Findings of the BEA 2026 Shared Task on Vocabulary Difficulty Prediction for English Learners

Mariano Felice and Lucy Skidmore
English Language Research & Impact
British Council, UK
name.surname@britishcouncil.org

Abstract

This paper reports findings from the BEA 2026 Shared Task on Vocabulary Difficulty Prediction for English Learners across three L1s (Spanish, German and Mandarin). The task featured open and closed tracks, using data from the British Council's Knowledge-based Vocabulary Lists (KVL). Submissions were received from 23 teams employing diverse modelling approaches, including transformers, Large Language Models, feature-based approaches and ensembles. Results were evaluated using RMSE, with winning systems significantly exceeding the baseline and establishing new state-of-the-art benchmarks. This paper offers an examination of the participating systems, performance across tracks and L1s, and the factors that can affect prediction accuracy.

1 Introduction

Vocabulary is a crucial aspect of language knowledge, shaping what learners can understand and produce. Establishing the difficulty of vocabulary is therefore essential for creating level-appropriate content and developing valid, reliable assessment instruments (Schmitt et al., 2020; François and Gala, 2024). However, determining word difficulty still relies on labour-intensive processes involving expert judgment and costly pretesting, which limits scalability and slows innovation (Dunn et al., 2026). As learning and assessment increasingly rely on digital platforms, the need for more efficient and scalable solutions is more pressing than ever.

While previous shared tasks have explored related problems such as Complex Word Identification (Paetzold and Specia, 2016; Yimam et al., 2018), Lexical Complexity Prediction (Shardlow et al., 2021), Lexical Simplification (Shardlow et al., 2024) and Item Difficulty Estimation (Yaneva et al., 2024), they were not designed with English language learners in mind and did not explore the influence of learners' L1 on L2 vocabulary difficulty. This shared task addresses that gap by bringing vocabulary difficulty prediction into the domain of language learning and assessment, offering a new perspective on a problem that has traditionally relied on psychometric calibration methods. In doing so, it serves as a testbed for state-of-the-art NLP models, with findings that have the potential to inform AI-powered solutions for item writing, content generation, adaptive testing, and personalised vocabulary learning.

2 Related Research

In the language learning and assessment domain, NLP-based vocabulary difficulty prediction spans a range of approaches, each with its own framing of 'difficulty'. Such work includes CEFR level classification of vocabulary in isolation (Alfter and Volodina, 2018; Settles et al., 2020) and in context (Aleksandrova and Pouliot, 2023; Bannò et al., 2025), classification of vocabulary difficulty using crowd-sourced ratings as labels (Degraeuwe, 2025) as well as prediction of continuous difficulty values derived from testing learners (Skidmore et al., 2025). This shared task follows the approach of the latter, which sits at the intersection of two areas of research in NLP: Question Difficulty Estimation from Text and Lexical Complexity Prediction.

Question Difficulty Estimation from Text (QDET) aims to predict the difficulty of test items directly from their text, without the need for response data. A growing body of research has explored QDET for high-stakes assessment, driven by its advantages in efficiency and scalability over traditional psychometric approaches (AlKhuzaey et al., 2024). While the latest methods in NLP have been successfully applied in domains related to content knowledge assessment (Yaneva et al., 2024), QDET approaches for language assessment, where difficulty is more closely tied to the linguistic properties of the item, still largely rely on sim-

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 964–984

pler models (AlKhuzaey et al., 2024; Peters et al., 2025). This is particularly true for L2 vocabulary item difficulty estimation, which commonly features 'vocabulary-only' datasets that are not well-suited to the latest contextualised word embedding models (Benedetto et al., 2023).

Lexical Complexity Prediction (LCP) is an extension of Complex Word Identification (CWI), which concerns the automatic detection of complex words, primarily to support downstream text simplification (North et al., 2025). LCP expands the binary classification task of CWI to regression, assigning each word a continuous complexity score. The definition of complexity varies by domain, ranging from crowd-sourced ratings to morphosyntactically derived values (North et al., 2023). The majority of LCP and CWI research focuses on L1 vocabulary complexity. However, in line with the increasing body of work advocating for personalised LCP (Lee and Yeung, 2018; Gooding and Tragut, 2022), where demographic features such as L1 background are accounted for, LCP for L2 learners is an increasingly popular focus of research (North and Zampieri, 2023; Degraeuwe, 2025).

3 Shared Task Description

This shared task was organised by researchers at the British Council as part of the BEA Workshop, held at ACL in July 2026. The task addresses vocabulary difficulty prediction for learners of English with diverse first language (L1) backgrounds. Building on prior work (Skidmore et al., 2025), this shared task framed vocabulary difficulty prediction as a regression problem, where systems predict the difficulty of a target English word given a vocabulary test item written in the learner's L1. Data for the task were drawn from the British Council's Knowledge-based Vocabulary Lists (KVL) (Schmitt et al., 2021, 2024), a multilingual dataset containing such items with psychometrically calibrated difficulty scores for learners from three L1 backgrounds: Spanish (ES), German (DE) and Mandarin (CN).

Participation was open to the research community following a limited timeline. Training and development data alongside baseline models were released on 26th January 2026, giving participants approximately eight weeks to develop their systems before test data were made available on 20th March. Teams then had one week to submit their final system outputs. Systems were evaluated on the test set by the organisers, with results announced on 3rd April. Further details on the tracks, data, baseline models and evaluation are outlined below and on the shared task website[1].

3.1 Tracks

The shared task included two tracks:

Closed Track: Designed to elicit models that rely on the provided dataset and standard, publicly available NLP tools. The goal was to explore optimal model design under controlled data conditions.

Data: Systems were permitted to only use the provided training data for each corresponding L1. Additional training data or combining data from different L1s was not allowed.
Models, tools and databases: Publicly available, 'off-the-shelf' pre-trained transformer models (e.g. BERT, RoBERTa) and their embeddings, standard NLP tools (e.g. taggers, parsers) and publicly available linguistic databases (e.g. WordNet, frequency lists) were allowed.
LLMs: Publicly available large language models were permitted strictly as a tool to extract features from the item text. For example, generating scalar features (e.g. imageability ratings), textual features (e.g. English translation of the L1 source), or deriving embeddings.

Open Track: The 'anything goes' category, which allowed for maximum flexibility and the use of external resources. The purpose of this track was to encourage creativity and explore the full potential of current AI technology.

Data: Any additional training data was permitted. This included public corpora, proprietary datasets, synthetic data, or combining different L1 subsets.
Models, tools and databases: No restrictions.
LLMs: Any system using an LLM to generate a difficulty prediction or identify the target word was classified as open track, regardless of whether the LLM output was used directly or passed to a downstream model.

3.2 Data

The dataset for the task was derived from the British Council's Knowledge-based Vocabulary Lists (KVL) (Schmitt et al., 2021, 2024), initially developed to collate difficulty rankings of English vocabulary for learners with L1 backgrounds of Spanish, German and Mandarin. The productive

[1] britishcouncil.org/data-science-and-insights/bea2026st

English word knowledge of over 100,000 learners was assessed using items testing form-based recall of individual lemmas in a translation format (Laufer and Goldstein, 2004). From approximately 3.3 million responses, difficulty estimates were derived separately for each L1, applying random-item-random-person (RPRI) Rasch models (De Boeck, 2008) built within a generalised linear mixed model (GLMM) framework (Dunn, 2024).

Below is an example test item in Spanish, where learners were required to input the remainder of the target English word "house"[2]:

casa Vivo en una casa grande que tiene tres dormitorios.
h _ _ _ _

For the shared task, participants were provided the 'Extended KVL Dataset for NLP' (Skidmore et al., 2025) for model training and development. This dataset contains 6,768 parallel English vocabulary test items, with prompts in the three L1s, totalling 20,304 items. The dataset was split into train and development subsets to facilitate comparison to baseline models for the task, however participants were allowed to use this development data to train their models if desired. For final model evaluation, an additional 748 vocabulary test items per L1 (2,244 total) were released. Table A5 in the Appendix summarises the dataset split. The dataset contained the following columns:

- `item_id`: An ID number from 1 to 6,768. Items with the same `item_id` across different L1s share the same English target word).
- `L1`: The L1 of the prompt ('es', 'de', or 'cn').
- `en_target_word`: The English target word.
- `en_target_pos`: The part of speech of the English target word.
- `en_target_clue`: A partial-spelling clue of the English target word.
- `L1_source_word`: The corresponding L1 source word(s).
- `L1_context`: The L1 contextualising prompt.
- `GLMM_score`: The GLMM difficulty estimate for the vocabulary test item. A lower score indicates a more difficult word.

Details for how to access the shared task data can be found on the British Council's website.

[2]The German and Mandarin versions had similar, yet distinct prompts, for example in German: "Haus Ich wohne in einem Haus mit Garten." And in Mandarin: "房子 我买了一座房子。"

3.3 Baseline Models

The baseline models used for the task are multilingual XLM-RoBERTa-base transformer encoders (Conneau et al., 2020), fine-tuned on the training subset of the data following (Skidmore et al., 2025). The input text includes the question in the same order as it is presented to the test takers (L1 source word, L1 context, EN clue), followed by the target answer (EN target word), as shown in the example:

```
casa </s> Vivo en una casa grande que
tiene tres dormitorios. </s> h____
</s> house
```

Four baseline models were provided to participants, three in the closed track (`baseline_closed_es`, `baseline_closed_de` and `baseline_closed_cn`) which were fine-tuned on the L1 subsets of the data separately, and one in the open track (`baseline_open_xx`) which was fine-tuned on all of the L1 subsets combined. These models are available via GitHub.[3] Hyperparameters for the baseline models can be found in Table A6 in the Appendix.

