Transformer-based readability classifiers are worse than you think: Evidence from cross-domain Arabic readability assessment

Sarah Alzubi
Brigham Young University
sarah.alzubi01@gmail.com

Robert Reynolds
Brigham Young University
robert_reynolds@byu.edu

Abstract

Arabic readability assessment is underexplored compared to English, and existing models are typically evaluated only within the training domain. We introduce the Jordanian School Textbook Corpus (JSTC), 82,512 segments from 240 textbooks spanning grades 1–12, and combine it with DARES to train XGBoost classifiers, fine-tuned CAMeLBERT transformers, and hybrid architectures evaluated both in-domain and on the BAREC out-of-domain benchmark. CAMeLBERT achieves strong in-domain performance (QWK = 0.830) but its cross-domain QWK collapses to 0.085, while XGBoost over 127 handcrafted linguistic features alone maintains the highest cross-domain QWK (0.240); adding [CLS] embeddings to those features actively harms transfer. Probing reveals that CAMeLBERT layers implicitly capture some linguistic features but higher-level signals overwhelm them, and Captum attribution identifies nouns and nominal particles such as al- as the most important tokens. The results argue for prioritizing linguistically-grounded features over contextual embeddings when cross-domain robustness is required.

1 Introduction

Readability assessment—the task of estimating the difficulty level of a text for a target audience—is fundamental to educational applications such as textbook selection, curriculum design, and adaptive learning systems (Vajjala, 2022).[1] While substantial progress has been made for English, Arabic readability research remains comparatively underresourced despite Arabic being a major world language spoken by over 400 million people (Cavalli-Sforza et al., 2018). Arabic presents unique challenges for readability assessment that cannot be addressed by simply transferring English-centric methods. Its morphological system is exceptionally rich: a single Arabic word can encode subject agreement, definiteness, case, mood, and multiple clitics, making morphosyntactic features far more informative for readability than in morphologically simpler languages. Orthographic variation, optional diacritics, and the diglossia between Modern Standard Arabic (MSA) and regional dialects further complicate both corpus construction and feature extraction.

Recent work in readability assessment has followed two broad paradigms. The first employs handcrafted linguistic features—lexical diversity measures, syntactic complexity metrics, and morphological statistics—fed to traditional classifiers (Deutsch et al., 2020; Al-Tamimi et al., 2014; El-Haj and Rayson, 2016). The second fine-tunes pretrained transformer models such as BERT on labeled readability data, using the [CLS] token representation as input to a classification head (Lee et al., 2021; Imperial, 2021; Ouassil et al., 2024). Hybrid approaches that combine both feature types have also been explored, with mixed results (Lee et al., 2021). For Arabic specifically, the transformer paradigm has gained traction with the release of pretrained Arabic language models such as CAMeLBERT (Inoue et al., 2021), enabling direct fine-tuning for downstream tasks.

A critical limitation of existing work, however, is that models are almost universally evaluated only in-domain—that is, on held-out data from the same distribution as the training set. This evaluation protocol can mask a fundamental problem: neural models may learn to exploit surface-level correlations and semantic peculiarities specific to the training corpus rather than capturing genuine text complexity. Practical educational deployment requires readability models that generalize across text types: a model trained on textbooks must also work

[1] This article is an abbreviated version of Alzubi (2026). Code and data are available in a private repo (due to copyright restrictions) at https://github.com/Sarah-git99/2026_BEA_Arabic_Readability.git. Contact the authors for access.

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 766–776

reliably for news articles, literature, and online content encountered by students. This cross-domain robustness has received almost no systematic attention for Arabic readability, and recent evidence from other languages suggests that neural models may fail dramatically under domain or format shift (Van Der Sluis and Van Den Broek, 2025; Lim et al., 2024).

The stakes of this question are not merely academic. As readability assessment tools are adopted in educational settings—to match learners with appropriate materials, to evaluate textbook difficulty across curricula, and to support adaptive learning platforms (Vajjala, 2022)—their cross-domain reliability becomes a practical concern. If these tools rely on models that perform well on benchmarks but fail in real-world deployment across diverse text types, the consequences could include mismatched materials, frustrated learners, and misallocated educational resources.

