Glite at BEA 2026 Shared Task 1: Holistic Difficulty Models Dominate, Feature Engineering Closes the Gap in L1-Aware Vocabulary Difficulty Prediction

Vassili Philippov[1], **Dmitrii Andreev**[1], **Pavel Katunin**[1], **Anton Nikolaev**[2]
[1]Glite, {vassili, dmitrii, pavel}@glite.ai
[2]School of Biosciences, The University of Sheffield, Sheffield, UK
a.nikolaev@sheffield.ac.uk

Abstract

This paper describes our submission to the BEA 2026 Shared Task on L1-Aware English Vocabulary Difficulty Prediction. We build per-L1 CatBoost regressors over 1,161 candidate linguistic, psycholinguistic, dictionary, and LLM-derived features drawn from 129 feature sets; out-of-fold predictions from fine-tuned encoder and decoder-LLM regression heads enter the model as additional features. Features are selected via Recursive Feature Elimination with nested cross-validation, producing compact per-L1 models of 29–150 features per run. For the closed track we introduce a per-feature-column compliance audit that classifies 57 of 129 feature sets as track-eligible under the organiser rulings, an audit that forced a rebuild of the selection and ensembling pipelines in the final week. We further show that decoder-LLM LoRA regression heads — LLaMA-3.1-8B being the single strongest model in our pool — provide the largest marginal gains in the open track, and that a simpler per-L1 CatBoost on RFE-selected features matches or exceeds Ridge-stacking ensembles over the same base models. Our systems ranked 1st in the closed track and 2nd in the open track on all three L1s (Spanish, German, Mandarin), reducing baseline RMSE by 29.9% in the closed track and 35.9% in the open track on average.

1 Introduction

Vocabulary knowledge is a cornerstone of L2 language proficiency and a primary determinant of reading comprehension, listening, and writing ability. Reliable per-item difficulty calibration is the bottleneck for adaptive language assessment, CEFR-aligned curriculum design, learner-knowledge prediction, and personalised learning. Traditional calibration requires costly pretesting against representative learner populations; data-driven NLP methods offer a scalable alternative, provided they can be shown to recover the psychometric difficulty that learners actually experience.

The BEA 2026 Shared Task introduces two features absent from earlier shared tasks on word difficulty. First, *L1-awareness*: items are presented with a localised prompt in the learner's first language, and the task is to predict difficulty for Spanish, German, and Mandarin Chinese L1 speakers separately. Second, labels are *psychometrically calibrated*: the target is a GLMM-based difficulty estimate drawn from the British Council's Knowledge-based Vocabulary Lists (Schmitt et al., 2024), fit over roughly 3.3M test responses from more than 100,000 learners. Unlike most public vocabulary benchmarks, the task measures *productive* knowledge with an exact-spelling requirement: the learner must produce the target from a partial-spelling clue (first letter + word length), an L1 translation, and a localised context — a setup that materially shifts what difficulty is being measured. Prior work addressed related problems — CWI (Paetzold and Specia, 2016; Yimam et al., 2018), LCP (Shardlow et al., 2021), and MLSP (Shardlow et al., 2024) — without learner-specific context or psychometric targets. The task's closed/open split — closed (no LLMs, no extra data) vs. open (unrestricted) — adapts the Restricted/Unrestricted distinction used in the BEA 2019 grammatical-error-correction shared task (Bryant et al., 2019), and isolates the contribution of resource access.

We describe a system that places first in the closed track and second in the open track across all three L1s. The contributions are:

- **A closed-track system that places 1st on all three L1s**, reducing baseline RMSE by 28.2–31.9% (average 29.9%).
- **An open-track system placing 2nd on all three L1s**, with an average baseline reduction of 35.9%.
- **A per-feature-column closed-track compliance framework** applied to 129 feature sets,

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 1091–1105

classifying 57 as eligible and 72 as ineligible under the organiser rulings — a 56% rejection rate that required rebuilding selection and ensembling pipelines in the final week of the competition.

- **Empirical evidence that end-to-end holistic predictors are the dominant signal.** A leave-one-family-out ablation shows removing the holistic-models family costs +0.102 RMSE (an order of magnitude more than any other), driven by decoder-LLM LoRA regression heads — LLaMA-3.1-8B alone attains 0.831 at 0.13% trainable parameters. The holistic family is considerably larger than comparator families (68 vs 5–14 sets), which amplifies the ablation effect; per-set performance in Table 7 confirms the gap is not due to size alone. The second-largest lever is the *virtual-learner* family (+0.006 RMSE): weak LLMs used as proxy learners rather than difficulty judges.
- **A candid post-mortem of a data-leakage incident** in which four feature sets leaked the target variable into LLM prompts or cross-fold predictions; the leaked ensemble scored RMSE 0.609 before invalidation, and was caught by structural verification of every feature artefact.
- **An autonomous research framework** (§4) that evaluated more than 1,000 candidate features across 270+ tracked experiments, with spec-verified artefacts and an immutable audit trail.

Section 2 reviews prior shared tasks; Section 3 describes the data; Section 4 describes the system; Section 5 details the experiments; Section 6 presents results; and Section 7 analyses per-L1 errors, what worked, and what did not.

2 Related Work

Shared tasks on word difficulty. Complex Word Identification framed difficulty as binary classification for L2 readers (Paetzold and Specia, 2016; Yimam et al., 2018). Lexical Complexity Prediction moved to continuous regression with multiple annotators per item (Shardlow et al., 2021), and the Multilingual Lexical Simplification Pipeline extended the setup to multiple languages and the downstream simplification task (Shardlow et al., 2024). None of these tasks conditions prediction on the learner's L1 or uses psychometrically calibrated targets; the BEA 2026 task (Felice and Skidmore, 2026; Schmitt et al., 2024) is, to our knowledge, the first shared task to combine both.

Track structure. Our closed/open split follows the Restricted vs. Unrestricted distinction introduced by the BEA 2019 grammatical-error-correction shared task (Bryant et al., 2019), which also defined a third Low-Resource track. BEA 2019 enforced track compliance through submitted system descriptions; our closed-track compliance framework (§4.4) is a finer-grained realisation that audits eligibility at the level of individual feature columns.

Methodology precedents. Gombert et al. (2024) won the BEA 2024 Item Difficulty shared task with a transformer-encoder regression that uses scalar mixing across encoder layers, multi-task training on item difficulty and response time, and rational-network regression heads. Several other BEA 2024 systems stacked multiple encoder regressors via linear meta-learners. We combine elements of both patterns while diverging in three respects: (i) the pool is enlarged to include decoder-LLM LoRA regression heads alongside encoder regressions; (ii) per-L1 gradient boosting replaces the linear meta-learner as the fusion layer; and (iii) features are selected by Recursive Feature Elimination with nested cross-validation rather than all-in stacking, which reduces dimensionality before the final fit and limits overfitting on the relatively small (∼16k-row) out-of-fold validation set.