3.4 Evaluation

Submissions were evaluated using two metrics: Root Mean Squared Error (RMSE) and Pearson correlation, chosen for comparability with prior work in LCP and QDET research (North et al., 2023; AlKhuzaey et al., 2024). RMSE was used as the primary metric for the official rankings while Pearson correlation provided a scale-independent measure to facilitate cross-L1 comparison.

4 Participation in the Shared Task

4.1 Teams and Submissions

In total, 23 teams participated in the shared task, from organisations across 16 countries: Bangladesh, Belgium, China, Germany, India, Indonesia, Ireland, Japan, Mexico, Romania, South Korea, Switzerland, United Arab Emirates, United Kingdom, United States of America and Uruguay. Teams were invited to submit up to three 'runs' per track and L1. The distribution of teams and runs is summarised in Table 1. Almost all teams submitted to the closed track (22 out of 23), with eight of those teams also submitting to the open track. One team (SurreyCTS) submitted only to the open track. Most teams provided submissions for all three L1s,

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

	Open		Closed	
L1	**Teams**	**Runs**	**Teams**	**Runs**
Spanish	8	23	21	55
German	8	23	20	52
Mandarin	9	25	21	52
Total	9	71	22	159

Table 1: Number of participating teams and runs.

whereas two teams (AIDA and BZPT) focused on just one L1, Spanish and Mandarin, respectively. Table A7 in the Appendix provides a breakdown of the submitted runs per team.

4.2 Summary of Approaches

Out of the 23 participating teams, 19 published an accompanying system description paper. Table A8 in the Appendix reports the tracks, L1s, approaches, and primary models explored, where submissions are tagged with up to four attributes: **transformers** (transformer-encoder models fine-tuned for prediction), **LLMs** (LLMs fine-tuned for prediction), **features** (any use of features), and **ensemble** (any combination of the above models).

Fine-tuning encoder-only transformer models was the most popular approach, explored by 17 out of 19 teams. The most widely used model family was XLM-RoBERTa (Conneau et al., 2020), particularly XLM-RoBERTa-large, which was used by eight teams, while RemBERT (Chung et al., 2021) and DeBERTA-v3 (He et al., 2023) were also popular choices. Of note, three teams built systems using mmBERT-base, a newer multilingual encoder-only transformer recently shown to outperform the previous generation of multilingual models (Marone et al., 2025). Encoders pretrained for related tasks were also explored, including COMET (Rei et al., 2020) (SurreyCTS) for machine translation quality estimation as well as the sentence embedding models Multilingual E5 (Wang et al., 2024) and BGE-M3 (Chen et al., 2024) (uogal, UOL@IDEM). At the architectural level, some teams moved beyond standard CLS-token pooling, experimenting with methods such as attention-weighted pooling, multi-layer concatenation of hidden states, and target-word-specific pooling (AIDA, Failure, SurreyCTS). Two teams experimented with modified training objectives: Jinnie's Lab employed multi-task learning, jointly predicting difficulty alongside an auxiliary POS tag; TOEBM employed a multi-objective training framework, supplementing the standard regression loss with contrastive objectives designed to better align the learned representations with the task structure.

Fine-tuning LLMs was explored by three of the open-track teams (Glite, Sakura, TeamXBC). All three systems used Low-Rank Adaptation (LoRA or QLoRA) (Hu et al., 2022) to fine-tune models from several decoder-only families including Qwen (Yang et al., 2024), Mistral (Jiang et al., 2023; Liu et al., 2026), LLaMA (Dubey et al., 2024), and GLM (Team GLM, 2024). Sakura additionally introduced a novel technique to fine-tune LLMs for continuous value prediction using soft targets.

Feature engineering was also widespread, with 15 of the 19 teams incorporating features into their models, typically fused architecturally with transformer representations (e.g. EduNLP, Failure) or used as inputs to gradient-boosted regressors (e.g. RETUYT-InCo). Some teams applied novel ways to integrate features: AIDA used features to calibrate transformer outputs by predicting their residual error, while uogal augmented transformer inputs with NMT-generated translations of L1 contexts as a complementary cross-lingual signal. Beyond these integration strategies, the features themselves broadly fell into three groups: target word, cross-lingual, and test item. Word-level features included surface properties (e.g. word length), psycholinguistic predictors (e.g. age of acquisition), and L2 learner-specific resources (e.g. CEFR levels) (Boosted Cats, SATLab). Cross-lingual features captured relationships between the L1 source word or context and the English target, most commonly with features related to orthographic and semantic similarity (e.g. RETUYT-InCo). Test item features captured difficulty related to the item content, such as the proportion of letters revealed by the spelling clue (e.g. Data Asgardians, EduNLP) and how recoverable the English target is from the L1 context (UOL@IDEM). Beyond traditional corpus-based and lexical resources, two notable extraction strategies were applied. The first involved probing language models as proxy test-takers, recording signals such as token surprisal and masked token prediction probability (Glite, Sakura). The second used LLMs as direct judges, prompting them to rate aspects of difficulty such as the item's quality (Glite) or the degree of lexical ambiguity (Sakura).

Model ensembling was adopted by 15 of the 19 teams. The most common strategy was averaging across multiple runs of models from the same

architecture, primarily transformer encoders (e.g. SAAKTH, Token Titans). Exceptions included an array of classical regressors (UGA Threshold) and an ensemble of fine-tuned LLMs (Sakura). Many teams combined architecturally diverse models, typically pairing transformers with feature-based models and weighting predictions toward the transformer (e.g. Data Asgardians, NLP-Explorers). Teams explored stacking ensembles, using a meta-learner such as linear regression or a shallow MLP to combine predictions (e.g. TOEBM, uogal). Three teams (Glite, Sakura, SATLab) used meta-learners that consumed both model predictions and handcrafted item-level features.

Open-track teams took advantage of the relaxed restrictions in several other ways. Three teams fine-tuned decoder-only LLMs for difficulty prediction. Five teams trained models jointly on all three L1s rather than on a separate per-L1 basis. Glite made use of a proprietary sense-level lexical resource with difficulty estimates. One team (uogal) experimented with pretraining on EFCAMDAT (Geertzen et al., 2013), a corpus of L2 English learner essays, though the resulting models were not included in their final system. Moreover, Glite and Sakura used commercial LLM APIs to generate features. Two teams made notable use of agentic AI: TeamXBC used AI coding agents to build their ML solution with varying degrees of human input, while Glite's autonomous research framework enabled the systematic evaluation of over 1,000 candidate features across 270+ tracked experiments.

5 Results

The official leaderboards for each track and L1 are included in Appendix B, where we report RMSE and Pearson correlation for each run. Given the small differences in RMSE across systems, we performed a statistical significance analysis to identify groups of systems with equivalent performance. Significance testing was conducted using a non-parametric bootstrap method (Efron and Tibshirani, 1993) with 10,000 iterations. To account for multiple comparisons, we applied a Bonferroni correction, setting the significance threshold to $\alpha_{adj} = 0.05/(n-1)$, where n is the number of submissions within a given track and L1. Systems were grouped sequentially, such that a new group was created only when a system performed significantly worse than the current group's leader. A summary of the results including the best entry per team is included in Table 2. Results show that for all tracks and L1s, most teams were able to beat the baseline with at least one of their submissions, often by a large margin.

The winning team for the closed track was Glite (Philippov et al., 2026), whose submitted runs took the top three positions across all three L1s. The best runs achieved an RMSE of 0.903, 0.885 and 0.776 for Spanish, German and Mandarin, respectively, with corresponding Pearson correlations of 0.877, 0.871 and 0.889. The Glite systems used a per-L1 CatBoost regressor (Prokhorenkova et al., 2018), combining scalar out-of-fold predictions from fine-tuned encoder models with hand-crafted features. These were selected from a large candidate pool via Recursive Feature Elimination (RFE) under nested cross-validation, applied separately for each L1. Features included word-level difficulty (e.g., surface form, frequency, morphology, concreteness, CEFR levels), cross-lingual features (e.g., cognates and false friends), and test item difficulty (e.g. spelling difficulty and candidate competition features). Feature extraction relied on traditional corpus-based and lexical resources as well as masked-LM pseudo-likelihood guessing to probe model behaviour as a proxy test-taker. The pool also included a small set of derived second-order features, such as pairwise interactions and model-disagreement metrics. Other systems falling within the top statistical grouping included uogal (German, Mandarin), who used ensembles of fine-tuned transformer models augmented with EN translations of L1 content, and Sakura (Mandarin), who combined a single fine-tuned encoder prediction with explainable features in an XGBoost regressor.

