Analyzing Traditional and Neural Approaches to Multilingual Readability Assessment

Anonymous EMNLP submission

Abstract

Transformer-based models excel at Automatic Readability Assessment (ARA), yet feature-based models remain in active use because their predictions tie back to linguistic properties. This matters because readability labels are subjective and rater-dependent, so high accuracy on noisy ground truth may reflect surface patterns rather than the linguistic structure that defines difficulty. We test whether transformers internalize the same features as traditional models across Arabic, English, French, Hindi, and Russian using the README++ dataset. Shapley Additive Explanations (SHAP) identify the features driving traditional classifiers, which we then use as TCAV concept sets to probe multilingual XLM-R and language-specific encoders. Transformers recover surface-length, syntactic, and lexical-diversity signals, and reflect the ordinal CEFR structure of the traditional models. Alignment varies by model family, language, and layer, with language-specific encoders tracking traditional models more clearly than XLM-R. High linear separability does not always imply directional influence, limiting linear probing for count-based readability features.

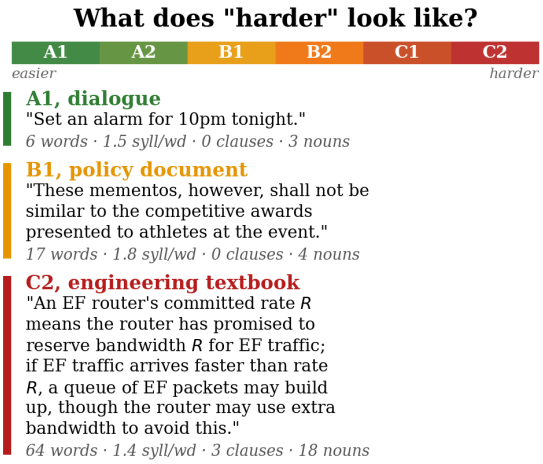

Figure 1: Sentences from README++ across CEFR levels with four representative features: word count, syllables per word (vocabulary), subordinate clauses (syntactic depth), and noun count (information density).

1 Introduction

Automatic Readability Assessment (ARA) predicts the difficulty of a text for a target reader. Difficulty is grounded in linguistic properties such as length, vocabulary, and syntactic structure, and these properties rise unevenly across reading levels as seen in Figure 1. Traditional feature-based classifiers remain in active use because their interpretability ties predictions back to these grounded linguistic features (Ribeiro-Flucht et al., 2024; Imperial and Ong, 2021). Transformer-based models achieve stronger performance but their internal representations are less transparent. Because readability labels are subjective and rater-dependent across small annotator pools (Naous et al., 2024; Vajjala, 2022), high accuracy on noisy ground truth may reflect surface patterns rather than the linguistic structure of difficulty, so accuracy alone is not a sufficient basis for trusting these models. Hybrid systems further suggest that explicit linguistic cues carry information transformers may not fully capture (Lee et al., 2021; Liu and Lee, 2023; Wilkens et al., 2024; Imperial, 2021). This raises a central question. Do transformer encoders learn the same readability-relevant linguistic signals that feature-based models rely on?

We investigate this question across Arabic, English, French, Hindi, and Russian using the README++ dataset (Naous et al., 2024). We first train traditional classifiers on handcrafted lexical, syntactic, and structural features, then apply Shapley Additive Explanations (SHAP) to identify which features drive their predictions (Lundberg and Lee, 2017). We use these SHAP-selected predictors as concept sets for Testing with Concept Activation Vectors (TCAV), probing both multilingual XLM-R and language-specific transformer

encoders (Kim et al., 2018). This pairing lets us compare *which features* drive traditional classifiers with *which directions* influence neural models.

Some transformer encoders recover classical readability signals, including surface length, syntactic complexity, and lexical diversity. Some also reflect the ordinal structure of CEFR levels observed in feature-based models. This alignment varies across languages, layers, and model families, with language-specific encoders generally producing clearer TCAV patterns than XLM-R. High linear separability does not always imply directional influence for count-based features, limiting linear probing as evidence of functional concept use.

Our contributions are threefold. First, we compare feature-based and transformer-based readability classification models across five languages. Second, we characterize how linguistic feature alignment varies by language, model family, and encoder layer. Third, we introduce a SHAP–TCAV framework for testing whether neural readability models encode the same interpretable signals used by traditional ARA systems.

2 Related Work

Feature-based, multilingual, and cross-lingual ARA. Early readability assessment relied on formula-based measures using surface features such as sentence length, word length, and unfamiliar word counts (Flesch, 1948; Kincaid et al., 1975; Dale and Chall, 1948; Mc Laughlin, 1969). Later machine learning approaches introduced richer handcrafted features, including lexical frequency, part-of-speech distributions, syntactic parse features, and discourse-level measures (Heilman et al., 2008; Vajjala and Meurers, 2013; Deutsch et al., 2020). These features remain useful because they connect predictions to interpretable linguistic properties. Uneven access to labeled data, readability standards, and NLP tools across languages limits multilingual ARA (Vajjala, 2022; Rao et al., 2021). Prior work addresses this through language-independent shallow features, language-specific models, and cross-lingual transfer using multilingual encoders (Shen et al., 2013; Budd et al., 2019; Gupta and Jones, 2023; Naous et al., 2024; Imperial et al., 2025). These studies evaluate predictive performance across languages but do not test whether neural models encode the same linguistic features used by traditional ARA systems.

Transformer-based and hybrid ARA. Transformer encoders have become central to ARA, often outperforming traditional feature-based models (Martinc et al., 2021; Imperial, 2021; Lee et al., 2021). Hybrid systems that combine transformer representations or predictions with handcrafted linguistic features often achieve strong results, especially in low-data or cross-domain settings (Lee et al., 2021; Liu and Lee, 2023; Wilkens et al., 2024). This creates a tension. If explicit features improve transformer models, they may contain information not fully captured by neural representations. If they do not, transformers may already encode those signals internally (Deutsch et al., 2020). Existing work addresses this question through downstream performance. We instead examine the representations directly, asking whether fine-tuned transformer readability models encode the same linguistic concepts that drive feature-based classifiers.

Feature attribution and concept-based probing. Interpretability methods provide tools for comparing explicit features and neural representations. SHAP assigns importance scores to input features and has been used to identify which linguistic predictors drive readability classifiers (Lundberg and Lee, 2017; Ribeiro-Flucht et al., 2024; Imperial and Ong, 2021). Probing methods, by contrast, test whether linguistic properties can be recovered from neural representations (Conneau et al., 2018; Tenney et al., 2019). High probing accuracy does not necessarily show that a model uses a concept for prediction. A feature may be linearly separable without being functionally influential (Maudslay and Cotterell, 2021; Bolukbasi et al., 2021). TCAV addresses this limitation by representing a human-defined concept as a direction in activation space and measuring whether movement along that direction affects model predictions (Kim et al., 2018). Prior work applies TCAV to NLP classification tasks such as toxicity and abusive language detection (Nejadgholi et al., 2022b,a, 2023), but not to systematically compare feature-based and transformer-based ARA across languages. Our work fills this gap by pairing SHAP-derived readability features with TCAV-based concept probing.