In this paper, we present the first systematic comparison of linguistic, neural, and hybrid readability models for Arabic under both in-domain and cross-domain conditions. Our contributions are:

1. **Jordanian School Textbook Corpus (JSTC)**: A new corpus of approximately 82,512 text segments from 240 Jordanian school textbooks (grades 1–12), with 127 extracted linguistic features across six categories.

2. **Cross-domain evaluation**: A rigorous comparison of seven model configurations on both 1) withheld in-domain textbook data and 2) the BAREC benchmark (Elmadani et al., 2025), revealing that fine-tuned transformers suffer catastrophic performance collapse under domain shift.

3. **Key finding**: XGBoost with linguistic features alone achieves the highest cross-domain Quadratic Weighted Kappa (QWK), substantially outperforming all transformer and hybrid configurations. Adding `[CLS]` embeddings to linguistic features actively degrades cross-domain performance.

2 Related Work

2.1 Arabic Readability Corpora and Methods

Early Arabic readability work adapted traditional formulas to Arabic's morphological properties. Al-Tamimi et al. (2014) proposed AARI, an automatic readability index for Arabic, and El-Haj and Rayson (2016) introduced OSMAN, a readability metric tailored to Arabic text. Forsyth (2014) applied statistical machine-learning methods for automatic readability detection for Modern Standard Arabic. Cavalli-Sforza et al. (2018) surveyed the state of Arabic readability research, identifying corpus scarcity as a major bottleneck.

More recently, several Arabic readability corpora have been developed. Nassiri et al. (2019) introduced MoSAR, a Modern Standard Arabic readability corpus for L1 learners. El-Haj et al. (2024) and Almujaiwel et al. (2025) released DARES, a dataset for Arabic readability estimation derived from Saudi school materials. Most recently, Elmadani et al. (2025) introduced BAREC, a large-scale benchmark spanning multiple domains and grade levels. Despite this progress, none of these efforts has systematically evaluated cross-domain generalization of their models.

2.2 Transformer and Hybrid Approaches

Lee et al. (2021) demonstrated that combining BERT soft labels with handcrafted linguistic features improves readability assessment for English, achieving state-of-the-art results on the CommonLit benchmark. Imperial (2021) showed that BERT embeddings alone can be effective for readability classification. For Arabic specifically, Ouassil et al. (2024) combined BERT with BiLSTM for readability assessment, and Khallaf and Sharoff (2021) applied both sentence embeddings and traditional linguistic features for readability classification.

2.3 Cross-Domain Robustness

Cross-domain evaluation of readability models has received limited but growing attention. Van Der Sluis and Van Den Broek (2025) showed that interpretable models exhibit better domain generalization in textual complexity modeling across multiple languages, providing theoretical motivation for feature-based approaches. Their analysis demonstrates that models with transparent, linguistically-grounded features avoid the domain-specific overfitting that characterizes opaque neural representations. Lim et al. (2024) investigated the robustness of hybrid models in cross-domain readability assessment, finding that linguistic features provide more stable transfer than purely neural representations. Deutsch et al. (2020) provided a thorough evaluation of linguistic features for English readability at the BEA workshop, establishing that traditional features remain competitive with neural

approaches, though they did not conduct cross-domain experiments.

To our knowledge, no prior study has systematically compared linguistic, neural, and hybrid readability models for Arabic under cross-domain conditions. Our work fills this gap, leveraging Arabic's rich morphological system to construct a particularly informative linguistic feature set and using the recently released BAREC benchmark to perform rigorous out-of-domain evaluation.

3 Data

3.1 Jordanian School Textbook Corpus (JSTC)

We construct the Jordanian School Textbook Corpus (JSTC) from 240 official textbooks in the Jordanian national curriculum, spanning Grades 1–12. These textbooks cover core subjects including Arabic Language, Islamic Education, and Digital Skills.