The winning team for the open track was Sakura (Nohejl et al., 2026), whose submitted runs took the top three positions for German and Mandarin, and the top two positions for Spanish. The best runs achieved an RMSE of 0.742, 0.723 and 0.630 for Spanish, German and Mandarin, respectively, with corresponding Pearson correlations of 0.919, 0.916 and 0.928. The approach was centred on a novel soft-target cross-entropy loss for fine-tuning LLMs to predict continuous difficulty values. Three LLMs from different families were fine-tuned using this method and combined via linear regression stacking with out-of-fold predictions. This ensemble was then enhanced with word-level difficulty features (e.g., production and reception frequency, CEFR level, word length), cross-lingual features (e.g., orthographic similarity and morphemic over-

lap), and test item difficulty features (e.g. spelling difficulty and lexical ambiguity, extracted using LLMs as raters). Prompt-based features were calibrated using temperature-scaled, probability-weighted inference (G-Scale) (Nohejl et al., 2025). This configuration was the best-performing for Spanish and Mandarin. The team also experimented with a further set of test item difficulty features combining LLM-prompted difficulty ratings and proxy test-taker estimates, in which LLMs were prompted to solve the items. This configuration took the top position for German. Also within the top statistical grouping, Glite applied the same approach as their closed-track submission, expanding the feature set to include both fine-tuned and prompted LLM difficulty ratings, proxy test-taker estimates, and LLM-based item judgements.

6 Analysis and Discussion

To investigate how performance varied across tracks and L1s, we calculated Mean Absolute Error (MAE), Mean Signed Error (MSE), average prediction Standard Deviation (Avg. SD) and average Pearson correlation (Table 3). Results show that MAE is consistently lower for systems in the open track (by 0.13 on average). This aligns with findings from teams that evaluated joint training across L1s (EduNLP, Jinnie's Lab, Sakura, TeamXBC), all of which reported improvements over per-L1 training, suggesting the open track gains reflect, at least in part, the benefit of shared signal across L1 subsets. This is most pronounced for Mandarin, which shows the lowest average error despite the orthographical distance from English. However, as noted by Data Asgardians, this is likely due to the reduced range of GLMM values for Mandarin (approximately -4 to 4), which is narrower than for Spanish (approximately -6 to 5) and German (approximately -5 to 4). In fact, looking at average Pearson correlation (a scale-independent measure) reveals that open-track performance for Mandarin is actually lower than for Spanish and German, confirming that its lower MAE is an artefact of the narrower scale.

MSE was positive in all cases, indicating that systems tend to underestimate difficulty on average. Open-track systems show less underestimation across all three L1s, though this is less pronounced for Mandarin — again, likely reflecting its narrower scale. Average SD improves for Spanish and German but increases slightly for Mandarin in the open track, suggesting that while overall accuracy improves, open-track systems do not achieve greater consensus. Together, these results confirm that the benefits of cross-lingual training are not uniform across L1s.

To examine the L1 and track level differences outlined in Table 3, we investigated individual predictions for each item in the test set. Figure 1 shows the spread of predictions for each test item (in yellow), sorted from easiest (highest GLMM score) to hardest (lowest GLMM score). Ground-truth values are shown in purple, contrasted against regression lines for each winning submission (pink) and mean system performance (dotted red). In addition, we include trend curves for the lowest and highest predictions (dotted orange) in order to better evaluate prediction drift.

System predictions show strong monotonic alignment with true difficulty, meaning that models can successfully capture the ordinal relationship of the target values. This is confirmed by the high Pearson correlation coefficients reported in Appendix B. Predictions overlap substantially with target values across the dataset, except at the extremes, where the items are increasingly difficult to predict. This is regardless of the difficulty range (notably larger for Spanish) and likely a consequence of the small number of samples in the dataset with extreme difficulty values. Several submission papers support this, reporting weaker accuracy at both ends of the GLMM scale. EduNLP found that 'very hard' items (GLMM < -3) contribute 25.4% of total squared error despite comprising only 5.1% of items. UOL@IDEM report that the top 20% of 'very easy' items are over-predicted by +1.03 to +1.28 GLMM points by their closed track models. TOEBM showed that prediction bias may be model dependent: XLM-RoBERTa underestimated difficult items, whereas mDeBERTa-v3 showed systematic overestimation and mmBERT was more consistent across difficulty levels. Ensembling with a Ridge regressor exploited these complementary biases to produce more consistent predictions.

Looking at the boundary lines, we observe that the spread of predictions in the closed track is broadly homogenous across all L1s, although Spanish and Mandarin do exhibit an opening towards more difficult items (lower GLMM values), indicating less consensus among systems. Given the small number of extremely difficult items in the datasets, it is unsurprising that systems struggle to predict them. In the open track, prediction spread

Spanish

Rank	Team	RMSE	Pearson
1	Glite	0.903	.877
2	uogal	0.975	.858
3	AIDA	0.976	.857
4	Sakura	0.983	.854
5	NLP-Explorers	1.040	.839
6	SAAKTH	1.045	.836
7	TOEBM	1.063	.826
8	Jinnie's Lab	1.084	.818
9	RETUYT-InCo	1.094	.843
10	SATLab	1.108	.813
11	UZHCL	1.121	.816
12	UOL@IDEM	1.132	.813
13	TeamXBC	1.146	.823
14	Token Titans	1.170	.812
15	EduNLP	1.176	.788
16	Data Asgardians	1.182	.820
17	Failure	1.219	.763
18	Boosted Cats	1.236	.755
19	UGA Threshold	1.236	.758
20	Mixed Signals	1.244	.760
21	**BASELINE**	**1.257**	**.765**
22	Unibuc	1.421	.664

German

Rank	Team	RMSE	Pearson
1	Glite	0.885	.871
2	uogal	0.903	.869
3	Sakura	0.963	.844
4	NLP-Explorers	0.992	.845
5	SAAKTH	0.994	.844
6	TOEBM	0.997	.832
7	Jinnie's Lab	1.011	.827
8	SATLab	1.036	.830
9	UOL@IDEM	1.037	.834
10	UZHCL	1.071	.826
11	TeamXBC	1.114	.837
12	Data Asgardians	1.117	.830
13	Boosted Cats	1.118	.783
14	Failure	1.118	.788
15	EduNLP	1.124	.793
16	UGA Threshold	1.166	.761
17	Mixed Signals	1.173	.784
18	**BASELINE**	**1.258**	**.773**
19	RETUYT-InCo	1.260	.713
20	Unibuc	1.305	.692
21	Token Titans	1.564	.760

Mandarin

Rank	Team	RMSE	Pearson
1	Glite	0.776	.889
2	Sakura	0.816	.874
3	uogal	0.820	.879
4	TOEBM	0.880	.853
5	NLP-Explorers	0.882	.861
6	UOL@IDEM	0.891	.860
7	SAAKTH	0.900	.860
8	Jinnie's Lab	0.905	.843
9	SATLab	0.921	.846
10	TeamXBC	0.968	.851
11	UZHCL	0.991	.839
12	Data Asgardians	1.006	.854
13	Boosted Cats	1.030	.790
14	UGA Threshold	1.031	.791
15	Mixed Signals	1.055	.790
16	EduNLP	1.058	.800
17	Failure	1.090	.763
18	RETUYT-InCo	1.106	.754
19	**BASELINE**	**1.140**	**.753**
20	Unibuc	1.176	.717
21	Token Titans	1.382	.718
22	BZPT	1.588	.590

(a) Closed Track.

Spanish

Rank	Team	RMSE	Pearson
1	Sakura	0.742	.919
2	Glite	0.754	.916
3	TeamXBC	0.876	.885
4	uogal	0.975	.858
5	SurreyCTS	1.034	.839
6	Jinnie's Lab	1.053	.829
7	SATLab	1.087	.819
8	EduNLP	1.143	.802
9	**BASELINE**	**1.198**	**.783**

German

Rank	Team	RMSE	Pearson
1	Sakura	0.723	.916
2	Glite	0.764	.905
3	TeamXBC	0.826	.888
4	uogal	0.903	.869
5	SurreyCTS	0.945	.854
6	Jinnie's Lab	0.990	.834
7	SATLab	1.010	.828
8	EduNLP	1.071	.816
9	**BASELINE**	**1.166**	**.786**

Mandarin

Rank	Team	RMSE	Pearson
1	Sakura	0.630	.928
2	Glite	0.660	.920
3	TeamXBC	0.722	.904
4	uogal	0.820	.879
5	SurreyCTS	0.861	.863
6	Jinnie's Lab	0.885	.851
7	SATLab	0.926	.847
8	EduNLP	0.992	.824
9	**BASELINE**	**1.034**	**.804**
10	BZPT	1.123	.744

(b) Open Track.

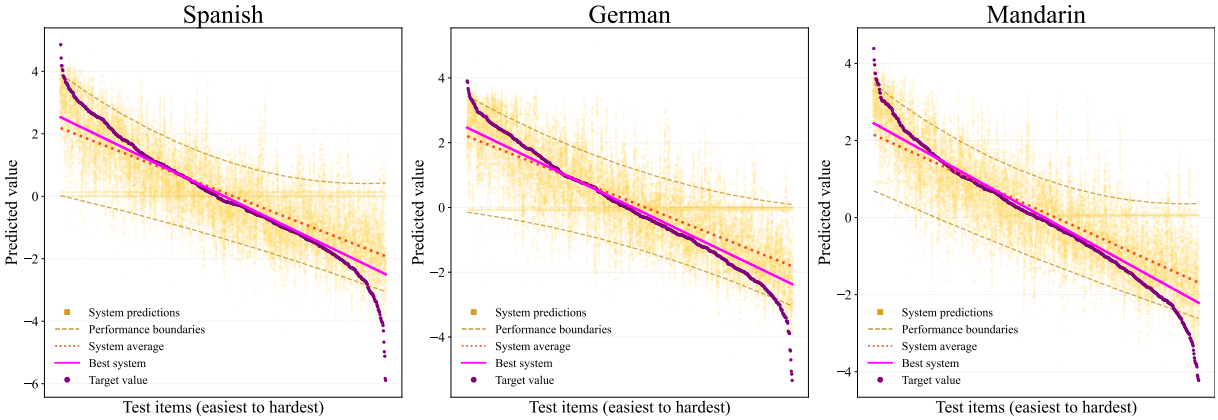
Table 2: Best results per team, track and L1.