3 Methodology

We study whether transformer-based readability models encode the same linguistic signals used by traditional feature-based classifiers. Our pipeline

Table 1: Dataset sizes by language and split.

Language	Train	Val	Test	Total
English	1,692	283	847	2,822
Arabic	1,166	195	584	1,945
French	1,001	167	501	1,669
Hindi	894	149	448	1,491
Russian	1,054	176	528	1,758
Total	5,807	970	2,908	9,685

has two stages. First, we train traditional classifiers on handcrafted linguistic features and use SHAP to identify the features most important for readability prediction. Second, we use those SHAP-selected features as TCAV concepts to probe multilingual and language-specific transformer encoders.

Data. We use README++, a multilingual sentence-level readability corpus covering Arabic, English, French, Hindi, and Russian (Naous et al., 2024). README++ annotates sentences on the six-level CEFR scale (A1–C2), which we treat as a six-class classification task with classes indexed $k \in \{0, \ldots, 5\}$ corresponding to A1 through C2. We use the publicly available portion of the dataset, containing 9,685 sentences, and create stratified train, validation, and test splits with a 60/10/30 ratio for each language. Table 1 reports the resulting split sizes. Class distributions are imbalanced and vary by language. English skews to A2–B2, Arabic to B1–B2, Russian to A1, French to B1, and Hindi is the most balanced. We use balanced class weights for all traditional classifiers.

For TCAV concept construction, we use the README++ training split for all languages. For English and Arabic, we enlarge the concept pools with CEFR-SP (Arase et al., 2022) and DARES (El-Haj et al., 2024), respectively. The English pool combines the README++ training split with the deduplicated Wiki-Auto and SCoRE subset of CEFR-SP, yielding roughly 12,000 sentences. The Arabic pool combines the README++ training split with DARES, yielding roughly 15,000 sentences. We use these supplementary corpora only to construct concept and random sets, not for classifier training, validation, or testing.

Linguistic features and feature selection. We extract handcrafted linguistic features using LFTK (Lee and Lee, 2023), with Stanza-based preprocessing through the `spacy-stanza` interface. LFTK extracts interpretable text properties such as surface length and density features, lexical diversity measures, part-of-speech composition, POS-based lexical variation, and entity-density features. For English, LFTK also provides additional English-specific features, including richer entity-density and lexical familiarity features based on external lexical resources. We exclude precomputed readability formula features so that the feature-based models do not directly encode the target construct. This yields 141 general LFTK features for non-English languages and 211 English features after excluding readability formulas.

Because many LFTK features are highly correlated, we apply correlation-based hierarchical clustering separately for each language on the training split. We first remove zero-variance features, compute pairwise Spearman correlations, define feature distance as $d(i, j) = 1 - |\rho_{ij}|$, and apply agglomerative clustering with complete linkage. We cut the dendrogram at $\tau = 0.30$, corresponding to $|\rho| > 0.70$ within a cluster, and keep one representative per cluster. The representative is the feature with the highest absolute Spearman correlation with the CEFR label. This reduces English from 211 to 59 features, and the non-English languages from 141 features to 32–35 selected features.

Traditional classifiers. We train three traditional classifiers per language on the selected linguistic features. The classifiers are Logistic Regression, Linear SVM, and Random Forest, following prior explainable ARA work (Ribeiro-Flucht et al., 2024). We standardize features for Logistic Regression and Linear SVM but not for Random Forest, since tree-based models do not require scaling. All classifiers use balanced class weights. We tune hyperparameters with grid search and 10-fold stratified cross-validation, using Quadratic Weighted Kappa (QWK) as the scoring metric. QWK suits CEFR prediction because it penalizes larger level errors more strongly than adjacent errors. We then refit the best model for each language on the full training split before computing SHAP values.

Transformer models. We compare multilingual XLM-R base (Conneau et al., 2020) with language-specific encoders. The language-specific models are BERT for English (Devlin et al., 2019), AraBERTv02 for Arabic (Antoun et al., 2020), CamemBERT for French (Martin et al., 2020), MuRIL for Hindi (Khanuja et al., 2021), and RuBERT for Russian (Kuratov and Arkhipov, 2019). We fine-tune each model separately per language

for six-way CEFR classification, following Naous et al. (2024). We train for 20 epochs with cross-entropy loss and Adam, select the learning rate from $\{10^{-5}, 10^{-6}, 10^{-7}\}$ by validation QWK, and use the best validation-QWK checkpoint. The search selects 10^{-5} for all languages. Each model uses a fresh six-way classification head on the `[CLS]` token representation and has 12 encoder layers, which we probe with TCAV.

SHAP-based feature ranking. We use SHAP to identify which features drive traditional classifiers (Lundberg and Lee, 2017). We use model-specific explainers that compute Shapley values exactly for each model class. For Logistic Regression and Linear SVM, we use `LinearExplainer` with correlation-dependent perturbation, using the training split as the background distribution. This estimates missing features conditionally on observed features rather than assuming feature independence, which accounts for feature correlations that clustering mitigated. For Random Forest, we use `TreeExplainer`, which computes SHAP values directly from the tree structure.

For each feature, we compute global importance as the mean absolute SHAP value across test examples and CEFR classes:

$$\mathrm{Imp}_m(i) = \frac{1}{nK} \sum_{j=1}^{n} \sum_{k=1}^{K} |S^m_{j,i,k}|,$$

where $S^m_{j,i,k}$ is the SHAP value of feature i for example j and class k under model m. The explainers produce SHAP values on different scales. Therfore we rank features within each classifier and aggregate the ranks using Borda count:

$$B(i) = \frac{1}{3} \sum_{m} r_m(i),$$

where $r_m(i)$ is the rank of feature i under model m. For each language, we select the 10 features with the lowest Borda scores and the 10 features with the highest Borda scores as TCAV concepts.

TCAV concept construction and probing. We use TCAV to test whether SHAP-selected readability features correspond to influential directions in transformer representations (Kim et al., 2018). For each selected feature, the positive concept set consists of 100 sentences with the highest values for that feature in the concept pool. Each concept therefore represents the property of having a high value of that feature: the concept for `t_char` is sentences with many characters, and the concept for `t_syll2` is sentences with many words of more than two syllables. TCAV then tests whether moving along this high-feature direction in activation space affects the model's CEFR predictions. Each random set contains 100 sentences sampled from the same pool. We draw 1,000 random sets per concept for English and Arabic, where the concept pools are enlarged with CEFR-SP and DARES, and 500 random sets per concept for French, Hindi, and Russian, where concept pools are limited to the README++ training split.