Text extraction from the textbook PDFs presented a significant challenge due to the complexity of Arabic script, including right-to-left writing, diacritical marks, complex ligatures, and mixed text-image layouts common in educational materials. Standard PDF text extraction tools produced unreliable output for Arabic content, so we employed the QARI-OCR model (Wasfy et al., 2025) for advanced character-level recognition of Arabic text from scanned pages. This OCR-based pipeline was essential for handling the typographic diversity of textbooks, which include tables, sidebars, and decorative elements alongside running text.

Preprocessing was performed using CAMeL Tools (Obeid et al., 2020). Text was dediacriticized to remove optional vowel markings that are inconsistently applied across textbooks, and letters were normalized (e.g., variant forms of *alif* and *ya'* were mapped to canonical forms) to reduce orthographic variation. Deduplication was applied to remove repeated passages appearing across multiple textbooks in the same curriculum (e.g., recurring introductory material); this process removed 5.42% of segments from JSTC. The resulting corpus comprises 82,512 text segments totaling over 2.1 million tokens. Segments preserve the natural textbook structure, including paragraphs, passages, bullet points, and standalone sentences, rather than being artificially re-segmented into uniform units.

Table 1 shows the distribution of segments and tokens across 12 grades, with lines between the four coarse-grained readability categories. The corpus exhibits substantial class imbalance, with high school grades contributing roughly five times as many segments as early elementary grades, reflecting the greater volume of advanced instructional material typical of readability corpora.

Coarse label	Grade	Segments	Tokens
L1	G1	2263	36063
	G2	3188	51103
	G3	3042	55950
L2	G4	6042	108681
	G5	5341	107208
	G6	4577	105726
L3	G7	7991	225743
	G8	7356	212545
	G9	10012	273309
L4	G10	16677	572634
	G11	15604	483499
	G12	11172	361304
Total		**82512**	**1764899**

Table 1: Distribution of segments and tokens in JSTC dataset.

3.2 DARES

To introduce curricular diversity into training, we supplement JSTC with DARES V1 (El-Haj et al., 2024), a readability corpus derived from the Saudi Arabian public school curriculum, grades 1–12. After deduplication (19.37% of DARES segments were removed), the combined training set comprises approximately 93,265 segments.

3.3 BAREC Cross-Domain Evaluation

For cross-domain evaluation, we use BAREC (Elmadani et al., 2025), a sentence-level multi-domain Arabic readability benchmark spanning three broad domains: Arts & Humanities, Social Sciences, and STEM. BAREC was used exclusively for cross-domain testing—none of its data was observed during training. This enables us to assess whether models capture generalizable readability signals or merely overfit to textbook-specific patterns. Both 12-class (individual grades) and 4-class (educational bands) label mappings are applied to BAREC for evaluation.

Evaluating on BAREC provides a particularly stringent test of cross-domain robustness because the BAREC texts differ from textbook prose along multiple dimensions simultaneously: topic, genre, register, text length, and intended audience.

A model that maintains predictive accuracy on BAREC demonstrates genuinely transferable readability representations.

4 Features and Models

We extract 127 handcrafted linguistic features from each text segment, organized into six categories designed to capture multiple dimensions of text complexity:

Lexical features include type-token ratio (TTR), moving-average TTR (MATTR; Covington and McFall, 2010), measure of textual lexical diversity (MTLD; McCarthy and Jarvis, 2010), average word length, and lexical density (ratio of content words to total words).

TF-IDF features comprise eight aggregated statistics inspired by Attia et al. (2023) computed from both surface forms and lemmatized forms, capturing rates of lexical distinctiveness within the corpus.

Syntactic features measure clause density, clause embedding depth, sentence length, and dependency parse structure using CamelParser 2.0 (Elshabrawy et al., 2023) with Universal Dependencies annotation.

Part-of-speech features capture the distribution of POS categories (nouns, verbs, adjectives, adverbs, pronouns, prepositions, and functional categories) as proportions within each segment.