External resources. Individual feature families rely on established psycholinguistic norms and L2-oriented vocabulary lists, including concreteness and prevalence norms (Brysbaert et al., 2014, 2019), the SCOPE psycholinguistic metabase (Gao et al., 2023), the CEFR-J wordlist (Tono, 2019), CEFR-annotated WordNet (Kikuchi et al., 2025), Chinese–English L2 AoA norms (Wang and Chen, 2020), the MorphoLex morphological database (Sánchez-Gutiérrez et al., 2018), the LADEC database of English compounds (Gagné et al., 2019), the language-specific SUBTLEX subtitle frequency corpora for Chinese, German, and Spanish (Cai and Brysbaert, 2010; Brysbaert et al., 2011; Cuetos et al., 2011), and the OpenSubtitles parallel corpus (Lison et al., 2018). Our open-track submission additionally uses a proprietary sense-level English lexical resource with per-sense difficulty estimates. We cite all of these in-line in §4.2.

Split	# items / L1	# L1s	# rows
Train	6,091	3	18,273
Dev	677	3	2,031
Test	748	3	2,244
Total	7,516	3	22,548

Table 1: Dataset statistics. Item ids are shared across L1s, so the same English word appears in all three L1 views. Train, dev, and test splits are disjoint by item id (ranges 1–6091, 6092–6768, and 6769–7516 respectively).

3 Task and Data

Task. Given an English target word presented to a learner whose L1 is Spanish (ES), German (DE), or Mandarin Chinese (CN), together with a localised prompt (a partial-spelling clue, a translation, and a context sentence in the L1), predict the word's psychometric difficulty as a continuous GLMM score (lower score = more difficult). A canonical Spanish example presents the cue *"h_____"* alongside a Spanish context sentence and the translation *casa*, with the target English word being *house*. Responses are graded correct only when the spelling matches exactly, so orthographically irregular forms carry a spelling-error component in the GLMM difficulty alongside the lexical-knowledge component.

Data. The task distributes item-level data for 7,516 English target words, each presented in three L1 variants. Table 1 summarises the splits. The labels are derived from a GLMM fit over approximately 3.3M real learner responses from more than 100,000 British Council KVL test-takers (Schmitt et al., 2024). Test labels were withheld during the shared-task evaluation period and have since been released publicly on the shared-task GitHub repository.[1]

Metrics. The shared task ranks systems by Root Mean Squared Error (RMSE, lower is better) averaged across L1s, with Pearson correlation reported as a secondary metric. We use RMSE as the primary decision metric throughout: all feature selection, ensemble pruning, and submission choices are made on K-fold RMSE.

Tracks. The *closed* track forbids large language models and additional training data beyond the shared task release, while allowing publicly available off-the-shelf pretrained transformers and linguistic databases. The *open* track places no restrictions. A team may submit up to three runs per track per L1.

4 System

4.1 Overview

Our submission is a per-L1 gradient-boosted regressor consuming a broad pool of heterogeneous features, including the out-of-fold (OOF) predictions of fine-tuned encoder and decoder-LLM regression heads. Figure 1 gives the end-to-end picture. The feature pool comprises 129 feature sets totalling 1,161 numeric columns, organised by seven domain families plus a small derived bucket (§4.2). Per L1, features are reduced with Recursive Feature Elimination under nested cross-validation, yielding compact sets of 29–150 features that train a single CatBoost regressor (Prokhorenkova et al., 2018) on the union of train and dev.

Autonomous research framework. Our submission was developed inside the Glite Autonomous Research Framework.[2] The framework enforces task isolation (one directory, one branch, one pull request per experiment), artefact specifications checked by verifiers before commit, and an immutable log of every experiment step. This enabled us to evaluate over 1,000 candidate features across 270+ tracked experiments, with automatic metric aggregation and append-only audit trails. The data-leakage incident described in §7.3 was detected precisely because every feature CSV is spec-verified and every fold-level score is traceable to its code revision.

4.2 Features

We organise our 129 feature sets around *what the feature measures about item difficulty* rather than how it was produced. Seven domain families — plus a small derived bucket — follow. Appendix B has per-family counts and the top-ranked sets per family (Table 7).

Target-word lexical properties (14 sets). Surface form, frequency, morphology, and semantic-lexical properties of the English target word itself: word length, corpus log-frequency, morphological decomposition (Sánchez-Gutiérrez et al., 2018), compound transparency (Gagné et al., 2019), neighbourhood density, concreteness (Brysbaert et al.,

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

[2] https://github.com/GliteTech/glite-arf

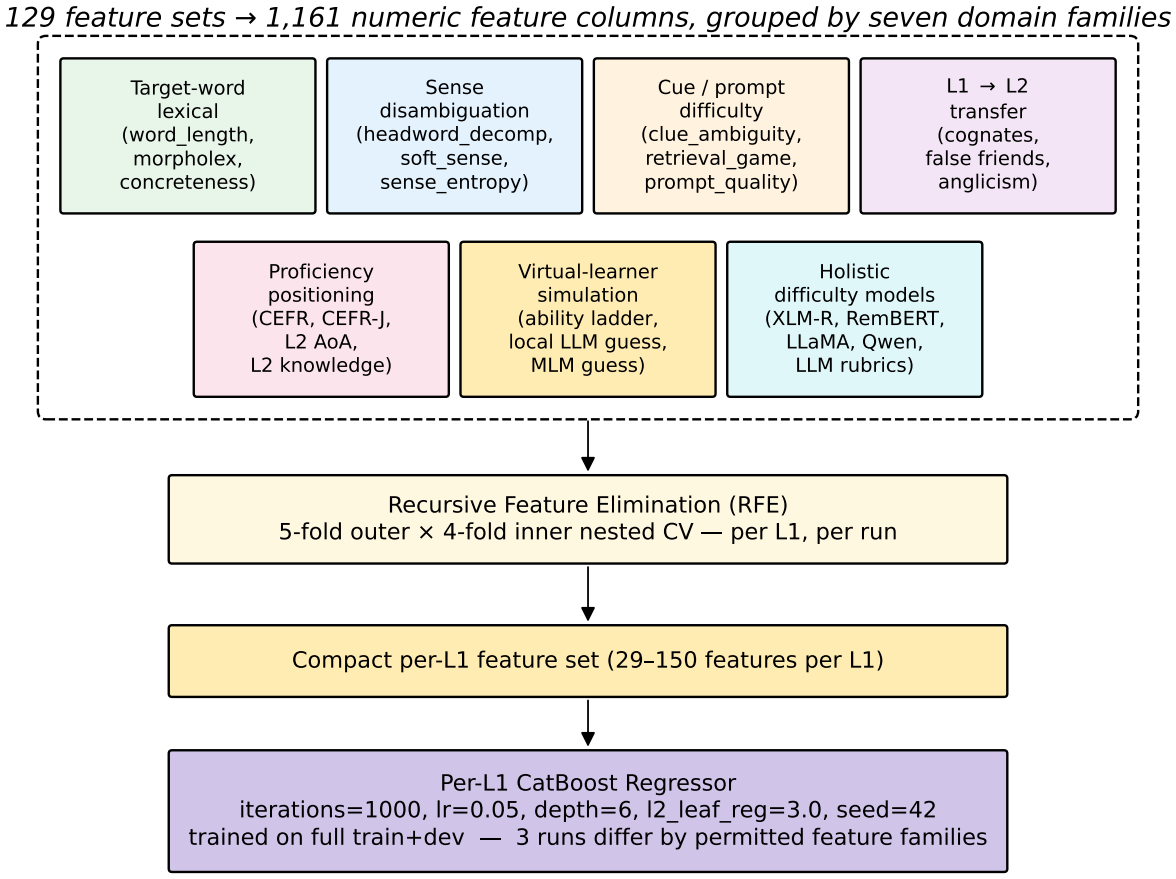

Figure 1: System architecture. Seven feature families feed into per-L1 RFE with nested CV, which yields compact per-L1 feature sets consumed by CatBoost. Transformer and decoder-LLM signals enter as out-of-fold predictions rather than via a separate stacking layer. Three submission runs differ in which feature sets and models are permitted.