For each transformer layer l, we extract the `[CLS]` representation $h_l(x)$ and train a logistic regression classifier to distinguish concept activation from random activations. The normalized classifier weight vector defines the Concept Activation Vector (CAV):

$$v^{(l)}_C = \frac{w^{(l)}_C}{\|w^{(l)}_C\|_2}.$$

We record CAV classifier accuracy as a measure of linear separability. We then measure whether moving in the concept direction increases the model's predicted probability for CEFR class k:

$$D_{C,k,l}(x) = \nabla_{h_l} p(y = k \mid x) \cdot v^{(l)}_C.$$

A positive value means moving the hidden representation along the concept direction raises the predicted probability for class k. The TCAV score is the fraction of class-k test examples with a positive directional derivative:

$$\mathrm{TCAV}_{C,k,l} = \frac{1}{|X_k|} \sum_{x \in X_k} \mathbf{1}\left[D_{C,k,l}(x) > 0\right].$$

CAV accuracy measures whether a concept is linearly recoverable from a layer. The TCAV score measures whether that concept direction positively influences the model's prediction for a CEFR class.

For probing, we extract activations and gradients at a maximum sequence length of 512 tokens, the architectural input limit of every target encoder. This recovers full context for the long-sentence tail. We compute gradients on the held-out test set, partitioned by true CEFR level, so each per-class pool X_k contains only test examples with the corresponding rating.

TCAV significance testing. Each concept is evaluated across multiple random sets, producing a distribution of TCAV scores for each concept–class–layer tuple (C, k, l). We test whether the mean TCAV score differs from the random baseline of 0.5 using a two-sided one-sample t-test:

$$H_0 : \mu_{C,k,l} = 0.5.$$

We assess significance at $\alpha = 0.01$ with Bonferroni correction. For each plotted analysis, we divide α by the maximum number of concept–class–layer comparisons shown in the corresponding graph and apply the adjusted threshold uniformly across plots. A TCAV result is significant only if its p-value falls below this Bonferroni-adjusted threshold.

4 Results

4.1 Model Performance

Table 2 reports test performance. Both transformer families outperform the traditional classifiers on QWK and accuracy across every language, with the language-specific encoder doing so on macro F1 as well in English, Arabic, and Russian. The improvement is largest on ordinal metrics. QWK ranges from 0.65–0.78 for the best traditional classifiers up to 0.83–0.86 for the best transformers. Accuracy improves from around 0.44–0.51 up to 0.53–0.64.

The two transformer families perform at broadly similar levels on QWK, but diverge more on accuracy and macro F1. XLM-R and BERT are nearly tied in English, and XLM-R and RuBERT are nearly tied in Russian, with RuBERT slightly higher on accuracy and XLM-R higher on macro F1. AraBERTv02 shows the clearest language-specific advantage, gaining 6 accuracy points and 6 macro F1 points over XLM-R in Arabic. The pattern reverses in Hindi and French, where XLM-R outperforms MuRIL and CamemBERT on accuracy and macro F1 despite similar QWK. The macro F1 gap is largest in French (0.53 versus 0.36) and Hindi (0.54 versus 0.41), likely from the balanced Hindi class distribution and CamemBERT's known sensitivity on smaller fine-tuning sets, where underperformance on any single class hurts the macro average more than in skewed languages.

4.2 What Feature-Based Models Use

Table 3 gives the full name and a short description for every feature that appears in any language's top 10. Surface-length features fill the top three positions in every language. Total characters (`t_char`) appears in the top three for every language, and average characters per word (`a_char_pw`) appears in the top four for every language (top three in English, Arabic, Hindi, and Russian, and rank four in French). Type-token ratio (`simp_ttr`) appears in the top 7 for all five, and the length-normalized Uber Index (`uber_ttr`) enters the top 10 for Arabic, Hindi, and Russian. POS counts fill the remaining slots. Adjectives (`n_adj`) are high-ranked in Arabic, French, Hindi, and Russian, and adpositions (`n_adp`) are high-ranked in English, French, Hindi, and Russian. English diverges mainly through two features only available in its enlarged pool. `t_syll2` counts multi-syllabic words, and `a_kup_pw` measures Kuperman age-of-acquisition per word.

Per-class SHAP beeswarms show that features invert polarity cleanly between the extreme CEFR classes (Figure 2). For A1 (Class 0), high values of `t_char`, `a_char_pw`, and related count features produce negative SHAP contributions. For C2 (Class 5), the same high values produce positive contributions. Intermediate classes transition smoothly between these poles. Classes 1 and 2 share the A1 polarity with decreasing magnitude. Classes 3 and 4 share the C2 polarity with increasing magnitude. Class 2 shows the most compressed range, marking the transition between the two regimes. This ordinal coherence is notable because the classifiers

Table 2: Test QWK, Accuracy, and Macro F1 per model and language. RF = Random Forest. LS = language-specific encoder. Best in bold.

Model	EN	AR	FR	HI	RU
		QWK			
LR	0.71	0.60	0.74	0.76	0.75
SVM	0.72	0.65	0.75	0.77	0.76
RF	0.75	0.65	0.78	0.76	0.76
XLM-R	0.82	0.82	**0.83**	**0.83**	**0.86**
LS	**0.83**	**0.85**	0.81	0.82	**0.86**
		Accuracy			
LR	0.40	0.43	0.40	0.44	0.42
SVM	0.45	0.44	0.40	0.44	0.42
RF	0.51	0.50	0.46	0.45	0.44
XLM-R	0.59	0.58	**0.53**	**0.54**	0.55
LS	**0.60**	**0.64**	0.51	0.46	**0.56**
		Macro F1			
LR	0.37	0.43	0.38	0.45	0.40
SVM	0.41	0.45	0.39	0.45	0.41
RF	0.46	0.49	0.44	0.45	0.38
XLM-R	0.50	0.59	**0.53**	**0.54**	**0.51**
LS	**0.54**	**0.65**	0.36	0.41	0.47

Table 3: Descriptions of all features appearing in any language's top 10. The Abbr. column gives the short label used in Table 4. Group abbreviations: SLD = surface length and density, POS = POS composition, LV = POS-based lexical variation, LD = lexical diversity, LF = lexical familiarity.