Track	L1	MAE	MSE	Avg. SD	Pearson
Closed	ES	.897	.108	.627	.773
Closed	DE	.869	.214	.607	.795
Closed	CN	.770	.214	.523	.815
Open	ES	.744	.061	.439	.850
Open	DE	.722	.098	.448	.849
Open	CN	.685	.155	.529	.826

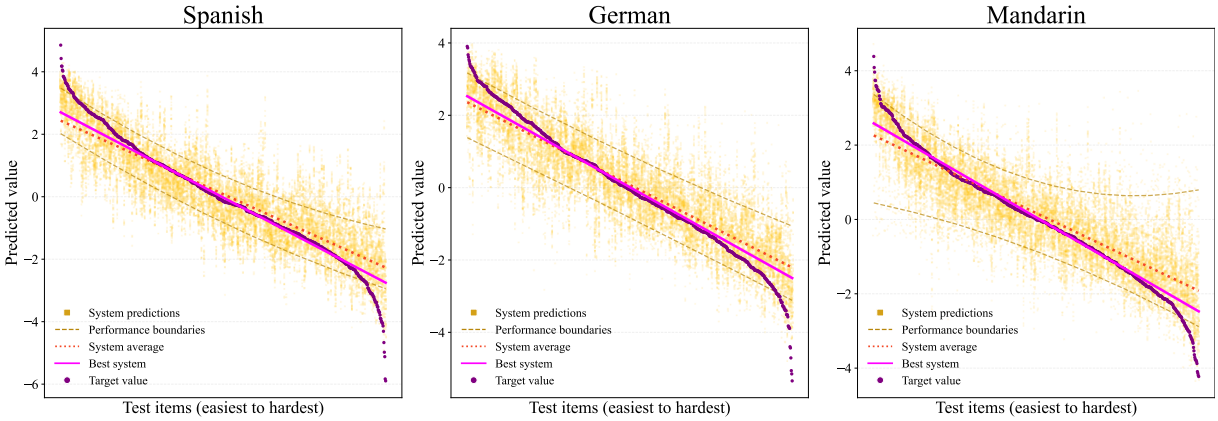
Table 3: Aggregated metrics per track and L1.

shrinks for all L1s, especially for Spanish and German. For Mandarin, however, dispersion increases at both ends of the difficulty scale in the open track, suggesting that additional signals may introduce noise rather than useful information for this L1 at the extremes. This finding may in part be due to the greater orthographical distance between Mandarin and English limiting cross-lingual generalisability compared to Spanish and German, reported in prior work (Skidmore et al., 2025) and corroborated in submissions for the shared task (UGA Threshold, Boosted Cats, SurreyCTS).

The wide vertical dispersion observed for some items indicates cases where the systems largely disagree. These are words that the systems find particularly confusing. Likewise, items with the highest error (where the yellow predictions are furthest away from the purple values) indicate the words that were the most difficult to predict. Table 4 includes some examples. The words that are easier to predict are either cognates, concrete nouns, or specific terminology, whereas the most difficult ones tend to be polysemous words (whose most frequent sense is not the target sense), abstract words, collocates or words with specific parts-of-speech (gerunds, adverbs, verbs). Item-level error analyses carried out by the teams corroborate this, with models consistently struggling most with words with uncommon senses or grammatical roles. In particular, EduNLP found that nominalised '-ing' forms showed a 58% higher RMSE than other words, such as "dining" "baking", "speaking" seen in Table 4, while Boosted Cats found that items where the tested POS differed from the dominant POS,



(a) Closed track results.



(b) Open track results.

Figure 1: System prediction distributions sorted by target difficulty (higher GLMM scores → easier items).

illustrated by "spring", "amount" and "received" in Table 4, increased prediction variance by 24–32%.

Considering the shared task approaches as a whole, there are several conclusions that can be drawn. Firstly, improvements over the XLM-RoBERTa-base baseline were driven primarily by model choice. In particular, switching to XLM-RoBERTa-large delivered improvements across multiple teams (e.g. TeamXBC). Architectural modifications such as multi-layer pooling and layer aggregation offered smaller improvements in some systems (Jinnie's Lab, SurreyCTS), though more complex modifications such as layerwise learning rate decay were largely ineffective or harmful (TeamXBC). Secondly, feature-based models were competitive with the baseline, particularly for Spanish and German where cross-linguistic similarity signals are informative (Boosted Cats, SAAKTH). However, ablation studies for hybrid models revealed that incorporating hand-crafted features provided only small improvements to already strong ensemble baselines (Glite, SAAKTH). Nevertheless, feature-based models offer a clear advantage in interpretability, providing insights into what drives difficulty for different learner groups, which is essential for item development in practice (Dunn et al., 2026). Finally, ensembling almost always helped, with model diversity proving more important than raw individual model strength (TOEBM, uogal). Interestingly, however, complex stacking strategies were found to overfit to the small development set provided (EduNLP, TeamXBC, uogal).

The two winning teams, Glite and Sakura, shared a common architectural approach to the task: combining out-of-fold predictions from multiple fine-tuned models with a broad set of hand-crafted features. Both were among the few to use features targeting item-level as well as word-level properties, in particular proxy test-taker signals from masked LMs and LLM-based judgements of item difficulty. They were among only three teams to experiment with fine-tuned and prompted LLMs as a core component of their systems, which proved highly effective but remains relatively underexplored across the shared task as a whole. Both systems also used cross-validation strategies for model training. To-

Category	Spanish	German	Mandarin
Least Confusing (lowest std dev)	breathing (n.), prepared (adj.), product (n.), demand (v.), defeat (v.)	notice (n.), necessary (adj.), pile (n.), reward (v.), admire (v.)	necessary (adj.), morality (n.), blame (v.), criminal (adj.), surrounding (adj.)
Most Confusing (highest std dev)	shape (v.), marked (adj.), dozen (num.), received (adj.), knight (v.)	crowned (v.), tweet (v.), received (adj.), walker (n.), shape (v.)	standing (n.), schoolmate (n.), remarkably (adv.), clear (adv.), house (v.)
Easiest to predict (lowest error)	perfectionist (n.), bow (n.), frequently (adv.), blame (v.), crowd (v.)	necessary (adj.), increasingly (adv.), dog (n.), genetically (adv.), album (n.)	surgery (n.), left (adj.), exercise (v.), silence (n.), luxury (n.)
Hardest to predict (highest error)	dining (n.), baking (n.), received (adj.), clear (adv.), unchallenged (adj.)	streak (n.), anger (v.), umbrella (n.), fund (n.), hydrated (adj.)	speaking (n.), stake (n.), swear (v.), clear (adv.), might (v.)

Table 4: Sample vocabulary items for different prediction categories.

gether, these factors suggest that the performance gap between the winning systems and the rest was driven by the combination of multiple model predictions, broad and well-chosen feature coverage, and evidence-based model development.

Across submissions, a picture emerged of vocabulary difficulty as shaped by both an underlying component tied to features of the English word and a L1-specific signal. This is supported by the fact that language-agnostic pipelines with no L1-specific features outperformed the baseline models (Failure). Improvements from joint L1 training (EduNLP, Jinnie's Lab, Sakura, TeamXBC) further confirm that much of this can be shared across learner groups, which supports findings of previous research (Skidmore et al., 2025). At the same time, per-L1 hand-crafted features also proved valuable, and analyses of feature importance revealed meaningful differences in what drives difficulty for different learner groups (Boosted Cats, Sakura). A similar pattern was evident in ensemble models, where different encoder combinations were preferred for different L1s (uogal). Cross-lingual transfer experiments reinforced this picture: Spanish and German models transferred to each other reasonably well, but Mandarin transferred poorly, reflecting fundamental differences in how difficulty is structured across typologically distant pairs (Boosted Cats, SAAKTH). Together, these findings affirm the value of L1-aware modelling with clear implications for scalable, personalised assessment and adaptive vocabulary learning.

7 Future Work

Looking ahead, several directions emerge from the findings of the shared task. The 'compression effect' of the target labels provide motivation to experiment with alternative labelling strategies, such as ordinal or banded difficulty rather than continuous GLMM scores. The L1 asymmetry between European languages and Mandarin makes a strong case for extending vocabulary difficulty prediction to more typologically diverse L1s, such as Arabic or Japanese, to test whether the findings here generalise beyond the specific language pairs studied. Finally, making use of learner response data to validate the insights produced by interpretable models would provide empirical grounding for conclusions that currently rest on model behaviour alone. Features identified as important by multiple systems such as spelling difficulty, polysemy of the target word, and cross-lingual transfer signals could be directly tested against learner responses, representing a valuable bridge between the NLP and language testing research communities.