2014), prevalence (Brysbaert et al., 2019), the SCOPE psycholinguistic metabase (Gao et al., 2023), and lemma-level features from a proprietary sense-level English lexical resource. These are the features a lexicographer would call out first when asked why a word is hard.

Sense-level disambiguation (7 sets). For polysemous words the tested sense is often not the dominant sense, so features that model *which* sense is tested matter. Each KVL item is first mapped to a sense in a proprietary English dictionary via an LLM word-sense-disambiguation prompt given the target, clue sentence, and L1 translation. The family then covers sense-level dictionary features (212 columns; solo Pearson 0.704), morphological headword decomposition, soft sense distributions, sense entropy, and a polysemy-contrast LLM feature.

Cue / prompt difficulty (6 sets). Features that analyse the *clue* (partial spelling + L1 translation + L1 context), not the target word itself: clue-ambiguity (dictionary words that match the same first-letter + word-length pattern), a retrieval-game information-theoretic model of candidate competition, semantic-competition counts of viable alternatives, LLM-rated prompt quality, prompt-framing variants, and a context-only guess feature. These capture how tightly the cue constrains the answer.

L1 → L2 transfer (14 sets). Cross-linguistic features about whether the learner's L1 helps or hurts: cognate distance, CogNet-detected cognates, a Spanish expert cognate list, anglicism detection, false-friend flags, LLM language-similarity ratings, phonological transfer, cross-lingual orthographic neighbourhood, Chinese script-complexity, L1-side frequency (corpus and subtitle), etymology, translation ambiguity, and LLM back-translation quality.

L2 proficiency positioning (8 sets). Empirical evidence of where the target word sits on

the learner- developmental scale: CEFR levels from EFLLex and Words-CEFR (Tono, 2019), the CEFR-J / JACET Asian-learner wordlists, sense-level CEFR from CEFR-annotated WordNet (Kikuchi et al., 2025), Brysbaert's L2 yes/no word-knowledge norms, L2 age of acquisition (Wang and Chen, 2020), empirical learner production rates, and Cambridge cloze-exam presence. Every set in this family is closed-track eligible.

Virtual-learner simulation (7 sets). A distinctive family in our system: features produced by letting a weaker or ability-graded model *attempt to answer* the item and recording whether it succeeds, its confidence, or its surprisal. The LLM here acts as a proxy learner, not as a difficulty judge. This family includes an ability ladder over four OpenAI models of graded size, guesses from four local small models, LLM clue- and description-based guessing, LLM answer entropy, token-surprisal features, and masked-LM pseudo-likelihood guessing. All are open-track-only except the masked-LM variant (closed-eligible).

Holistic difficulty models (68 sets). End-to-end predictors that read the full item and emit a scalar difficulty estimate, entering the pipeline as out-of-fold features. Covers fine-tuned encoders — XLM-R (Conneau et al., 2020), RemBERT (Chung et al., 2021), ELECTRA-base (Clark et al., 2020), and mBERT per L1 with TAPT (Gururangan et al., 2020) and R-Drop (Liang et al., 2021) under multiple seeds; scaled encoders (XLM-R-XL 3.5B, XLM-R-XXL 10.7B) trained with LoRA (Hu et al., 2022); decoder-LLM LoRA regression heads for LLaMA-3.1-8B (Dubey et al., 2024), Qwen-2.5-7B (Yang et al., 2024), and Mistral-7B (Jiang et al., 2023); LLM rubric variants; and LLM embeddings. LoRA uses rank 16, $\alpha = 16$, targeting `q_proj/k_proj/v_proj` (0.13% trainable) with a scalar regression head on the last-token hidden state. LLaMA-3.1-8B achieves the single strongest K-fold RMSE (0.831), beating the 10.7B XLM-R-XXL encoder (0.966) despite fewer parameters; Mistral-7B scores 0.951 on average but suffers a fold-4 collapse (§7.3).

Derived / second-order (5 sets). Features whose inputs are other features' outputs: pairwise interactions, model-disagreement metrics, per-item CatBoost feature-group ablation sensitivities, and a Ridge distillation from the SemEval-2021 LCP dataset.

4.3 Feature selection and final model

Selection. For each L1 and each submission run, we run Recursive Feature Elimination with nested cross-validation (5-fold outer, 4-fold inner). The outer folds partition item ids; inner folds select the smallest feature set whose K-fold RMSE is within 0.001 of the optimum. The resulting per-L1 feature counts are summarised in Appendix A.

Final model. For each (track, run, L1) combination we train a single CatBoost regressor with `iterations=1000`, `learning_rate=0.05`, `depth=6`, `l2_leaf_reg=3.0`, and `seed=42` on the union of the train and dev splits. No early stopping is used. The CatBoost feature matrix contains the RFE-selected subset of linguistic, psycholinguistic, and dictionary features, plus the OOF predictions of the fine-tuned transformer and decoder-LLM regression heads treated as ordinary numeric features.

Architecture alternatives considered. A 10-model Ridge meta-learner over three global CatBoost variants plus seven transformer OOF columns (per-L1 $\alpha = 500, 500, 100$) scored K-fold RMSE 0.745 versus 0.740 for the CatBoost pipeline. Its coefficients corroborated two observations used later in §7.3: gradient-boosted features dominate ($\approx 88\%$ of total weight), and Mistral-7B receives near-zero weight. We submitted the simpler CatBoost pipeline; a greedy pruning over the 63-model Ridge pool (Figure 2) further confirms that six models match the full ensemble.

Three submission runs per track. The three runs differ in which feature sets and models are permitted at RFE time. Run 1 is the full RFE-optimal pipeline. Run 2 is a resource-robust variant: for the open track it excludes the proprietary dictionary; for the closed track it excludes the large XLM-R-XL encoder. Run 3 is the most conservative: transformers-only for the open track, and a larger feature-rich RFE set for the closed track. Appendix A lists per-run feature counts and K-fold RMSE.