Feature	Abbr.	Group	Description
`t_char`	char	SLD	Total characters in the sentence
`t_syll2`	syll2	SLD	Words with more than two syllables (English only)
`t_stopword`	stop	SLD	Total stop words
`a_char_ps`	char/s	SLD	Avg characters per sentence
`a_char_pw`	char/w	SLD	Avg characters per word
`n_noun`	noun	POS	Count of nouns
`n_adj`	adj	POS	Count of adjectives
`n_adp`	adp	POS	Count of adpositions
`n_det`	det	POS	Count of determiners
`n_cconj`	cconj	POS	Count of coord. conjunctions
`n_ucconj`	uconj	POS	Unique coord. conjunctions
`n_udet`	udet	POS	Unique determiners
`n_uadv`	uadv	POS	Unique adverbs
`n_part`	part	POS	Count of particles
`a_verb_pw`	verb/w	POS	Avg verbs per word
`a_verb_ps`	verb/s	POS	Avg verbs per sentence
`a_noun_pw`	noun/w	POS	Avg nouns per word
`a_pron_pw`	pron/w	POS	Avg pronouns per word
`a_punct_ps`	punct/s	POS	Avg punctuation per sentence
`simp_adj_var`	adj-var	LV	Unique adj. lemmas / adj. tokens
`simp_cconj_var`	cconj-var	LV	Unique cconj lemmas / cconj tokens
`root_adp_var`	adp-var	LV	Unique adp lemmas / $\sqrt{\text{adp tokens}}$
`root_verb_var`	verb-var	LV	Unique verb lemmas / $\sqrt{\text{verb tokens}}$
`simp_ttr`	TTR	LD	Type-token ratio: unique / total
`uber_ttr`	Uber	LD	Uber Index: $\log^2(\text{total}) / \log(\text{total}/\text{unique})$
`a_kup_pw`	AoA/w	LF	Avg Kuperman age-of-acquisition per word (English only)

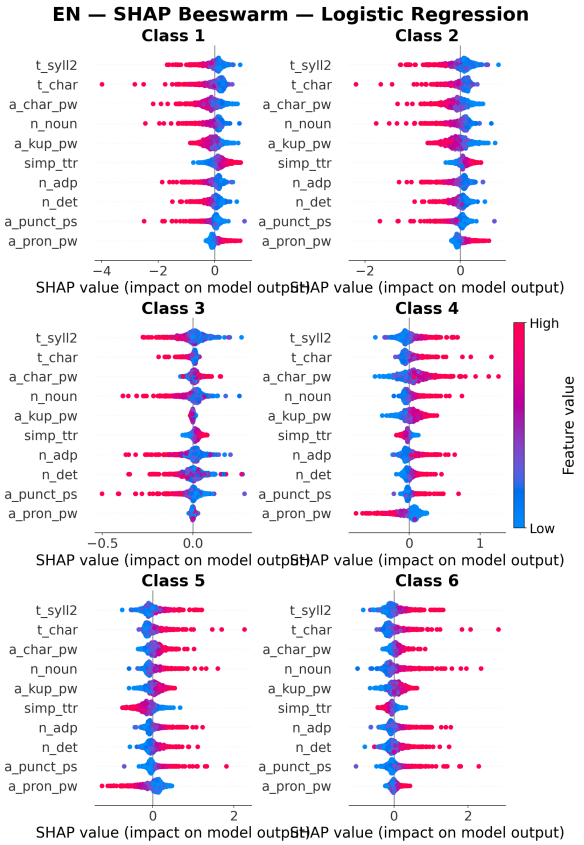

Figure 2: English Logistic Regression SHAP beeswarm, one panel per CEFR level (Class 0 = A1, Class 5 = C2). Color encodes feature value (blue = low, red = high), and horizontal position encodes SHAP contribution. Surface-length features show inverted polarity between A1 and C2, indicating an ordinally coherent representation despite no ordinal loss.

use flat cross-entropy with no ordinal constraint. The same pattern holds for the SVM and Random Forest classifiers and for the other four languages.

Table 4 shows the top 10 features by Borda rank per language, with the direction in which each feature pushes predictions. Feature directions match linguistic intuition. Longer sentences and longer words push toward harder CEFR levels, consistent with classical readability formulas. More adjectives, adpositions, and conjunctions push toward harder levels, indexing modification, prepositional phrases, and clause coordination. Raw `simp_ttr` decreases mechanically with sentence length and pushes toward easier text. Length-normalized `uber_ttr` pushes the other way in Arabic, Hindi, and Russian, capturing genuine vocabulary richness once length is controlled for. For English, `a_kup_pw` pushes upward, meaning sentences built from later-learned words rate as harder, and `a_pron_pw` trends downward, since pronoun-heavy text is typically conversational and easier.

4.3 CAV Separability

Across all five languages, CAVs trained on the top 10 SHAP features reach higher classifier accuracy than CAVs for the bottom 10. Figure 3 shows CAV accursies for Arabic and English. The gap is visible for both transformer families. In the language-specific encoders, top-feature CAV accuracy generally rises in early layers. English BERT and Arabic AraBERTv02 reach near-perfect separability by mid-network and plateau, and Russian RuBERT reaches its plateau by layer 4. French CamemBERT and Hindi MuRIL are non-monotonic, with MuRIL peaking in early layers and dipping through the middle and final layers. In XLM-R, accuracy is noisier across layers. English XLM-R shows a dip around layer 6 before recovering at later layers. Hindi XLM-R is flatter across layers, while Hindi MuRIL peaks higher in early layers but dips in mid and final layers. The features driving the feature-

Table 4: Top 10 features per language by Borda rank, with Logistic Regression SHAP direction. Abbreviations defined in Table 3.

Rank	EN	AR	FR	HI	RU
1	syll2↑	char↑	char/s↑	char/w↑	char↑
2	char↑	char/w↑	char↑	char↑	char/w↑
3	char/w↑	stop↑	TTR↓	noun↑	adj↑
4	noun↑	adj↑	char/w↑	adp↑	noun/w↑
5	AoA/w↑	uconj↑	adp↑	cconj↑	Uber↑
6	TTR↓	TTR↓	adj↑	adj↑	adp↑
7	adp↑	part↑	udet↑	TTR↓	TTR↓
8	det↑	verb/w↓*	verb/s↑	adj-var↑	stop↑
9	punct/s↑	Uber↑*	uadv↑	Uber↑	cconj↑
10	pron/w↓*	cconj-var↑*	punct/s↑	adp-var↑	verb-var↑

Note. ↑ means higher feature values push toward higher CEFR levels. ↓ means higher values push toward lower CEFR levels. * marks a general trend not consistent across all class-level panels.

based classifiers are the same features most linearly distinguishable in transformer activation spaces.

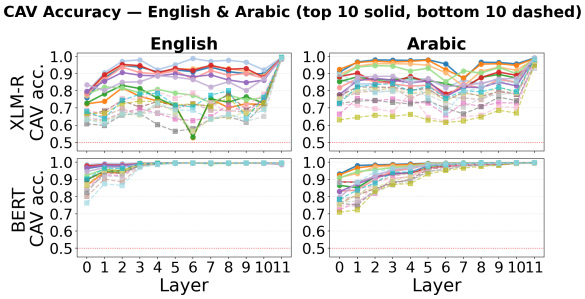

Figure 3: Mean CAV classifier accuracy across 12 layers for English and Arabic. Top row: XLM-R. Bottom row: language-specific encoder. Solid lines show top 10 SHAP features; dashed lines show bottom 10.

