002	**0.830**
ScalarMix+SelfAttn	1.067	0.808
MeanLayer+SelfAttn	1.054	0.813
MaxLayer+SelfAttn	1.057	0.814
MeanToken	**1.001**	**0.830**
ScalarMix+MeanToken	1.053	0.811
MeanLayer+MeanToken	1.018	0.824
MaxLayer+MeanToken	1.013	0.826

(b) Open track: German

Model	RMSE	ρ
baseline	1.081	0.819
SelfAttn	1.068	**0.824**
ScalarMix+SelfAttn	1.153	0.792
MeanLayer+SelfAttn	1.141	0.802
MaxLayer+SelfAttn	1.090	0.820
MeanToken	**1.066**	**0.824**
ScalarMix+MeanToken	1.134	0.798
MeanLayer+MeanToken	1.087	0.817
MaxLayer+MeanToken	1.078	0.821

(c) Open track: Spanish

Table 5: Comparison of multi-task mmBERT models with variant augmentations in the closed track.

Model	RMSE	ρ
baseline	**0.936**	**0.832**
SelfAttn	0.939	0.830
ScalarMix+SelfAttn	0.977	0.815
MeanLayer+SelfAttn	0.963	0.824
MaxLayer+SelfAttn	0.999	0.808
MeanToken	0.942	0.829
ScalarMix+MeanToken	1.088	0.764
MeanLayer+MeanToken	0.943	0.830
MaxLayer+MeanToken	0.952	0.825

(a) Closed Track: Mandarin

Model	RMSE	ρ
baseline	1.058	0.804
SelfAttn	1.059	0.810
ScalarMix+SelfAttn	1.213	0.740
MeanLayer+SelfAttn	1.123	0.781
MaxLayer+SelfAttn	1.093	0.797
MeanToken	**1.050**	**0.812**
ScalarMix+MeanToken	1.072	0.805
MeanLayer+MeanToken	1.061	0.810
MaxLayer+MeanToken	1.052	**0.812**

(b) Closed track: German

Model	RMSE	ρ
baseline	**1.126**	**0.802**
SelfAttn	1.120	**0.805**
ScalarMix+SelfAttn	1.325	0.715
MeanLayer+SelfAttn	1.404	0.675
MaxLayer+SelfAttn	1.177	0.782
MeanToken	1.112	0.801
ScalarMix+MeanToken	1.138	0.798
MeanLayer+MeanToken	1.117	**0.805**
MaxLayer+MeanToken	**1.115**	0.807

(c) Closed track: Spanish

Table 6: Shared task performance and hyperparameter results of submitted models. Models were fine-tuned on the train set and evaluated on the development set.

Configuration	**Track**	**L1**	**Learning rate**	**Weight decay**	**Warmup step**	**RMSE**	ρ
MTL+SelfAttn	open	CN	4e-05	0.0	100	0.885	0.851
MTL+SelfAttn	open	DE	4e-05	0.0	100	1.053	0.829
MTL+SelfAttn	open	ES	4e-05	0.0	100	1.002	0.830
MTL+MeanToken	open	CN	4e-05	0.1	200	0.895	0.847
MTL+MeanToken	open	DE	4e-05	0.1	200	1.056	0.829
MTL+MeanToken	open	ES	4e-05	0.1	200	0.990	0.834
MTL+MeanLayer+MeanToken	open	CN	5e-05	0.1	200	0.899	0.845
MTL+MeanLayer+MeanToken	open	DE	5e-05	0.1	200	1.061	0.826
MTL+MeanLayer+MeanToken	open	ES	5e-05	0.1	200	1.009	0.827
MTL+MaxLayer+MeanToken	closed	CN	4e-05	0.1	100	0.905	0.843
MTL+MeanToken	closed	CN	5e-05	0.0	100	0.932	0.832
MTL+SelfAttn	closed	CN	5e-05	0.0	100	1.185	0.710
MTL+MeanLayer+MeanToken	closed	DE	5e-05	0.0	200	1.011	0.827
MTL+MaxLayer+MeanToken	closed	DE	5e-05	0.0	200	1.032	0.818
MTL	closed	DE	4e-05	0.0	100	1.039	0.815
MTL	closed	ES	5e-05	0.1	200	1.084	0.818
MTL+MeanToken	closed	ES	4e-05	0.1	200	1.114	0.807
MTL+SelfAttn	closed	ES	3e-05	0.0	200	1.142	0.796

Search space for hyperparameters: learning rate (1e-5, 2e-5, 3e-5, 4e-5, 5e-5), weight decay (0, 0.1), warmup step (100, 200).