4.4 Closed-track compliance engineering

The closed track prohibits generative LLMs, paid APIs, additional training data, and cross-L1 training, while allowing publicly available off-the-shelf pretrained transformer encoders, standard NLP tools, and public linguistic databases (WordNet, CEFR wordlists, SUBTLEX, EFLLex, MorphoLex,

LADEC, SCOPE, etc.). Feature selection and final CatBoost training were performed separately for each L1, using only that L1's provided train and dev rows. Because features are generated from heterogeneous code paths and our feature pool is large, we developed a per-column compliance schema that records, for each feature column, which external models, APIs, datasets, and computation paths were used to produce it. Every feature set is audited against five organiser rulings issued in March 2026: (1) difficulty-in-prompt leakage, (2) LLMs predicting the target word, (3) paid embedding APIs, (4) LLMs estimating learner production probability, and (5) LLMs used purely for rubric scoring without the target. Seventeen features were reclassified after these rulings, and two further reclassifications came from our internal audit (inadvertent training-data contamination and external-data usage). Of 129 feature sets (1,161 columns), 57 (249 columns) remained closed-track-eligible after the audit — a 56% rejection rate that forced a full rebuild of the RFE and submission pipelines in the final week. A 469-line closed-track LLM-usage disclosure submitted with our system enumerates all 14 pretrained models used in the closed-track submission; all are encoder-only or encoder-decoder, with no generative LLMs and no paid APIs. We view this per-column audit as a generalisable methodological contribution for any shared task with resource restrictions.

5 Experiments

Folds. We use a single K=5 fold assignment over item ids, shared across L1s, so that any English word appears in the same fold in all three L1 views. This prevents cross-L1 leakage via joint training. The fold assignments are persisted deterministically and reused across every experiment in the project.

Hyperparameters. CatBoost uses `iterations=1000`, `lr=0.05`, `depth=6`, `l2_leaf_reg=3.0`, `seed=42`. Encoder regressors (XLM-R, RemBERT, ELECTRA, mBERT) use batch size 16 (or 4 for XL/XXL), max length 128, AdamW with linear warmup and decay, 3 epochs for per-L1 variants with TAPT+R-Drop ($\lambda = 1.0$), and seed sweep over $\{123, 456, 789\}$. LoRA adapters target `q_proj/k_proj/v_proj` with rank 16, $\alpha = 16$, dropout 0.1; decoder-LLM regression heads read the last-token hidden state through an MLP of width 2048 with GELU. Full hyperparameters are listed in Appendix C.

Hardware and compute. CatBoost models and small-encoder fine-tuning ran locally on Mac Silicon (48 GB unified memory); scaled encoders (XLM-R-XL 3.5B, XLM-R-XXL 10.7B) and the 7–8B decoder-LLM regression heads (LLaMA-3.1-8B, Qwen-2.5-7B, Mistral-7B) were trained on rented A100-class GPUs. Total compute was approximately 100 wall-hours and roughly $500 in commercial LLM API spend.

6 Results

Main result. Our system placed 1st in the closed track on all three L1s and 2nd in the open track on all three L1s (Table 2). The average RMSE reduction over the official baseline is 29.9% closed (28.2% ES, 29.6% DE, 31.9% CN) and 35.9% open (37.1% ES, 34.5% DE, 36.2% CN).

Ablation by domain family (Table 3). Table 3 reports two complementary analyses. *Leave-one-out* (left block) removes one family at a time and re-runs RFE; *family alone* (right block) keeps only that family and runs RFE over it. Together these measure *necessity* and *sufficiency*. *Holistic difficulty models* are both: removing the family costs $+0.102$ RMSE (an order of magnitude more than any other), and alone the family attains K-fold RMSE 0.772 / Pearson 0.900 with just 29 features — within 0.03 of the full system. *Virtual-learner simulation* is the clearest second-tier contributor: $+0.006$ RMSE to remove and Pearson 0.743 alone. The other five families show a striking asymmetry: their leave-one-out costs are small ($+0.001$ to $+0.003$), yet alone they reach Pearson 0.68–0.78 — classical-tabular signal that is largely absorbed by the holistic models when present but remains substantive when isolated. *Cue / prompt difficulty* is the strongest standalone classical family (r 0.782).

Top features per L1. The top-15 RFE features per L1 (Appendix D, Table 10) reveal a universal core: all four LLaMA-3.1-8B seeds, all four Qwen-2.5-7B seeds, the ability-ladder mean success rate, the LLM-based counterfactual cue-sensitivity score, and the learner's L2 vocabulary-knowledge rank — all appear in the top 15 for every L1. L1-specific features surface in the mid-ranks: surprisal and interaction features for ES, an additional XLM-R-XL variant for DE, and a frequency-based cue-sensitivity variant plus a closed-eligible per-L1 XLM-R-large-TAPT model for CN.

Track	System	ES		DE		CN		Avg
		RMSE	r	RMSE	r	RMSE	r	RMSE
Closed	**1. Ours**	**0.903**	**0.877**	**0.885**	**0.871**	**0.776**	**0.889**	**0.855**
Closed	2. Best other team[†]	0.975	0.858	0.903	0.869	0.816	0.874	0.898
Closed	3. Best other team[‡]	0.976	0.857	0.963	0.844	0.820	0.879	0.920
Closed	Baseline	1.257	0.765	1.258	0.773	1.140	0.753	1.218
Open	1. Sakura	**0.742**	**0.919**	**0.723**	**0.916**	**0.630**	**0.928**	**0.698**
Open	**2. Ours**	0.754	0.916	0.764	0.905	0.660	0.920	0.726
Open	3. TeamXBC	0.876	0.885	0.826	0.888	0.722	0.904	0.808
Open	Baseline	1.198	0.783	1.166	0.786	1.034	0.804	1.133

Table 2: Official test-set leaderboard excerpt, summarised at team level using each team's best submitted run per L1. RMSE is the primary ranking metric (lower is better); r is Pearson. **Bold** marks the best per-column entry for that track. Closed track had 18–19 unique participating teams per L1; open track had 6–7. [†]The 2nd-best closed-track team is *uogal* for ES and DE, and *Sakura* for CN. [‡]The 3rd-best closed-track team is *AIDA* for ES, *Sakura* for DE, and *uogal* for CN.

	Leave-one-out		Family alone		
Configuration	K-fold RMSE	Δ RMSE	#feats	K-fold RMSE	Pearson
Full (all 7 families)	**0.740**	—	—	—	—
− Holistic difficulty models	0.842	+0.102	29	0.772	0.900
− Virtual-learner simulation	0.746	+0.006	75	1.194	0.743
− L2 proficiency positioning	0.743	+0.003	34	1.286	0.684
− Cue / prompt difficulty	0.742	+0.002	47	1.105	0.782
− Sense disambiguation	0.742	+0.002	62	1.233	0.718
− L1 $\rightarrow$ L2 transfer	0.741	+0.001	54	1.261	0.696
− Target-word lexical	0.741	+0.001	121	1.276	0.693

Table 3: Open-track ablation by domain family (§4.2). Left block (*leave-one-out*) deletes the named family and re-runs RFE over all remaining features, reporting the rebuilt system's K-fold RMSE and its gap to the full system (full system: 66 RFE-selected features, K-fold RMSE 0.740). Right block (*family alone*) keeps only the named family and runs RFE over its features, reporting the optimal feature count, K-fold RMSE, and Pearson of the family-only system. Same folds, hyperparameters, and RFE procedure throughout. Rows sorted by Δ RMSE (most damaging removal first).