4.4 TCAV Directionality

For four of the five languages, TCAV fractions in at least one model family reproduce the SHAP-derived directions, with class C2 fractions for count-based features sitting above 0.5 and class A1 fractions sitting below. Rather than reporting individual layers, we look for sequential layer groups where the directional pattern holds for most top-10 features, since a signal that appears at one layer but disappears at the next is hard to interpret as a stable property of the model. The cleanest example is English BERT (Figure 4, lower half). Across layers 0–5, count-based features (`t_syll2`, `t_char`, `a_char_pw`, `n_noun`, `n_adp`, `n_det`, `a_kup_pw`, `a_punct_ps`) show high TCAV for C2 and low TCAV for A1. Moving the representation toward longer, denser sentences increases the model's C2 probability and decreases its A1 probability. Two features invert. `simp_ttr` and `a_pron_pw` show low TCAV for C2 and high TCAV for A1, matching their SHAP polarity.

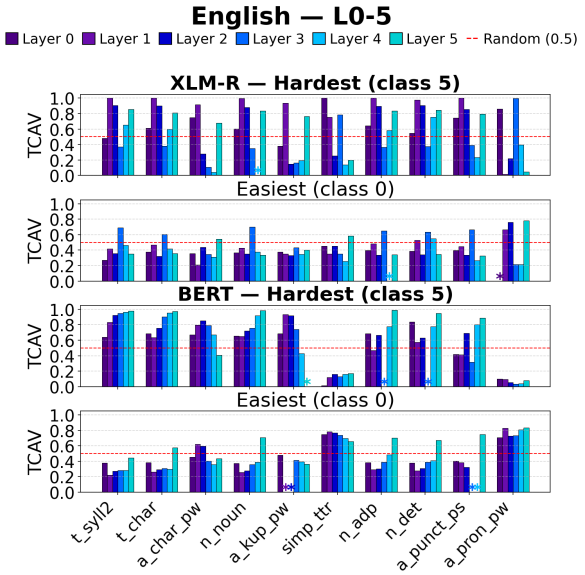

Figure 4: English TCAV fractions for the top 10 SHAP features, layers 0–5. For each model, the upper panel shows C2 (hardest) and the lower panel shows A1 (easiest). Asterisks mark non-significant results.

English XLM-R (Figure 4, upper half) shares almost none of this consistency. `a_char_pw` drops from ∼0.7 at layer 0 to ∼0.1 at layer 4 for C2. `simp_ttr` reaches 1.0 at layer 0 but falls to 0.2 by layer 5. The general C2-vs-A1 split remains visible on average, but the layer-wise instability lines up with the mid-layer dip in CAV accuracy from Section 4.3 (Figure 3), suggesting that XLM-R's intermediate representations encode these concepts in this way less reliably. The same broad pattern, BERT-family models showing cleaner directional structure than XLM-R within a coherent layer group, holds for Russian and French, though the layer groups differ. Russian RuBERT shows count features high for C2 and `simp_ttr` inverted across layers 4–6, while Russian XLM-R holds this pattern only at layers 4–5 and degrades at layer 6 for several features. French CamemBERT builds progressively across layers 6–11, with count features rising for C2 toward the later layers in this range and `simp_ttr` staying low for C2 throughout the group, while French XLM-R is unstable across the same range. The pattern does not extend uniformly across all 12 layers in either language, only within these mid-to-late groups.

Hindi reverses the model-family pattern (Figure 5). XLM-R produces the cleaner pattern across layers 3–5, with count features high for C2, `simp_ttr` inverted, and almost all A1 fractions below 0.5. `a_char_pw` sits at ∼0.4 at layer 3 and

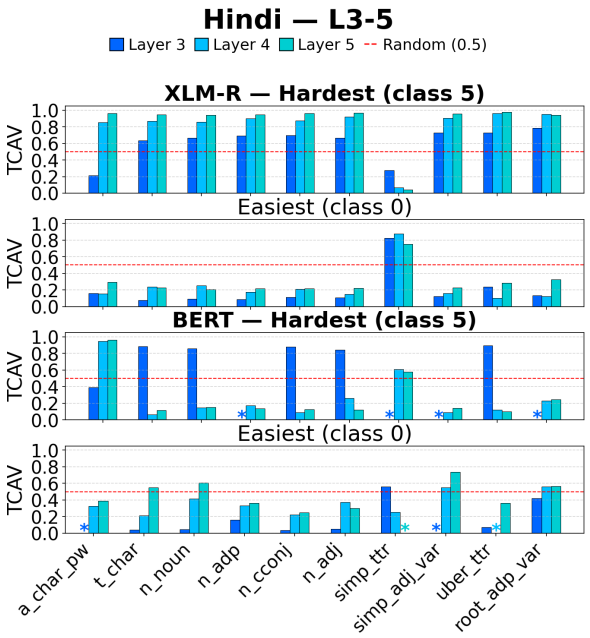

Figure 5: Hindi TCAV fractions for the top 10 SHAP features, layers 3–5. For each model, the upper panel shows C2 (hardest) and the lower panel shows A1 (easiest). Asterisks mark non-significant results.

jumps to over 0.9 at layer 4. This is the only language where XLM-R produces clearer directional structure than the language-specific encoder, which we attribute to the smaller Hindi dataset and balanced class distribution affecting CAV stability.

Arabic breaks the pattern entirely. CAV accuracies are high for both transformer families (top features reach 0.90–0.98 in Figure 3), so the concepts are separable in activation space. The TCAV fractions for C2 and A1, however, do not consistently separate above and below the 0.5 baseline within any layer group we examined, for either AraBERTv02 or XLM-R. The concepts are present but are not the axes along which Arabic predictions move. This does not mean the models fail to encode readability for Arabic. TCAV tests a specific hypothesis about linear directional influence, and the absence of a clean pattern may reflect a more distributed or non-linear encoding of count-based concepts that linear CAVs do not capture.

5 Discussion

The feature-based classifiers and fine-tuned transformers converge on the same signals and the same directions. Because we defined the TCAV concepts from the feature-based side and applied them to transformers that never saw feature values as inputs, the agreement in both separability and polarity indicates that transformers rediscover this structure rather than encoding something categorically different.

How clearly the agreement surfaces depends on language and model family. Language-specific encoders give stable layer-wise trajectories while XLM-R is noisier in the middle layers, with Hindi reversing the pattern. Arabic stands apart in a different way. Separability is high but directionality is absent across both transformer families, meaning the signals sit in activation space without being the axes along which predictions vary. For count-based concepts, this likely reflects a non-linear or distributed encoding that escapes linear probing.

One task-level property cuts across both analyses. The feature-based beeswarms grade smoothly along CEFR, and the transformer TCAV results show the same C2-vs-A1 inversion at the extremes, despite flat cross-entropy training in both cases. We designed neither analysis to produce ordinal structure, so the ordinal shape comes from the task itself. Explicit ordinal losses may therefore offer smaller gains than is sometimes assumed.