8 Conclusions

This paper presented an overview of the BEA 2026 Shared Task on Vocabulary Difficulty Prediction for English Learners, predicting difficulty scores for English lexical items across three L1s (Spanish, German, and Mandarin). The task comprised closed and open tracks, restricting and permitting additional data and models respectively. A total of 23 teams participated using a variety of approaches, including classical machine learning, transformer models, LLMs and ensembles. Most systems outperformed our transformer-based baseline, with winners in each track doing so by a large margin.

System analysis showed that ensembles combining linguistic features with fine-tuned model predictions achieved the best results. Systems tended to underestimate difficulty and struggled the most with polysemous words. Open-track systems consistently outperformed their closed-track counterparts, suggesting that joint training across L1s improves predictive accuracy. However, the degree of impact of these approaches varies across L1s and difficulty levels and warrants further investigation.

We believe the insights gained from the shared task will significantly contribute to the design of more robust vocabulary prediction models, ultimately supporting better test item development and more adaptive language learning applications.

9 Limitations

Most of the limitations of this work stem from the datasets used to train and evaluate the systems. The data was restricted to three L1s (German, Spanish, Mandarin), so conclusions about system performance must be interpreted with caution, as they may not generalise to learners with different linguistic backgrounds. Furthermore, the test items across these L1s were not standardised, potentially introducing variation in difficulty estimates beyond learner background. Some participants found that the small development set (677 items per L1) limited reliable model selection and ensemble calibration, which may have impacted the overall performance of some systems. In addition, the use of the probabilistic values derived from the GLMM framework as observed difficulty values comes with its own caveats, as discussed by Schmitt et al. (2024).

Since GLMM scores are estimations rather than direct observations, RMSE values should be interpreted with care. To provide a more robust assessment of model performance, we also report Pearson correlation, which captures the linear relationship between predictions and true values and is less sensitive to numerical fluctuations.

The GLMM difficulty values in our datasets are also not directly comparable across L1s, since they were derived from different population samples. This is something that must be taken into account when training systems using a combination of data from all L1s, such as our open track baseline.

For limitations pertaining to individual submissions, please refer to the corresponding description papers.

10 Ethics statement

The dataset for this shared task was collected by the British Council through a large-scale crowdsourcing initiative (Schmitt et al., 2021). Participation was voluntary and participants agreed for their response data to be used for research purposes. All data was anonymised prior to analysis. Vocabulary difficulty scores were derived from aggregated responses using psychometric modelling, ensuring that individual identities cannot be inferred.

A primary objective of this shared task is the promotion of responsible AI practices in language education, with an emphasis on transparency, reproducibility, and fairness. For this reason, all the data, baseline models, evaluation scripts and documentation have been made publicly available for research and educational purposes. The insights from the submitted models and their respective description papers also aim to foster greater transparency and explainability in the domain of AI for education.

Acknowledgements

The authors would like to thank the original research team behind the Knowledge-based Vocabulary Lists, whose datasets have been fundamental to this shared task. Special thanks go to Karen J. Dunn for her continued guidance and support. We also wish to thank all participants of the BEA 2026 Shared Task for their engagement and contributions.

References

Abid Al Hossain and Kamruzzaman Khan Alve. 2026. Failure at BEA 2026 Shared Task 1: One pipeline, three L1s: A unified language-agnostic system for vocabulary difficulty prediction. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, San Diego, California. Association for Computational Linguistics.

Desislava Aleksandrova and Vincent Pouliot. 2023. CEFR-based contextual lexical complexity classifier in English and French. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 518–527, Toronto, Canada. Association for Computational Linguistics.

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, New Orleans, Louisiana. 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.

Stefano Bannò, Kate M. Knill, and Mark J. F. Gales. 2025. Exploiting the English vocabulary profile for L2 word-level vocabulary assessment with LLMs. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 632–646, Vienna, Austria. Association for Computational Linguistics.

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 Comput. Surv.*, 55(9).

Yves Bestgen. 2026. SATLab at BEA 2026 Shared Task 1: Predicting the difficulty of English words for three L1 learners using primarily psycholinguistic features. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, San Diego, California. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Xiaobin Chen. 2026. TeamXBC at BEA 2026 Shared Task 1: How AI (and I) won the shared task: Vibe and agentic coding solutions for practical machine learning problems. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, San Diego, California. Association for Computational Linguistics.

Seok Hyeon Cho, JunHyeok Choi, Sangeun Ji, and Sung Won Han. 2026. AIDA at BEA 2026 Shared Task 1: A two-stage framework for L1-aware vocabulary difficulty prediction with representation diversity and residual calibration. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, San Diego, California. Association for Computational Linguistics.

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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.

Emma Dalbo. 2026. UGA Threshold at BEA 2026 Shared Task 1: Predicting vocabulary acquisition difficulty with hand-crafted SLA-based features. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, San Diego, California. Association for Computational Linguistics.

Paul De Boeck. 2008. Random item IRT models. *Psychometrika*, 73(4):533–559.

Jasper Degraeuwe. 2025. You shall know a word's difficulty by the family it keeps: Word family features in personalised word difficulty classifiers for L2 Spanish. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 312–325, Vienna, Austria. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and 1 others. 2024. The LlaMA 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Karen J. Dunn. 2024. Random-item Rasch models and explanatory extensions: A worked example using L2 vocabulary test item responses. *Research Methods in Applied Linguistics*, 3(3):100143.

Karen J. Dunn, Lucy Skidmore, and Thomas Rogers. 2026. When measurement meets machine learning: interpretability and scalability in modelling item difficulty for language assessment. *Frontiers in Education*, 11.

Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Thomas François and Núria Gala. 2024. Graded resources for learning and teaching foreign languages: An overview. *ITL – International Journal of Applied Linguistics*, 175(1):8–24.

Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. 2013. Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In *Proceedings of the 31st Second Language Research Forum (SLRF)*, pages 240–254.

Sian Gooding and Manuel Tragut. 2022. One size does not fit all: The case for personalised word complexity models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 353–365, Seattle, United States. Association for Computational Linguistics.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, and 1 others. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Nouran Khallaf and Serge Sharoff. 2026. UOL@IDEM at BEA 2026 Shared Task 1: Neural fusion and feature-rich modeling for L1-aware vocabulary difficulty prediction. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, San Diego, California. Association for Computational Linguistics.

Tayyab Latif, Asifa Bibi, Sabur Butt, and Sidorov Grigori. 2026. NLP-Explorers at BEA 2026 Shared Task 1: DeBERTa–CatBoost weighted ensemble approach for L1-specific vocabulary difficulty prediction. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, San Diego, California. Association for Computational Linguistics.

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

John Lee and Chak Yan Yeung. 2018. Personalizing lexical simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 224–232, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Wicaksono Leksono, Joanito Agili Lopo, Tsamarah R. N. Nugraha, Ahmad Cahyono Adi, and Muhamad Oriza Nurfajri. 2026. TOEBM at BEA 2026 Shared Task 1: Improving lexical difficulty prediction with context-aligned contrastive learning and ridge ensembling. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, San Diego, California. Association for Computational Linguistics.

Zhe Li, Pauline Aguinalde, and Jinnie Shin. 2026. Jinnie's Lab at BEA 2026 Shared Task 1: Precalibration of vocabulary item difficulty with multilingual transformers and multi-task learning. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, San Diego, California. Association for Computational Linguistics.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, and 101 others. 2026. Ministral 3. *Preprint*, arXiv:2601.08584.

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.

Jonas Mayer Martins, Zhuojing Huang, Aaricia Herygers, and Lisa Beinborn. 2026. BoostedCats at BEA 2026 Shared Task 1: What makes a word hard to learn? Modeling L1 influence on English vocabulary difficulty. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, San Diego, California. Association for Computational Linguistics.

Karthik Mattu, Adit Dhall, Arshad Naguru, Shubh Sehgal, Thejas Nagesh Gowda, and Hakyung Sung. 2026. SAAKTH at BEA 2026 Shared Task 1: L1-aware English vocabulary difficulty prediction with hybrid transformer and psycholinguistic features. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, San Diego, California. Association for Computational Linguistics.

Adam Nohejl, Akio Hayakawa, Yusuke Ide, and Taro Watanabe. 2025. A Japanese dataset and efficient multilingual LLM-based methods for Lexical Simplification and Lexical Complexity Prediction. *Journal of Natural Language Processing*, 32(4):1129–1188.

Adam Nohejl, Xuanxin Wu, Yusuke Ide, Maria Angelica Riera Machin, and Yi-Ning Chang. 2026. Sakura at BEA 2026 Shared Task 1: What makes vocabulary difficult? In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, San Diego, California. Association for Computational Linguistics.

Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2025. Deep learning approaches to lexical simplification: A survey. *Journal of Intelligent Information Systems*, 63:111–134.

Kai North and Marcos Zampieri. 2023. Features of lexical complexity: insights from L1 and L2 speakers. *Frontiers in Artificial Intelligence*, 6.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical complexity prediction: An overview. *ACM Comput. Surv.*, 55(9).

Gustavo Paetzold and Lucia Specia. 2016. 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.

Anubhab Parashar and Sandeep Mathias. 2026. Token Titans at BEA 2026 Shared Task 1: Multilingual lexical complexity prediction via fine-tuned XLM-RoBERTa with ensemble decoding. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, San Diego, California. Association for Computational Linguistics.