K-fold $\rightarrow$ test calibration. Sixteen of 18 submitted runs scored *better* test RMSE than K-fold RMSE (mean $\Delta = -0.022$ closed, -0.012 open); test Pearson exceeded K-fold Pearson on every comparable run (Appendix B). The two runs where test was worse than K-fold have small magnitude ($+0.001$ and $+0.015$ RMSE). We therefore find no empirical evidence of substantial K-fold overfitting across our 1,000+ feature search, and this pattern is consistent with the leak audit described in §7.3, since a residual leak would inflate K-fold relative to test.

Greedy ensemble pruning. For the 10-model Ridge architecture considered in §4.3 we independently ran greedy forward selection over a larger 63-model Ridge pool (Appendix A, Figure 2). The best single model achieves 0.814 (a CatBoost RFE variant). Adding the second (LLaMA-3.1-8B) drops RMSE by 0.056. Six models suffice to match the full 63-model ensemble (0.745 vs 0.745). The optimum is reached at 39 models (0.739). Beyond that, the next 24 additions are neutral or harmful, and the 63rd model added — `mistral_7b` — is the single worst addition, raising RMSE by 0.004.

Gap to first place in the open track. Sakura placed first on all three open-track L1s with test RMSE 0.742 / 0.723 / 0.630 (ES / DE / CN). Our second-place gaps are small on Spanish (+0.012) but widen on German (+0.041) and Mandarin (+0.030). We suspect Sakura's fine-tuned-LLM pipeline captures L1-specific morphology — Germanic compounding in particular — that our classical tabular features under-represent; closing these gaps is a concrete direction for future work.

Closed-track margin is wider. Our closed-track first-place margins over the second-best team (*uogal* for ES/DE, *Sakura* for CN) are 0.072 / 0.018 / 0.040 RMSE — roughly $3\times$ our open-track gap to Sakura on ES and CN, and about $4\times$ on DE.

This asymmetry is consistent with the closed-track restrictions penalising systems whose open-track advantage depends on paid LLM APIs: once generative LLMs are removed, the per-column compliance audit (§4.4) and our 57-set eligible pool produce a feature base competitive enough to lead each L1.

7 Analysis

7.1 Per-L1 error patterns

The three L1s show systematically different residual structure. Spanish items with Latinate roots are under-predicted as easier than they are; items with Germanic roots or noun morphology shared with French cognates are the opposite. German residuals are reduced most by the length-ratio feature and by the XLM-R-XL OOF column, which carries German-sensitive morphology signal absent from the 7B decoder-LLMs. Mandarin has the lowest RMSE on all systems — the CN target distribution is narrower — and benefits least from cognate-style features, instead drawing on the character-level Chinese script-complexity feature and per-L1 RemBERT OOF. All three L1s exhibit regression-to-mean compression of roughly 20% (slope ≈ 0.80) that shrank between ensemble versions but did not vanish (quantified further in the Limitations section).

7.2 What worked

Holistic predictors are necessary and nearly sufficient. Seven of the top-10 RFE-ranked features for every L1 are decoder-LLM OOF columns. A single LLaMA-3.1-8B head (0.831 K-fold RMSE) beats XLM-R-XXL (0.966, 42% more parameters) at 0.13% trainable parameters; adding Qwen-2.5-7B drops RMSE a further 0.010. The ceiling analysis in Table 3 sharpens this: an RFE pass restricted to the 68 holistic sets converges to 29 features at K-fold RMSE 0.772 / Pearson 0.900 — within 0.03 RMSE of the full system. In greedy forward selection LLaMA is the second model added (-0.056 RMSE) and Qwen the third (-0.010).

Virtual-learner simulation is a distinct signal. Removing the *virtual-learner* family costs $+0.006$ RMSE (Table 3), larger than any classical-tabular or cue/sense family; alone the family reaches Pearson 0.743. Letting a weak LLM *attempt* the item therefore adds signal the decoder-LLM regression heads do not already absorb — a different way of using LLMs from the dominant regression-head paradigm.

Simpler ensemble beat Ridge stacking. A single per-L1 CatBoost consuming transformer and LLM OOF predictions as plain features matched or exceeded every Ridge-stacking alternative; greedy pruning shows only 6 of 63 models suffice to saturate the Ridge ensemble.

Column-level compliance preserved signal. Operating the March 2026 compliance audit at the column level rather than the feature-set level preserved 249 columns across 57 closed-eligible sets that a set-level audit would have discarded, including the per-L1 RemBERT and XLM-R-large-TAPT OOF columns that dominate closed-track feature importance.

7.3 What did not work

Data-leakage post-mortem. Our ensemble v13 reached K-fold RMSE 0.609 — implausibly strong. Four leaking feature sets were identified: two passed the GLMM target into LLM prompts asking the model to judge annotation quality; a third used cross-L1 out-of-fold predictions with an item-id split shared across L1s; a fourth produced residual-prediction features circular with the training loss. Quarantining all four and rebuilding every downstream artefact yielded a clean v14 at 0.802. We cannot rule out undetected remaining leaks, but the K-fold $\rightarrow$ test calibration (§6, Appendix B) is consistent with the audit: test RMSE beats K-fold on 16 of 18 runs, whereas a residual leak would invert that pattern.

Mistral-7B is actively harmful in the ensemble. Of three decoder LLMs trained with the same LoRA recipe, LLaMA (0.831) and Qwen (0.840) were useful, but Mistral-7B (0.951) was not. Its average hides a fold-4 collapse to 1.36 (folds 0–3: 0.83–0.89), suggesting an initialisation pathology rather than a data issue. In greedy forward selection over the 63-model Ridge pool Mistral is the final addition and adds +0.004 RMSE — the single largest harmful contribution — and a parallel 10-model Ridge independently assigns it near-zero weight. The submitted system excludes Mistral.

Seven recurring failure patterns. Across 37 negative experiments we observed seven recurring patterns: (i) transformer architectural surgery (scalar-mix, multi-task, CharCNN) made no difference over a tuned XLM-R-base; (ii) adversarial

training did not help at 20k rows; (iii) tabular features saturated at $\sim$20 sets; (iv) derived features (SHAP, counterfactual sensitivity, residual regressors) tended to be circular with the target; (v) only 3 of 13 psycholinguistic databases added marginal signal; (vi) HPO and bagging moved Pearson by ≤ 0.001 per 100 trials; (vii) Pinyin/script normalisation destroyed information XLM-R's tokeniser handles natively.