The pairing of SHAP and TCAV is the methodological point. The two methods provide independent evidence from explicit feature usage on one side and representation geometry on the other, and neither is causal on its own. Their agreement narrows the space of plausible explanations. Their disagreement, as in Arabic, locates the limit precisely rather than dismissing it as model failure. The framework extends to any domain where interpretable baselines and neural models coexist.

6 Conclusion

Fine-tuned transformer encoders pick up many of the linguistic signals that drive feature-based readability, and they do so along the same directions. This holds for English, Russian, Hindi, and French. Arabic is the informative exception that makes the framework's value clear. Where SHAP and TCAV disagree, the disagreement is itself a finding, pointing here to readability signals that exist in the representation but not as linear axes.

Limitations

The experiments use a single dataset (README++) at the sentence level on the CEFR scale, and we do not average transformer results across random seeds. The concept vocabulary is cross-lingually asymmetric because general-purpose LFTK fea-

tures are narrower for non-English languages, particularly for entity density and lexical familiarity. TCAV tests directional influence along pre-specified linear directions, cannot reveal concepts the investigator did not specify, and is sensitive to random-set choice and concept entanglement. The comparison covers one multilingual and one language-specific encoder per language and does not generalize to larger or differently structured models. Natural next steps are a richer concept vocabulary for morphologically rich languages, causal methods such as activation patching or concept steering, and extension to instruction-tuned LLMs and document-level ARA.

Ethics Statement

This work uses publicly available readability corpora and does not involve human subjects beyond the existing annotations in those corpora. All data sources are cited, and all fine-tuned models are derived from publicly released checkpoints under permissive licenses. The methods studied here are interpretability tools for classification models and do not introduce direct misuse risks. One indirect consideration is that readability models trained on limited-domain data may produce biased judgments when deployed in educational settings. The cross-lingual analysis highlights uneven feature availability across languages, which is itself an equity concern for multilingual ARA.

References

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based Model for Arabic Language Understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resources Association.

Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. CEFR-Based Sentence Difficulty Annotation and Assessment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6206–6219, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tolga Bolukbasi, Adam Pearce, Ann Yuan, Andy Coenen, Emily Reif, Fernanda Viégas, and Martin Wattenberg. 2021. An Interpretability Illusion for BERT. ArXiv:2104.07143 [cs].

Ray Budd, Tamas Marius, Paul Gatewood, and Doug Jones. 2019. Using K-Means in SVR-Based Text Difficulty Estimation. In *8th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2019)*, pages 84–88. ISCA.

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.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Edgar Dale and Jeanne S. Chall. 1948. A Formula for Predicting Readability: Instructions. *Educational Research Bulletin*, 27(2):37–54.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic Features for Readability Assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, Seattle, WA, USA → Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv:1810.04805 [cs].

Mo El-Haj, Sultan Almujaiwel, Damith Premasiri, Tharindu Ranasinghe, and Ruslan Mitkov. 2024. DARES: Dataset for Arabic Readability Estimation of School Materials. In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, pages 103–113, Torino, Italia. ELRA and ICCL.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Esther Gupta and Douglas Jones. 2023. Analyzing the Trade Space in Multi-lingual Automatic Text Difficulty Estimation. In *9th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 76–80. ISCA.

Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 2008. An Analysis of Statistical Models and Features for Reading Difficulty Prediction. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 71–79, Columbus, Ohio. Association for Computational Linguistics.

Joseph Marvin Imperial. 2021. BERT Embeddings for Automatic Readability Assessment. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 611–618, Held Online. INCOMA Ltd.

Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Muñoz Sánchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R. Jablonkai, Ekaterina Kochmar, Robert Joshua Reynolds, Eugénio Ribeiro, Horacio Saggion, Elena Volodina, Sowmya Vajjala, Thomas François, Fernando Alva-Manchego, and Harish Tayyar Madabushi. 2025. UniversalCEFR: Enabling Open Multilingual Research on Language Proficiency Assessment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9703–9755, Suzhou, China. Association for Computational Linguistics.

Joseph Marvin Imperial and Ethel Ong. 2021. Under the Microscope: Interpreting Readability Assessment Models for Filipino. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 1–10, Shanghai, China. Association for Computational Lingustics.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. MuRIL: Multilingual Representations for Indian Languages. ArXiv:2103.10730 [cs].

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). ArXiv:1711.11279 [stat].

J. P. Kincaid, Jr Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Number: RBR875.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. ArXiv:1905.07213 [cs].

Bruce W. Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on Text Readability Assessment: A Transformer Meets Handcrafted Linguistic Features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bruce W. Lee and Jason Lee. 2023. LFTK: Handcrafted Features in Computational Linguistics. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 1–19, Toronto, Canada. Association for Computational Linguistics.

Fengkai Liu and John Lee. 2023. Hybrid Models for Sentence Readability Assessment. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 448–454, Toronto, Canada. Association for Computational Linguistics.

Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and Unsupervised Neural Approaches to Text Readability. *Computational Linguistics*, 47(1):141–179.

Rowan Hall Maudslay and Ryan Cotterell. 2021. Do Syntactic Probes Probe Syntax? Experiments with Jabberwocky Probing. ArXiv:2106.02559 [cs].

G. Harry Mc Laughlin. 1969. SMOG Grading-a New Readability Formula. *Journal of Reading*, 12(8):639–646.

Tarek Naous, Michael J Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. 2024. ReadMe++: Benchmarking Multilingual Language Models for Multi-Domain Readability Assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12230–12266, Miami, Florida, USA. Association for Computational Linguistics.

Isar Nejadgholi, Esma Balkir, Kathleen Fraser, and Svetlana Kiritchenko. 2022a. Towards Procedural Fairness: Uncovering Biases in How a Toxic Language Classifier Uses Sentiment Information. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 225–237, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Isar Nejadgholi, Kathleen Fraser, and Svetlana Kiritchenko. 2022b. Improving Generalizability in Implicitly Abusive Language Detection with Concept Activation Vectors. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5517–5529, Dublin, Ireland. Association for Computational Linguistics.

Isar Nejadgholi, Svetlana Kiritchenko, Kathleen C. Fraser, and Esma Balkir. 2023. Concept-Based Explanations to Test for False Causal Relationships Learned by Abusive Language Classifiers. In *The*

7th Workshop on Online Abuse and Harms (WOAH), pages 138–149, Toronto, Canada. Association for Computational Linguistics.

Simin Rao, Hua Zheng, and Sujian Li. 2021. Cross-Lingual Leveled Reading Based on Language-Invariant Features. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2677–2682, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Luisa Ribeiro-Flucht, Xiaobin Chen, and Detmar Meurers. 2024. Explainable AI in Language Learning: Linking Empirical Evidence and Theoretical Concepts in Proficiency and Readability Modeling of Portuguese. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 199–209, Mexico City, Mexico. Association for Computational Linguistics.