Sydney Peters, Nan Zhang, Hong Jiao, Ming Li, and Tianyi Zhou. 2025. Review of text-based approaches to item difficulty modeling in large-scale assessments. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Coordinated Session Papers*, pages 37–47, Wyndham Grand Pittsburgh, Downtown, Pittsburgh, Pennsylvania, United States. National Council on Measurement in Education (NCME).

Vassili Philippov, Dmitrii Andreev, Pavel Katunin, and Anton Nikolaev. 2026. Glite at BEA 2026 Shared Task 1: Holistic difficulty models dominate, feature engineering closes the gap in L1-aware vocabulary

difficulty prediction. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, San Diego, California. Association for Computational Linguistics.

Adrian Pineda, Sabur Butt, and Héctor Gibrán Ceballos Cancino. 2026. Data Asgardians at BEA 2026 Shared Task 1: A hybrid transformer–feature ensemble for L1-aware English vocabulary difficulty prediction. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, San Diego, California. Association for Computational Linguistics.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. Catboost: unbiased boosting with categorical features. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 6639–6649, Red Hook, NY, USA. Curran Associates Inc.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Santiago Robaina, Aiala Rosá, and Luis Chiruzzo. 2026. RETUYT-INCO at BEA 2026 Shared Task 1: Feature-enriched mDeBERTa for word difficulty prediction. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, San Diego, California. Association for Computational Linguistics.

Norbert Schmitt, Karen Dunn, Barry O'Sullivan, Laurence Anthony, and Benjamin Kremmel. 2021. Introducing Knowledge-based Vocabulary Lists (KVL). *Tesol Journal*, 12(4).

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

Norbert Schmitt, Paul Nation, and Benjamin Kremmel. 2020. Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching*, 53(1):109–120.

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, 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.

Avinash Kumar Sharma. 2026. EduNLP at BEA 2026 Shared Task 1: Multi-model ensemble with feature-augmented transformers for vocabulary difficulty prediction. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, San Diego, California. 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.

Bernardo Stearns, John P. McCrae, Thomas Gaillat, and Jefkine Kafunah. 2026. uogal at BEA 2026 Shared Task 1: Ensemble of multilingual encoders with NMT augmentation for L1-aware vocabulary difficulty prediction. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, San Diego, California. Association for Computational Linguistics.

Team GLM. 2024. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools. *arXiv preprint arXiv:2406.12793*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

Georgina Jennifer Willoughby, Jordan Painter, Diptesh Kanojia, Emily Frances Wells, and Constantin Orasan. 2026. SurreyCTS at BEA 2026 Shared Task 1: Semantic funnelling and entropy-based multilingual lexical difficulty prediction. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, San Diego, California. Association for Computational Linguistics.

Victoria Yaneva, Kai North, Peter Baldwin, Le An Ha, Saed Rezayi, Yiyun Zhou, Sagnik Ray Choudhury, Polina Harik, 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, Mexico City, Mexico. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, and 1 others. 2024. Qwen2.5: A party of foundation models. *arXiv preprint arXiv:2412.15115*.

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.

A Supplementary tables

L1	Train	Development	Test
Spanish	6,091	677	748
German	6,091	677	748
Mandarin	6,091	677	748

Table A5: An overview of the dataset splits used in the shared task, totalling 7,516 vocabulary items per L1 background.

Model name	L1	Learning rate	Weight decay	Warmup ratio	Batch size	Dropout rate
`baseline_closed_es`	ES	3e-5	0.1	0.1	32	0.1
`baseline_closed_de`	DE	3e-5	0.0	0.1	32	0.1
`baseline_closed_cn`	CN	3e-5	0.1	0.1	32	0.1
`baseline_open_xx`	XX	3e-5	0.1	0.1	32	0.1

Table A6: Optuna hyperparameters for the baseline models, following Skidmore et al. (2025).

	Open			Closed		
Team	**ES**	**DE**	**CN**	**ES**	**DE**	**CN**
AIDA	–	–	–	3	–	–
BZPT	–	–	2	–	–	1
Boosted Cats (HuDS lab)	–	–	–	3	3	3
Data Asgardians	–	–	–	3	3	3
EduNLP	3	3	3	3	3	3
Failure	–	–	–	1	1	1
Glite	3	3	3	3	3	3
Jinnie's Lab	3	3	3	3	3	3
Mixed Signals	–	–	–	3	3	3
NLP-Explorers	–	–	–	3	3	3
RETUYT-InCo	–	–	–	2	1	1
SAAKTH	–	–	–	3	3	3
SATLab	3	3	3	3	3	3
Sakura	3	3	3	3	3	3
SurreyCTS	3	3	3	–	–	–
TOEBM	–	–	–	3	3	3
TeamXBC	3	3	3	3	3	3
Token Titans	–	–	–	2	3	2
UGA Threshold	–	–	–	3	3	3
UOL@IDEM	–	–	–	3	3	3
UZHCL	–	–	–	1	1	1
Unibuc	–	–	–	2	2	2
uogal	2	2	2	2	2	2
Total	**23**	**23**	**25**	**55**	**52**	**52**

Table A7: Number of runs submitted per team, track, and L1.

Team Name	Track	L1	Approach(es)	Model(s)
AIDA (Cho et al., 2026)	Closed	Spanish	Transformers; features; ensemble	mDeBERTa-v3-base; RemBERT; XLM-RoBERTa-large; LightGBM
Boosted Cats (HuDS lab) (Martins et al., 2026)	Closed	All	Features	CatBoost
BZPT	Both	Mandarin	—	—
Data Asgardians (Pineda et al., 2026)	Closed	All	Transformers; features	XLM-RoBERTa-large
EduNLP (Sharma, 2026)	Both	All	Transformers; features; ensemble	XLM-RoBERTa-base; LightGBM
Failure (Al Hossain and Alve, 2026)	Closed	All	Transformers; features; ensemble	XLM-RoBERTa-base
Glite (Philippov et al., 2026)	Both	All	Transformers; features; ensemble; LLMs	CatBoost; XLM-RoBERTa-base; XLM-RoBERTa-large; XLM-RoBERTa-XL; XLM-RoBERTa-XXL; RemBERT; ELECTRA-base; mBERT; LLaMA-3.1-8B; Qwen2.5-7B
Jinnie's Lab (Li et al., 2026)	Both	All	Transformers	mmBERT-base
Mixed Signals	Closed	All	—	—
NLP-Explorers (Latif et al., 2026)	Closed	All	Transformers; features; ensemble	DeBERTa-v3-large; XLM-RoBERTa-large; CatBoost
RETUYT-InCo (Robaina et al., 2026)	Closed	All	Transformers; features; ensemble	XGBoost; mDeBERTa-v3-base
SAAKTH (Mattu et al., 2026)	Closed	All	Transformers; features; ensemble	XLM-RoBERTa-large; XGBoost
Sakura (Nohejl et al., 2026)	Both	All	Transformers; features; ensemble; LLMs	mmBERT-base; GLM-4-32B; Qwen2.5-32B; Ministral-3-14B; XGBoost; linear regression
SATLab (Bestgen, 2026)	Both	All	Transformers; features; ensemble	LightGBM; XLM-RoBERTa-base
SurreyCTS (Willoughby et al., 2026)	Open	All	Transformers; features; ensemble	RemBERT; COMET
TeamXBC (Chen, 2026)	Both	All	Transformers; ensemble; LLMs	XLM-RoBERTa-large; RemBERT; Qwen2.5-7B
TOEBM (Leksono et al., 2026)	Closed	All	Transformers; ensemble	XLM-RoBERTa-base; mDeBERTa-v3-base; mmBERT-base
Token Titans (Parashar and Mathias, 2026)	Closed	All	Transformers; ensemble	XLM-RoBERTa-large
UGA Threshold (Dalbo, 2026)	Closed	All	Features; ensemble	scikit-learn models: Ridge; SGD (ElasticNet); Random Forest; Gradient Boosting; Extra Trees; MLP
Unibuc	Closed	All	—	—
uogal (Stearns et al., 2026)	Both	All	Transformers; features; ensemble	Multilingual-E5-large; XLM-RoBERTa-large; RemBERT; InfoXLM-Large
UOL@IDEM (Khallaf and Sharoff, 2026)	Closed	All	Transformers; features	Multilingual-E5-large; BGE-M3
UZHCL	Closed	All	Transformers; ensemble	XLM-RoBERTa-large

Table A8: Participating teams and chosen approaches. The models reported include those used to make predictions (including ensemble members), but exclude models used for feature extraction. Any additional experiments conducted by the teams beyond the submitted runs are not included.