8 Conclusion

Our system placed 1st closed / 2nd open on all three BEA 2026 L1s. The architecture is a per-L1 CatBoost on RFE-selected features plus OOF predictions from fine-tuned transformer and decoder-LLM regression heads. Three findings stand out: decoder-LLM LoRA heads dominate open-track importance (LLaMA-3.1-8B strongest); column-level compliance preserved 249 closed-eligible columns a set-level audit would have rejected; and six models match a 63-model Ridge ensemble, with Mistral-7B actively degrading it. Future work: distil the decoder-LLM signal into closed-track-legal form, and extend to more L1s.

Limitations

Our open-track system relies on a proprietary sense-level English lexical resource with difficulty estimates; reproducing the exact open-track numbers requires licensing of this resource. The closed-track system uses only public resources and can be reimplemented from the feature descriptions and hyperparameters given in §4 and Appendix C. We evaluated only three L1s — Spanish, German, and Mandarin Chinese — all drawn from the British Council KVL test-taker population; generalisation to typologically distant L1s such as Arabic or Japanese, and to learner populations outside paid-test-taker demographics, is untested. Our feature search explored more than 1,000 candidate features using autonomous agents under human supervision; while every artefact is spec-verified and every cross-validation fold is deterministic, the scale of the search creates an *a priori* risk of overfitting to the K-fold protocol. In practice this risk did not materialise: 16 of 18 submitted runs scored better on the held-out test set than on K-fold (mean Δ RMSE = -0.017), and test Pearson exceeded K-fold Pearson on every comparable run (Appendix B) — indicating that nested cross-validation plus dev-set diagnostics were sufficient for calibration at this search scale. Our best ensembles exhibit roughly 20% regression-to-mean compression (slope ≈ 0.80): the hardest items are systematically under-predicted by -0.55 GLMM units and the easiest over-predicted by $+0.43$; this pattern shrank between ensemble versions but did not disappear. Several of our strongest open-track features rely on commercial LLM APIs (approximately $500 total across the project) and on GPU-hosted decoder-LLM regression heads requiring 55–83 GB of VRAM, available only via cloud GPU rental; despite our best submission reducing baseline RMSE by 35.9% in the open track and 29.9% in the closed track, per-item residuals of 0.6–0.9 RMSE remain, indicating the task is far from solved.

Acknowledgments

We thank the BEA 2026 Shared Task organisers — Lucy Skidmore, Mariano Felice, and Karen Dunn — for designing the task and releasing the KVL-based dataset, and the British Council for making the underlying psychometric data publicly available under a Creative Commons licence. We also thank our colleagues at Glite for supporting the development of the autonomous research framework used to coordinate the experiments reported here, and the anonymous reviewer whose comments on the submission version of this paper improved the exposition of §3 and Figure 1.

References

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA 2019 shared task on grammatical error correction. In *Proceedings of BEA 2019*.

Marc Brysbaert, Matthias Buchmeier, Markus Conrad, Arthur M. Jacobs, Jens Bölte, and Andrea Böhl. 2011. The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, 58(5):412–424.

Marc Brysbaert, Paweł Mandera, Samantha F. McCormick, and Emmanuel Keuleers. 2019. Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*.

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

Qing Cai and Marc Brysbaert. 2010. SUBTLEX-CH: Chinese word frequencies based on film subtitles. *PLoS ONE*, 5(6):e10729.

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. In *Proceedings of ICLR*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *Proceedings of ICLR*.

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 ACL*.

Fernando Cuetos, María Glez-Nosti, Analia Barbón, and Marc Brysbaert. 2011. SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicológica*, 32:133–143.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The LLaMA 3 herd of models. *arXiv:2407.21783*.

Mariano Felice and Lucy Skidmore. 2026. Findings of the BEA 2026 shared task on vocabulary difficulty prediction for English learners. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, San Diego, California. Association for Computational Linguistics.

Christina L. Gagné, Thomas L. Spalding, and Daniel Schmidtke. 2019. LADEC: The large database of english compounds. *Behavior Research Methods*.

Chuanji Gao, Svetlana V. Shinkareva, and Rutvik H. Desai. 2023. SCOPE: The South Carolina psycholinguistic metabase. *Behavior Research Methods*, 55:2853–2884.

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 BEA 2024*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*.

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 *Proceedings of ICLR*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, and 1 others. 2023. Mistral 7B. *arXiv:2310.06825*.

Masato Kikuchi, Masatsugu Ono, Toshioki Soga, Tetsu Tanabe, and Tadachika Ozono. 2025. CEFR-annotated WordNet: LLM-based proficiency-guided semantic database for language learning. *arXiv preprint arXiv:2510.18466*.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. 2021. R-Drop: Regularized dropout for neural networks. In *Proceedings of NeurIPS*.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of LREC*.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 task 11: Complex word identification. In *Proceedings of SemEval-2016*.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost: unbiased boosting with categorical features. In *Proceedings of NeurIPS*.

Claudia H. Sánchez-Gutiérrez, Hugo Mailhot, S. Hélène Deacon, and Maximiliano A. Wilson. 2018. MorphoLex: A derivational morphological database for english. In *Behavior Research Methods*.

Norbert Schmitt, Karen Dunn, Barry O'Sullivan, Laurence Anthony, and Benjamin Kremmel. 2024. *Knowledge-based Vocabulary Lists*, volume 5 of *British Council Monographs on Modern Language Testing*. University of Toronto Press.

Matthew Shardlow, Richard Evans, Gustavo Paetzold, and Marcos Zampieri. 2021. SemEval-2021 task 1: Lexical complexity prediction. In *Proceedings of SemEval-2021*.

Matthew Shardlow, Sanja Štajner, Kai North, Tharindu Ranasinghe, and Marcos Zampieri. 2024. The BEA 2024 shared task on the multilingual lexical simplification pipeline. In *Proceedings of BEA 2024*.

Yukio Tono. 2019. CEFR-J wordlist version 1.5. Tokyo University of Foreign Studies, Tono Laboratory. `http://www.cefr-j.org/download.html`.

Jue Wang and Baoguo Chen. 2020. A database of Chinese-English bilingual speakers: Ratings of the age of acquisition and familiarity. *Frontiers in Psychology*, 11:554785.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, and 1 others. 2024. Qwen2.5: A party of foundation models. *arXiv:2412.15115*.

Seid Muhie Yimam, Chris Biemann, Sanja Štajner, Sian Gooding, and Matthew Shardlow. 2018. A report on the complex word identification shared task 2018. In *Proceedings of BEA 2018*.

A Submitted runs

Tables 4–5 give the per-run feature composition and per-L1 K-fold RMSE for every submitted run.

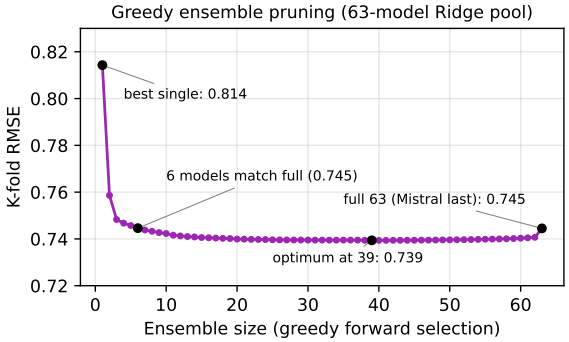

Figure 2: Greedy forward selection over a 63-model Ridge pool. Six models already match the full ensemble; the optimum is reached at 39; the last model added (Mistral-7B) is the single most harmful.