Wade Shen, Jennifer Williams, Tamas Marius, and Elizabeth Salesky. 2013. A Language-Independent Approach to Automatic Text Difficulty Assessment for Second-Language Learners. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 30–38, Sofia, Bulgaria. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers the Classical NLP Pipeline. ArXiv:1905.05950 [cs].

Sowmya Vajjala. 2022. Trends, Limitations and Open Challenges in Automatic Readability Assessment Research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5366–5377, Marseille, France. European Language Resources Association.

Sowmya Vajjala and Detmar Meurers. 2013. On The Applicability of Readability Models to Web Texts. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 59–68, Sofia, Bulgaria. Association for Computational Linguistics.

Rodrigo Wilkens, Patrick Watrin, Rémi Cardon, Alice Pintard, Isabelle Gribomont, and Thomas François. 2024. Exploring hybrid approaches to readability: experiments on the complementarity between linguistic features and transformers. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2316–2331, St. Julian's, Malta. Association for Computational Linguistics.

A Additional SHAP Results

Figures 6–9 show SHAP beeswarms from the Logistic Regression classifier for the four languages not shown in the main text. Each panel corresponds to one CEFR class (Class 0 = A1, Class 5 = C2), with color encoding feature value (blue = low, red = high) and horizontal position encoding SHAP contribution. The ordinal polarity inversion between A1 and C2 reported in the main text holds across all four languages.

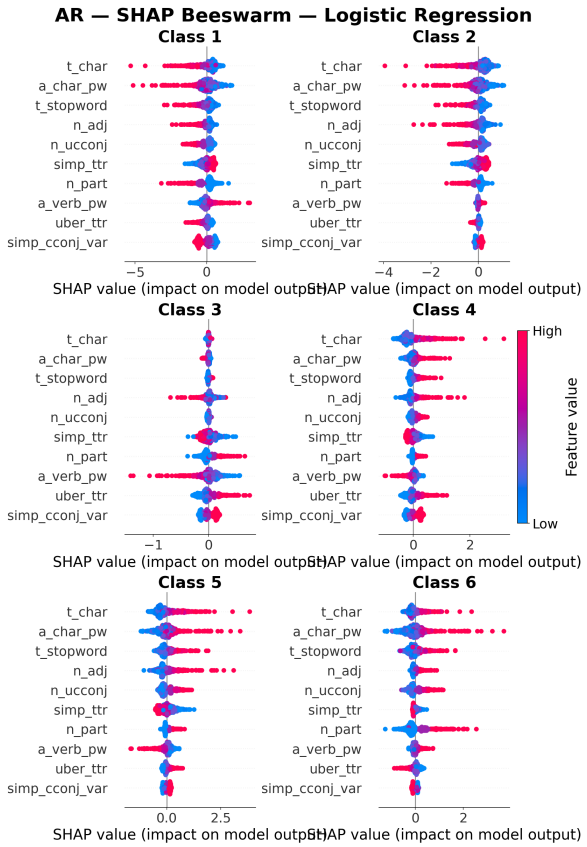

Figure 6: Arabic Logistic Regression SHAP beeswarm, one panel per CEFR level.

B Additional TCAV Results

Figures 10 and 11 show TCAV fractions for the top 10 SHAP features in French and Russian, across the mid-to-late layer groups where the directional pattern is most coherent. Each figure shows XLM-R in the top half and the language-specific encoder in the bottom half. Within each half, the upper panel shows C2 (hardest) and the lower panel shows A1 (easiest). Asterisks mark non-significant results.

C Additional CAV Accuracy Results

Figure 12 shows mean CAV classifier accuracy across 12 layers for French, Hindi, and Russian, complementing the Arabic and English plots in the main text. The same gap between top 10 and bottom 10 SHAP features holds across these languages.

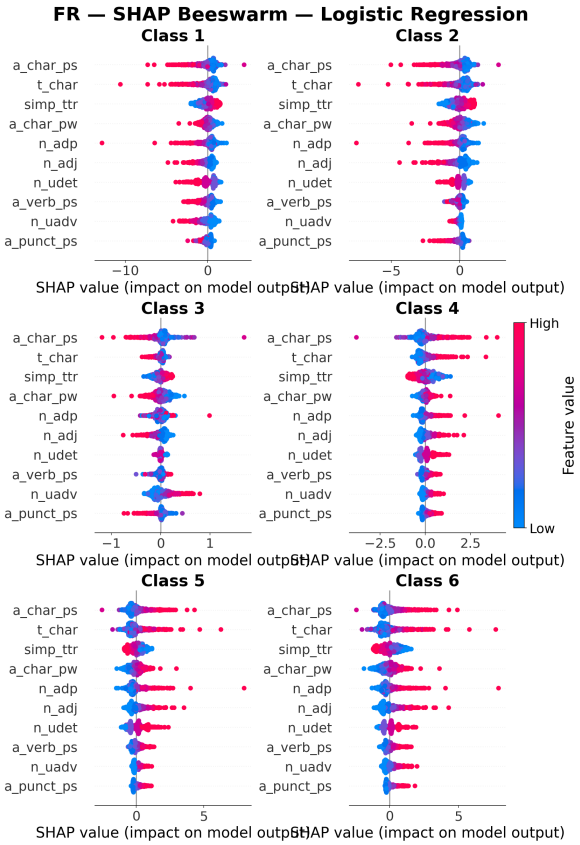

Figure 7: French Logistic Regression SHAP beeswarm, one panel per CEFR level.

D Use of AI Assistants

We used AI assistants during this work, primarily Claude (including Claude Code) and ChatGPT. For writing, AI assistants were used to edit and polish grammar and for LaTeX formatting. For coding, Claude Code was used to assist in writing and iterating on the Python code for our experiments. All experimental design, choice of methodology, concept formulation (SHAP–TCAV framework, feature selection, probing setup), interpretation of results, and the scientific content of the paper are the authors' own. The authors reviewed, tested, and verified AI-assisted code and text before inclusion, and take full responsibility for the correctness of the experiments and claims in this paper.

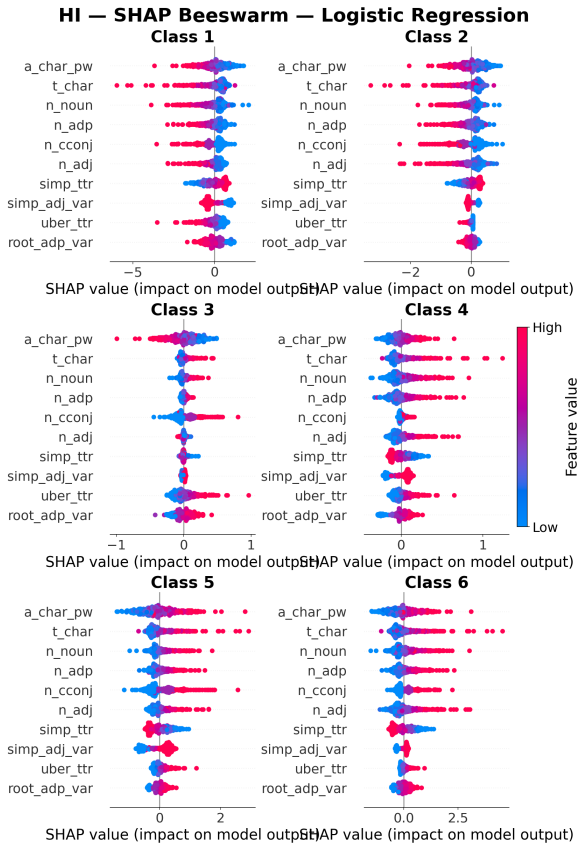

Figure 8: Hindi Logistic Regression SHAP beeswarm, one panel per CEFR level.