B Official leaderboards

Group	Rank	Team Name	Prediction File	RMSE	Pearson
1	1	Glite	predictions_run_3.csv	0.903	0.877
	2	Glite	predictions_run_2.csv	0.910	0.876
	3	Glite	predictions_run_1.csv	0.920	0.872
2	4	uogal	8enc_opt_weights_..._dev_preds.csv	0.975	0.858
	5	AIDA	predictions_run_1.csv	0.976	0.857
	6	uogal	8enc_dmeta_elasticnet_..._preds.csv	0.977	0.859
	7	Sakura	predictions_closed_max.csv	0.983	0.854
3	8	AIDA	predictions_run_2.csv	1.039	0.840
	9	NLP-Explorers	predictions_run_2.csv	1.040	0.839
	10	NLP-Explorers	predictions_run_1.csv	1.040	0.839
	11	NLP-Explorers	predictions_run_3.csv	1.041	0.838
	12	SAAKTH	predictions_run_3.csv	1.045	0.836
	13	SAAKTH	predictions_run_2.csv	1.053	0.835
	14	TOEBM	predictions_run_3.csv	1.063	0.826
	15	AIDA	predictions_run_3.csv	1.066	0.835
	16	Jinnie's Lab	mbert-base_mtl_mlp_cls_es_preds.csv	1.084	0.818
	17	SAAKTH	predictions_run_1.csv	1.087	0.819
	18	TOEBM	predictions_run_2.csv	1.089	0.816
	19	RETUYT-InCo	predictions_ensemble.csv	1.094	0.843
	20	SATLab	Dev.csv	1.108	0.813
	21	SATLab	Boot.csv	1.110	0.811
	22	Jinnie's Lab	mbert-base_mtl_mlp_mean_es_preds.csv	1.114	0.807
4	23	UZHCL	grand_ensemble_ft_preds.csv	1.121	0.816
	24	UOL@IDEM	run_1.csv	1.132	0.813
	25	UOL@IDEM	run_2.csv	1.134	0.808
	26	UOL@IDEM	run_3.csv	1.140	0.813
	27	Jinnie's Lab	mbert-base_mtl_mlp_mha_1_es_preds.csv	1.142	0.796
	28	TeamXBC	TeamXBC_predictions_run_2.csv	1.146	0.823
	29	Sakura	predictions_explainable.csv	1.156	0.789
	30	Token Titans	predictions_bert_1.csv	1.170	0.812
	31	EduNLP	predictions_closed_ensemble.csv	1.176	0.788
	32	EduNLP	predictions_closed_stretched.csv	1.178	0.788
	33	Data Asgardians	predictions_hp_solo.csv	1.182	0.820
5	34	Data Asgardians	predictions_hp_hybrid.csv	1.186	0.825
	35	Data Asgardians	predictions_closed_solo.csv	1.190	0.814
	36	TeamXBC	TeamXBC_predictions_run_1.csv	1.192	0.816
	37	SATLab	Nonly.csv	1.212	0.778
	38	Failure	predictions_run_1.csv	1.219	0.763
	39	EduNLP	predictions_closed_transformer.csv	1.220	0.787
	40	Boosted Cats	submission_es_run1.csv	1.236	0.755
	41	Boosted Cats	submission_es_run3.csv	1.236	0.755
	42	UGA Threshold	predictions_run_1.csv	1.236	0.758
	43	Boosted Cats	submission_es_run2.csv	1.237	0.754
	44	Mixed Signals	num_es_base_ipa_pos_preds.csv	1.244	0.760
	45	**BASELINE**	**baseline_closed_track.csv**	**1.257**	**0.765**
	46	UGA Threshold	predictions_run_3.csv	1.261	0.746
	47	Mixed Signals	numeric_es_base_preds.csv	1.264	0.752
	48	UGA Threshold	predictions_run_2.csv	1.268	0.740
	49	Mixed Signals	num_es_base_ipa_pos_mean_preds.csv	1.278	0.744
	50	Sakura	predictions_traditional.csv	1.305	0.721
6	51	RETUYT-InCo	predictions_xgboost.csv	1.323	0.713
	52	TOEBM	predictions_aya_llamma_ens.csv	1.324	0.719
	53	Unibuc	predictions_run_2.csv	1.421	0.664
	54	Unibuc	predictions_run_1.csv	1.425	0.659
7	55	TeamXBC	TeamXBC_predictions_run_3.csv	1.882	0.093
	56	Token Titans	predictions_debert_2.csv	1.885	0.043

Table A9: Full results for Closed Track: Spanish.

Group	Rank	Team Name	Prediction File	RMSE	Pearson
1	1	Glite	predictions_run_3.csv	0.885	0.871
	2	Glite	predictions_run_1.csv	0.887	0.871
	3	Glite	predictions_run_2.csv	0.895	0.868
	4	uogal	8enc_dmeta_elasticnet_..._preds.csv	0.903	0.869
	5	uogal	8enc_opt_weights_..._dev_preds.csv	0.944	0.865
2	6	Sakura	predictions_closed_max.csv	0.963	0.844
	7	NLP-Explorers	predictions_run_3.csv	0.992	0.845
	8	NLP-Explorers	predictions_run_1.csv	0.992	0.845
	9	NLP-Explorers	predictions_run_2.csv	0.993	0.845
	10	SAAKTH	predictions_run_3.csv	0.994	0.844
	11	TOEBM	predictions_run_3.csv	0.997	0.832
	12	SAAKTH	predictions_run_1.csv	1.000	0.831
	13	Jinnie's Lab	mbert-base_mtl_mlp_Mean_..._preds.csv	1.011	0.827
	14	SAAKTH	predictions_run_2.csv	1.012	0.843
3	15	Jinnie's Lab	mbert-base_mtl_mlp_Max_..._preds.csv	1.032	0.818
	16	SATLab	Boot.csv	1.036	0.830
	17	UOL@IDEM	run_2_freq.csv	1.037	0.834
	18	Jinnie's Lab	mbert-base_mtl_mlp_cls_de_preds.csv	1.039	0.815
	19	SATLab	Dev.csv	1.045	0.826
	20	UZHCL	grand_ensemble_ft_preds.csv	1.071	0.826
	21	TOEBM	predictions_run_2.csv	1.076	0.811
	22	UOL@IDEM	run_3_surprisal.csv	1.078	0.818
	23	UOL@IDEM	run_1_allfeatures.csv	1.079	0.819
	24	SATLab	Nonly.csv	1.103	0.812
	25	TeamXBC	TeamXBC_predictions_run_1.csv	1.114	0.837
	26	Data Asgardians	predictions_hp_hybrid.csv	1.117	0.830
	27	Boosted Cats	submission_de_run2.csv	1.118	0.783
4	28	Failure	predictions_run_1.csv	1.118	0.788
	29	Boosted Cats	submission_de_run1.csv	1.118	0.782
	30	TeamXBC	TeamXBC_predictions_run_2.csv	1.121	0.835
	31	Boosted Cats	submission_de_run3.csv	1.121	0.781
	32	EduNLP	predictions_closed_ensemble.csv	1.124	0.793
	33	EduNLP	predictions_closed_stretched.csv	1.125	0.793
	34	Sakura	predictions_explainable.csv	1.126	0.779
	35	Data Asgardians	predictions_closed_hybrid.csv	1.140	0.829
	36	UGA Threshold	predictions_run_3.csv	1.166	0.761
	37	Mixed Signals	numeric_ipa_de_1frz_preds.csv	1.173	0.784
	38	Mixed Signals	num_de_base_ipa_pos_meanings_preds.csv	1.174	0.784
	39	UGA Threshold	predictions_run_1.csv	1.176	0.757
	40	Data Asgardians	predictions_hp_solo.csv	1.177	0.822
5	41	Mixed Signals	numeric_de_base_ipa_pos_preds.csv	1.180	0.773
	42	UGA Threshold	predictions_run_2.csv	1.181	0.753
	43	Sakura	predictions_traditional.csv	1.195	0.747
	44	EduNLP	predictions_closed_transformer.csv	1.202	0.789
	45	TOEBM	predictions_aya_llamma_ensemble.csv	1.237	0.735
6	46	**BASELINE**	**baseline_closed_track.csv**	**1.258**	**0.773**
	47	RETUYT-InCo	predictions_xgboost.csv	1.260	0.713
	48	Unibuc	predictions_run_1.csv	1.305	0.692
	49	Unibuc	predictions_run_2.csv	1.309	0.692
7	50	Token Titans	predictions_run_1.csv	1.564	0.760
	51	Token Titans	predictions_ensemble.csv	1.569	0.772
	52	Token Titans	predictions_run_2.csv	1.569	0.772
8	53	TeamXBC	TeamXBC_predictions_run_3.csv	1.796	0.299

Table A10: Full results for Closed Track: German.