B Feature catalogue

Due to space constraints, we do not list the full 129-set catalogue here. Table 6 summarises per-family statistics, and Table 7 drills into the top feature sets in each family.

Table 7 drills into the top feature sets in each family, showing the single-set K-fold Pearson correlation and RMSE against the GLMM target. Decoder-LLM regression heads dominate, and the top closed-track-eligible signals are per-L1 encoder OOFs (RemBERT and XLM-R-large TAPT); classical linguistic and psycholinguistic features saturate around Pearson 0.55–0.59 individually, well below the transformer band.

K-fold vs test generalisation. Table 8 gives the full per-run comparison between our K-fold CV metrics (on the 5-fold item-id split of train+dev) and the official test-set metrics released by the organisers. All 18 submitted runs are listed. Sixteen of 18 runs have negative Δ RMSE (the two exceptions, open-track ES Run 3 and DE Run 1, are small in magnitude); all comparable Pearson values improve on test. K-fold Pearson is missing for Runs 2–3 where documentation recorded only RMSE.

C Hyperparameters

D Per-L1 RFE importance

Table 10 gives the top-15 RFE-ranked features per L1 for the open-track submitted system. Our RFE pipeline records the full elimination trajectory (elimination order, per-round RMSE, optimal cut) per L1 and per track for each submitted run.

E Representative LLM prompts

We illustrate two representative prompt templates covering the two main families of LLM-derived features.

Rubric scoring (holistic LLM rater).

```
You are an expert L2 English teacher.
Rate how difficult the target word is
for an adult learner whose first language
is {{L1}}, on a 0-10 scale. 0 = a near-
beginner will know it; 10 = only an
advanced learner will.

Target word: {{word}}
Part of speech: {{pos}}
Context sentence (English): {{clue}}
L1 translation(s): {{l1_word}}

Output only the integer score.
```

Ability-ladder (open track only). Four OpenAI models of graded ability are each asked to attempt the item; the feature records the fraction that recovered the target. The target word is *not* supplied; only the clue, L1 translation, and context are. Because the feature invokes generative LLMs, it is ineligible for the closed track regardless of whether the model ever sees the label.

F Compliance audit

Table 11 lists the categories of the 17 features reclassified as ineligible after the March 2026 organiser rulings. Per-column compliance metadata was maintained throughout for every feature set, including a 469-line closed-track LLM-usage disclosure submitted with our system.

G LLM API cost breakdown

Table 12 lists the ten most expensive LLM-API-based feature experiments and their positive/negative status. Total LLM spend across the project was approximately $498. Local MLX inference (llama1b, gemma2b, phi3b, qwen1b for the local-LLM guessing feature) and GPU training were separate and not included in this table.

		# features			K-fold RMSE			
Run	Strategy	ES	DE	CN	ES	DE	CN	Avg
1	Compact (RFE-optimal)	58	49	48	0.775	0.768	0.676	0.740
2	No proprietary dict.	66	86	47	0.781	0.770	0.683	0.745
3	Transformers only	8	9	37	0.822	0.796	0.703	0.774

Table 4: Open-track runs.

		# features			K-fold RMSE			
Run	Strategy	ES	DE	CN	ES	DE	CN	Avg
1	Compact (RFE-optimal)	29	49	30	0.925	0.916	0.805	0.882
2	No XLM-R-XL	30	36	33	0.923	0.912	0.805	0.880
3	Feature-rich	50	100	150	0.926	0.916	0.814	0.885

Table 5: Closed-track runs.

				Solo r		Family alone	
Family	Sets	Cols	Closed-elig.	Best set	Mean set	K-fold RMSE	Pearson
Target-word lexical	14	149	12	0.660	0.370	1.276	0.693
Sense disambiguation	7	248	1	0.704	0.374	1.233	0.718
Cue / prompt difficulty	6	50	2	0.569	0.520	1.105	0.782
L1 $\rightarrow$ L2 transfer	14	57	9	0.442	0.307	1.261	0.696
Proficiency positioning	8	36	8	0.586	0.507	1.286	0.684
Virtual-learner simulation	7	108	1	0.576	0.458	1.194	0.743
Holistic difficulty models	68	432	24	0.884	0.752	0.772	0.900
Derived / second-order	5	81	0	0.835	0.740	—	—
Total	129	1,161	57	0.884		0.740	—

Table 6: Per-domain-family totals under the seven-family taxonomy (plus a small Derived/second-order bucket for features that operate over other families' outputs). *Solo r* columns report the best and mean single-set K-fold Pearson within the family, taken from each set's `feature_info.json`; *Family alone* reports an RFE pass over the family-only feature pool (Table 3 right block). The *Total* row's *Family alone* RMSE (0.740) is the full-system K-fold result over all seven families, for reference.

Feature set	#cols	Pearson	K-fold RMSE
Target-word lexical properties			
Proprietary dictionary lemma-level features	82	0.660	1.330
AoA, prevalence, lexical decision	6	0.543	1.487
Corpus frequency / Zipf scale	3	0.520	1.511
SCOPE psycholinguistic metabase	18	0.505	1.530
MorphoLex morphological decomposition	8	0.400	1.623
Sense-level disambiguation			
Proprietary dictionary sense-level features	212	0.704	1.257
Posterior distribution over senses	7	0.547	1.481
Sense assignment uncertainty	11	0.503	1.523
Entropy over possible senses	4	0.302	1.688
Morphological decomposition of headword	5	0.219	1.730
Cue / prompt difficulty			
Count of viable target alternatives	4	0.569	1.456
Candidate competition from spelling clue	15	0.550	1.473
Dictionary words matching spelling pattern	7	0.537	1.641
LLM guesses from context only	10	0.493	1.541
Properties of L1 translation hint	6	0.491	1.543
L1 → L2 transfer			
LLM back-translation quality	7	0.442	1.588
LLM-rated L1/English similarity	5	0.440	1.566
Log-frequency of L1 source word	3	0.395	1.629
Phonological distance to English	2	0.394	1.647
Expert-annotated cognate list	8	0.381	1.615
L2 proficiency positioning			
Brysbaert L2 yes/no word-knowledge	5	0.586	1.434
EFLLex CEFR-level frequency trajectories	6	0.574	1.448
CEFR-J / JACET Asian-learner levels	3	0.569	1.446
CEFR level (EFLLex, Words-CEFR)	6	0.566	1.458
L2 age-of-acquisition norms	3	0.561	1.464
Virtual-learner simulation			
LLM token surprisal on target	10	0.576	1.447
Local small models attempt the item	24	0.574	1.455
Weak-LLM guess from description hints	10	0.565	1.505
Masked-LM guess of target from context	35	0.452	1.578
Four OpenAI models of graded ability	15	0.424	1.596
Holistic difficulty models			
LLaMA-3.1-8B LoRA regression head	1	0.883	0.831
Qwen-2.5-7B LoRA regression head	1	0.881	0.840
XLM-R-XXL (10.7B) LoRA + R-Drop	1	0.843	0.966
Mistral-7B LoRA regression head	1	0.833	0.951
XLM-R-large TAPT + R-Drop, per L1	1	0.821	1.062
XLM-R-XL (3.5B) LoRA + R-Drop	1	0.820	1.034
RemBERT fine-tuned per L1	–	0.814	1.024
XLM-R-XL LoRA + R-Drop, per L1	1	0.806	1.072
Derived / second-order			
CatBoost ablation across 6 feature groups	6	0.835	0.984
Pairwise feature interactions	40	0.827	0.993
Ensemble of LLM rubric difficulty ratings	22	0.786	1.096