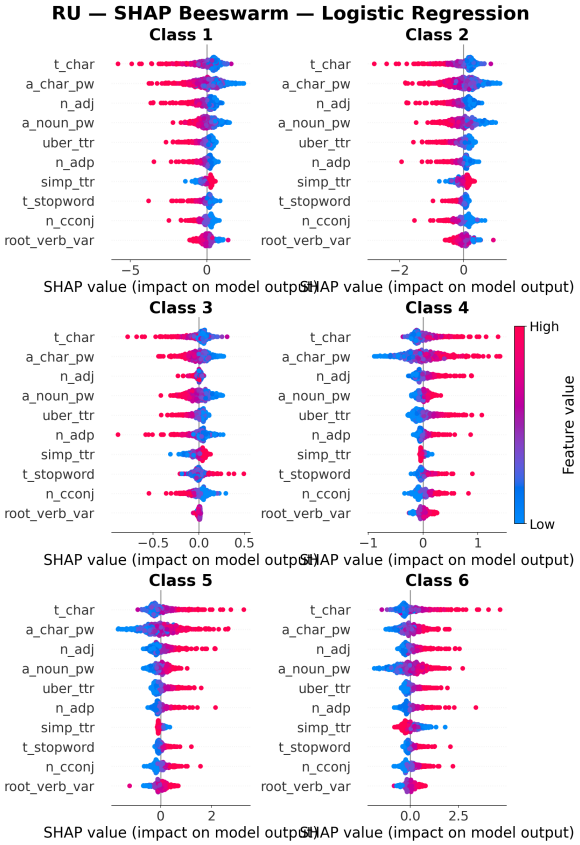

Figure 9: Russian Logistic Regression SHAP beeswarm, one panel per CEFR level.

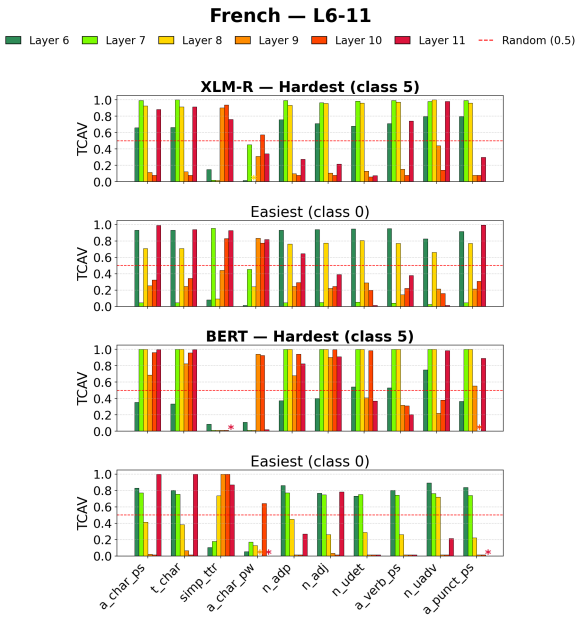

Figure 10: French TCAV fractions for the top 10 SHAP features, layers 6–11. Top half: XLM-R. Bottom half: CamemBERT. For each model, the upper panel shows C2 and the lower panel shows A1.

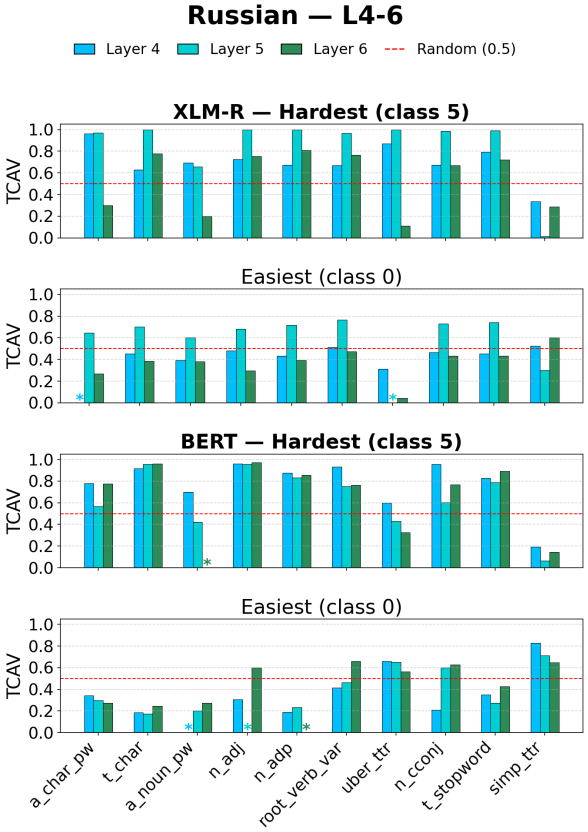

Figure 11: Russian TCAV fractions for the top 10 SHAP features, layers 4–6. Top half: XLM-R. Bottom half: RuBERT. For each model, the upper panel shows C2 and the lower panel shows A1.

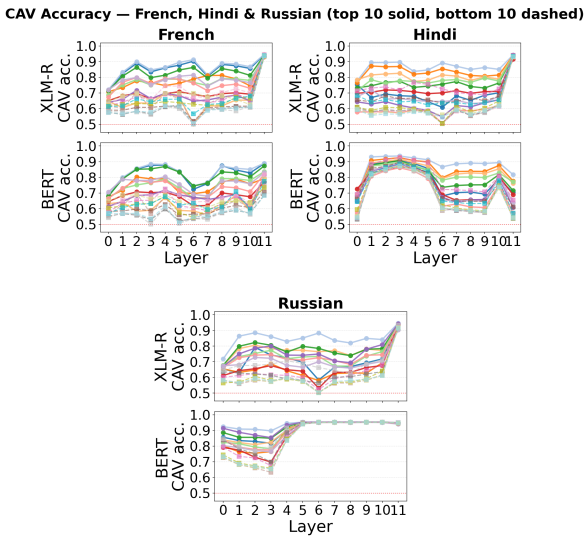

Figure 12: Mean CAV classifier accuracy across 12 layers for French, Hindi, and Russian. Top row: XLM-R. Bottom row: language-specific encoder. Solid lines show top 10 SHAP features; dashed lines show bottom 10.