Group	Rank	Team Name	Prediction File	RMSE	Pearson
1	1	Glite	predictions_run_2.csv	0.776	0.889
	2	Glite	predictions_run_3.csv	0.785	0.886
	3	Glite	predictions_run_1.csv	0.788	0.885
	4	Sakura	predictions_closed_max.csv	0.816	0.874
	5	uogal	8enc_dmeta_elasticnet_..._preds.csv	0.820	0.879
2	6	uogal	8enc_opt_weights_..._dev_preds.csv	0.841	0.880
	7	TOEBM	predictions_run_3.csv	0.880	0.853
	8	NLP-Explorers	predictions_run_2.csv	0.882	0.861
	9	NLP-Explorers	predictions_run_1.csv	0.883	0.860
	10	NLP-Explorers	predictions_run_3.csv	0.885	0.859
	11	UOL@IDEM	run_2_freq.csv	0.891	0.860
3	12	SAAKTH	predictions_run_3.csv	0.900	0.860
	13	Jinnie's Lab	mbert-base_mtl_mlp_Max_..._cn_preds.csv	0.905	0.843
	14	SAAKTH	predictions_run_2.csv	0.911	0.859
	15	SAAKTH	predictions_run_1.csv	0.913	0.851
	16	UOL@IDEM	run_3_top1.csv	0.919	0.858
	17	Sakura	predictions_explainable.csv	0.920	0.837
	18	SATLab	Dev.csv	0.921	0.846
	19	SATLab	Boot.csv	0.924	0.843
	20	UOL@IDEM	run_1_allfeatures.csv	0.930	0.856
	21	TOEBM	predictions_run_2.csv	0.930	0.834
	22	Jinnie's Lab	mbert-base_mtl_mlp_mean_cn_preds.csv	0.932	0.832
4	23	TeamXBC	TeamXBC_predictions_run_1.csv	0.968	0.851
	24	UZHCL	grand_ensemble_ft_preds.csv	0.991	0.839
	25	TeamXBC	TeamXBC_predictions_run_2.csv	1.002	0.859
	26	Data Asgardians	predictions_hp_hybrid.csv	1.006	0.854
	27	TeamXBC	TeamXBC_predictions_run_3.csv	1.006	0.839
	28	Data Asgardians	predictions_hp_solo.csv	1.008	0.850
	29	Data Asgardians	predictions_closed_solo.csv	1.013	0.842
	30	Boosted Cats	submission_cn_run3.csv	1.030	0.790
	31	UGA Threshold	predictions_run_1.csv	1.031	0.791
	32	Boosted Cats	submission_cn_run1.csv	1.034	0.789
	33	TOEBM	predictions_aya_llamma_ensemble.csv	1.037	0.789
	34	Boosted Cats	submission_cn_run2.csv	1.039	0.786
5	35	Mixed Signals	num_ipa_cn_1frz_5b_preds.csv	1.055	0.790
	36	EduNLP	predictions_closed_ensemble.csv	1.058	0.800
	37	EduNLP	predictions_closed_stretched.csv	1.059	0.800
	38	UGA Threshold	predictions_run_3.csv	1.061	0.778
	39	Mixed Signals	numeric_cn_base_ipa_pos_preds.csv	1.062	0.792
	40	Mixed Signals	num_cn_base_ipa_pos_mean_preds.csv	1.062	0.777
	41	UGA Threshold	predictions_run_2.csv	1.072	0.770
	42	Sakura	predictions_traditional.csv	1.078	0.767
	43	Failure	predictions_run_1.csv	1.090	0.763
	44	SATLab	Nonly.csv	1.094	0.826
	45	RETUYT-InCo	predictions_xgboost.csv	1.106	0.754
	46	EduNLP	predictions_closed_transformer.csv	1.112	0.801
6	47	**BASELINE**	**baseline_closed_track.csv**	**1.140**	**0.753**
	48	Unibuc	predictions_run_1.csv	1.176	0.717
	49	Unibuc	predictions_run_2.csv	1.178	0.715
	50	Jinnie's Lab	mbert-base_mtl_mlp_mha_1_cn_preds.csv	1.185	0.710
7	51	Token Titans	predictions_run2.csv	1.382	0.718
8	52	Token Titans	predictions_run1.csv	1.436	0.777
9	53	BZPT	predictions_bert.csv	1.588	0.590

Table A11: Full results for Closed Track: Mandarin.

Group	Rank	Team Name	Prediction File	RMSE	Pearson
1	1	Sakura	predictions_finetuned_llms_plus.csv	0.742	0.919
	2	Sakura	predictions_open_max.csv	0.743	0.920
	3	Glite	predictions_run_2.csv	0.754	0.916
	4	Glite	predictions_run_1.csv	0.755	0.916
	5	Sakura	predictions_finetuned_llms.csv	0.760	0.915
2	6	Glite	predictions_run_3.csv	0.837	0.896
	7	TeamXBC	TeamXBC_predictions_run_3.csv	0.876	0.885
3	8	uogal	8enc_opt_weights_..._dev_preds.csv	0.975	0.858
	9	uogal	8enc_dmeta_elasticnet_..._preds.csv	0.977	0.859
	10	TeamXBC	TeamXBC_predictions_run_2.csv	1.003	0.846
4	11	SurreyCTS	predictions_run_3.csv	1.034	0.839
	12	SurreyCTS	predictions_run_1.csv	1.046	0.833
	13	Jinnie's Lab	mbert-base_mtl_mlp_mha_1_xx_preds.csv	1.053	0.829
	14	Jinnie's Lab	mbert-base_mtl_mlp_mean_xx_preds.csv	1.056	0.829
	15	Jinnie's Lab	mbert-base_mtl_mlp_Mean_..._xx_preds.csv	1.061	0.826
5	16	SurreyCTS	predictions_run_2.csv	1.083	0.829
	17	SATLab	Dev.csv	1.087	0.819
	18	SATLab	Mean.csv	1.088	0.819
	19	SATLab	Boot.csv	1.091	0.817
	20	TeamXBC	TeamXBC_predictions_run_1.csv	1.114	0.832
	21	EduNLP	predictions_open_ensemble.csv	1.143	0.802
	22	EduNLP	predictions_open_improved_es.csv	1.147	0.801
	23	EduNLP	predictions_open_mega_stretched.csv	1.150	0.804
6	24	**BASELINE**	**baseline_open_track.csv**	**1.198**	**0.783**

Table A12: Full results for Open Track: Spanish.

Group	Rank	Team Name	Prediction File	RMSE	Pearson
1	1	Sakura	predictions_open_max.csv	0.723	0.916
	2	Sakura	predictions_finetuned_llms_plus.csv	0.726	0.915
	3	Sakura	predictions_finetuned_llms.csv	0.731	0.914
2	4	Glite	predictions_run_2.csv	0.764	0.905
	5	Glite	predictions_run_1.csv	0.769	0.904
	6	Glite	predictions_run_3.csv	0.790	0.898
3	7	TeamXBC	TeamXBC_predictions_run_3.csv	0.826	0.888
4	8	uogal	8enc_dmeta_elasticnet_..._preds.csv	0.903	0.869
5	9	uogal	8enc_opt_weights_..._dev_preds.csv	0.944	0.865
	10	SurreyCTS	predictions_run_3.csv	0.945	0.854
	11	TeamXBC	TeamXBC_predictions_run_2.csv	0.947	0.856
	12	SurreyCTS	predictions_run_1.csv	0.968	0.844
	13	Jinnie's Lab	mbert-base_mtl_mlp_mean_xx_preds.csv	0.990	0.834
	14	Jinnie's Lab	mbert-base_mtl_mlp_mha_1_xx_preds.csv	1.002	0.830
6	15	SurreyCTS	predictions_run_2.csv	1.004	0.844
	16	Jinnie's Lab	mbert-base_mtl_mlp_Mean_..._xx_preds.csv	1.009	0.827
	17	SATLab	Boot.csv	1.010	0.828
	18	SATLab	Mean.csv	1.014	0.827
	19	SATLab	Dev.csv	1.022	0.824
	20	TeamXBC	TeamXBC_predictions_run_1.csv	1.042	0.848
7	21	EduNLP	predictions_open_ensemble.csv	1.071	0.816
8	22	EduNLP	predictions_open_mega_stretched.csv	1.089	0.815
9	23	**BASELINE**	**baseline_open_track.csv**	**1.166**	**0.786**
10	24	EduNLP	predictions_open_improved_de.csv	1.397	0.671

Table A13: Full results for Open Track: German.

Group	Rank	Team Name	Prediction File	RMSE	Pearson
1	1	Sakura	predictions_finetuned_llms_plus.csv	0.630	0.928
	2	Sakura	predictions_open_max.csv	0.631	0.927
	3	Sakura	predictions_finetuned_llms.csv	0.640	0.925
	4	Glite	predictions_run_1.csv	0.660	0.920
	5	Glite	predictions_run_2.csv	0.662	0.920
2	6	Glite	predictions_run_3.csv	0.679	0.916
3	7	TeamXBC	TeamXBC_predictions_run_3.csv	0.722	0.904
4	8	uogal	8enc_dmeta_elasticnet_..._preds.csv	0.820	0.879
	9	TeamXBC	TeamXBC_predictions_run_2.csv	0.830	0.880
5	10	uogal	8enc_opt_weights_..._dev_preds.csv	0.841	0.880
	11	SurreyCTS	predictions_run_3.csv	0.861	0.863
	12	SurreyCTS	predictions_run_1.csv	0.883	0.852
6	13	SurreyCTS	predictions_run_2.csv	0.885	0.866
	14	Jinnie's Lab	mbert-base_mtl_mlp_mha_1_xx_preds.csv	0.885	0.851
	15	Jinnie's Lab	mbert-base_mtl_mlp_mean_xx_preds.csv	0.895	0.847
	16	Jinnie's Lab	mbert-base_mtl_mlp_Mean_..._xx_preds.csv	0.899	0.845
	17	SATLab	Boot.csv	0.926	0.847
	18	SATLab	Mean.csv	0.926	0.847
	19	SATLab	Dev.csv	0.933	0.845
7	20	TeamXBC	TeamXBC_predictions_run_1.csv	0.943	0.864
	21	EduNLP	predictions_open_ensemble.csv	0.992	0.824
	22	EduNLP	predictions_open_mega_stretched.csv	1.007	0.825
8	23	**BASELINE**	**baseline_open_track.csv**	**1.034**	**0.804**
	24	BZPT	predictions_tool-augmented.csv	1.123	0.744
9	25	EduNLP	predictions_open_improved_cn.csv	1.421	0.623
10	26	BZPT	predictions_prompting.csv	1.845	0.044

Table A14: Full results for Open Track: Mandarin.