Table 7: Top feature sets per domain family ranked by single-set K-fold Pearson correlation against the GLMM difficulty target. Seed variants collapsed to the base row. Families describe *what the feature measures about difficulty*: surface properties of the target word, sense disambiguation, cue / prompt difficulty, L1 transfer, proficiency positioning, virtual-learner simulation, or holistic difficulty models. Holistic models (fine-tuned transformers + LLM rubrics) dominate outright; virtual-learner probes and proficiency-positioning norms carry the next-strongest signal.

Track	L1 / Run	Approach	#feat	K-fold RMSE	Test RMSE	Δ RMSE	K-fold r	Test r	Δr
Closed	ES / 1	Compact	29	0.9250	0.920	−0.005	0.869	0.872	+0.003
Closed	ES / 2	No XLM-R-XL	30	0.9227	0.910	−0.013	–	0.876	–
Closed	ES / 3	Feature-rich	50	0.9264	0.903	−0.023	0.868	0.877	+0.009
Closed	DE / 1	Compact	49	0.9158	0.887	−0.029	0.857	0.871	+0.014
Closed	DE / 2	No XLM-R-XL	36	0.9122	0.895	−0.017	–	0.868	–
Closed	DE / 3	Feature-rich	100	0.9162	0.885	−0.031	0.857	0.871	+0.014
Closed	CN / 1	Compact	30	0.8046	0.788	−0.017	0.876	0.885	+0.009
Closed	CN / 2	No XLM-R-XL	33	0.8050	0.776	−0.029	–	0.889	–
Closed	CN / 3	Feature-rich	150	0.8136	0.785	−0.029	0.873	0.886	+0.013
Open	ES / 1	Compact	58	0.7754	0.755	−0.020	0.910	0.916	+0.006
Open	ES / 2	No prop. dict.	66	0.7813	0.754	−0.027	–	0.916	–
Open	ES / 3	Trans. only	8	0.8217	0.837	+0.015	–	0.896	–
Open	DE / 1	Compact	49	0.7678	0.769	+0.001	0.902	0.904	+0.002
Open	DE / 2	No prop. dict.	86	0.7702	0.764	−0.006	–	0.905	–
Open	DE / 3	Trans. only	9	0.7961	0.790	−0.006	–	0.898	–
Open	CN / 1	Compact	48	0.6756	0.660	−0.016	0.915	0.920	+0.005
Open	CN / 2	No prop. dict.	47	0.6830	0.662	−0.021	–	0.920	–
Open	CN / 3	Trans. only	37	0.7029	0.679	−0.024	–	0.916	–

Table 8: K-fold (internal) vs test (official) metrics for all 18 submitted runs. Negative Δ RMSE means the test score was better than K-fold; positive Δ Pearson means the same.

Model	Params	Training
CatBoost	1k trees	lr 0.05, depth 6, l2 3.0
XLM-R-base	278M	TAPT+R-Drop, bs 16, 3 epochs
XLM-R-large	550M	TAPT+R-Drop, bs 8, 3 epochs
XLM-R-XL	3.5B	LoRA r16 a16, R-Drop
XLM-R-XXL	10.7B	LoRA r16 a16, R-Drop
RemBERT	575M	per-L1, bs 16, seeds 3
ELECTRA-base	110M	per-L1, bs 16, 3 epochs
mBERT	178M	per-L1, bs 16, 3 epochs
LLaMA-3.1-8B	7.5B	LoRA r16 a16, q/k/v_proj
Qwen-2.5-7B	7.1B	LoRA r16 a16, q/k/v_proj
Mistral-7B	7.1B	LoRA r16 a16 (excluded)

Table 9: Per-family hyperparameters. All decoder LLMs use a scalar regression head on the last-token hidden state, MLP width 2048 with GELU.

#	ES	DE	CN
1	LLaMA s789	LLaMA	Qwen s789
2	Qwen s456	LLaMA s456	LLaMA s123
3	LLaMA	Qwen s123	Qwen s456
4	LLaMA s123	LLaMA s123	Qwen
5	Qwen s789	Qwen s789	Qwen s123
6	Qwen s123	XLM-R-XXL	LLaMA s789
7	al-mean	LLaMA s789	LLaMA
8	Qwen	Qwen	LLaMA s456
9	LLaMA s456	al-mean	al-mean
10	Mistral	Mistral	Mistral
11	cue-sens	cue-sens	XLM-R-lg s456
12	XLM-R-XXL	XLM-R-XL	cue-sens
13	surp-rate	Qwen s456	cue-sens-freq
14	l2-rank	l2-rank	l2-rank
15	surp-int	al-mini	al-mini

Table 10: Top-15 RFE-ranked features per L1 (open track). LLaMA and Qwen entries denote the LoRA regression OOF predictions of the corresponding seeded model; *al-mean* is the ability-ladder mean success rate; *cue-sens* is the LLM-based cue-sensitivity score; *l2-rank* is the learner's L2 vocabulary-knowledge rank. Abbreviated to fit.

Ruling / source	# features
LLM predicts target word (Cat 2)	7
Difficulty in same API call (Cat 1)	2
Paid embedding API (Cat 3)	3
LLM estimates production prob. (Cat 5)	1
R4 contamination (internal audit)	2
Difficulty rating (internal audit)	1
External training data (internal audit)	1
Total	17

Table 11: Compliance-audit reclassifications.

Feature	Cost (USD)	Result
LLM token surprisal (3 weak models)	40.92	+
Guess from description (3 weak models)	44.34	+
LLM-rated L1/English similarity (GPT)	30.23	+
False-friend flags (GPT-4)	36.97	+
Anglicism detection (GPT-4)	34.30	+
Guess from clue (3 weak models)	63.48	+
LLM rubric (4o-mini)	1.93	**best ROI**
LLM rubric (reasoning)	39.15	marginal
LLM pseudo-item generation	32.80	negative
Misc. batch experiments	163.0	7/10 +
Total (approx.)	498	—

Table 12: Representative LLM API expenditure. The cheapest positive feature ($1.93) improved RMSE by 0.015, a better return per dollar than any other.