Morphosyntactic features exploit Arabic's rich morphological system, including person, number, and gender agreement; definiteness marking (the *al-* prefix); verb mood and tense; and case marking—all extracted via CAMeL Tools (Obeid et al., 2020) using the D3TOK tokenization scheme.

Semantic features include Latent Dirichlet Allocation (LDA) topic distributions at multiple granularities ($K \in \{50, 100, 150, 200\}$) and derived measures of semantic richness, clarity, focus, and breadth, along with word frequency quartile analysis that captures lexical sophistication.

Together, these 127 features are designed to capture readability signals at multiple linguistic levels—from surface lexical statistics through syntactic structure to semantic content—while exploiting Arabic's rich morphological system in ways that generic transformer representations may not.

4.1 Models

We evaluate seven model configurations spanning statistical, neural, and hybrid approaches. Of the many statistical classifiers tested, we report results for XGBoost, which consistently outperformed alternatives such as logistic regression, SVM, and Random Forests in preliminary experiments.

XGBoost-Ling: XGBoost (Chen and Guestrin, 2016) trained on the 127 linguistic features.

XGBoost-CLS: XGBoost trained on the 768-dimensional `[CLS]` embedding extracted from a frozen CAMeLBERT model.

XGBoost-Ling+CLS: XGBoost trained on the concatenation of linguistic features and `[CLS]` embeddings.

CAMeLBERT-MSA: CAMeLBERT (Inoue et al., 2021) pretrained on Modern Standard Arabic, fine-tuned on the readability task.

CAMeLBERT-MIX: CAMeLBERT pretrained on a larger mixed-dialect corpus, fine-tuned on the readability task.

CAMeLBERT-MSA+H1: CAMeLBERT-MSA with the H1 subset of linguistic features (features which had very low correlation with probing model outputs) projected and fused via concatenation before the classification head.

CAMeLBERT-MIX+H3: Same architecture as CAMeLBERT-MSA+H1, but using CAMeLBERT-MIX and the H3 set of linguistic features (features which had very low, low, and moderate correlation with probing model outputs).

4.2 Experimental Setup

All models were trained on the combined JSTC and DARES data with a standard train/validation/test split. In-domain evaluation was performed on a held-out textbook test set drawn from the same corpus; cross-domain evaluation was performed on the full BAREC corpus, which was not observed during training. This strict separation ensures that cross-domain results reflect genuine generalization rather than data leakage.

We evaluate at two granularities: a **fine-grained** task with 12 classes (individual grade levels G1–G12) and a **coarse-grained** task with 4 classes (the educational bands in Table 1). The fine-grained task is particularly challenging because adjacent grade levels may differ only subtly in linguistic complexity, while the coarse-grained task groups these levels into broader educational stages.

For XGBoost models, we use the XGBoost

gradient-boosted tree classifier (Chen and Guestrin, 2016) with hyperparameters selected via cross-validation on the training set: n_estimators=600, learning_rate=0.03, max_depth=6. XGBoost is well-suited to this task because its tree-based architecture can capture non-linear interactions between linguistic features (e.g., the combination of high lexical density and long sentence length may indicate a different difficulty level than either feature alone).

For CAMeLBERT models, we fine-tune from the pretrained checkpoints with a classification head on top of the `[CLS]` representation. CAMeLBERT-MSA was pretrained on Modern Standard Arabic text, while CAMeLBERT-MIX was pretrained on a larger corpus that includes dialectal Arabic, potentially providing broader linguistic coverage at the cost of MSA specificity.

The hybrid models employ a projection fusion mechanism: the 127-dimensional linguistic feature vector is projected via a learned linear layer to match the 768-dimensional `[CLS]` representation. The projected features are then concatenated with the `[CLS]` output and passed to the classification head. This architecture allows the model to learn to weight both signal sources during training, but—as our results show—tends to prioritize the higher-dimensional `[CLS]` component. The models were trained using a maximum sequence length of 512, a learning rate of 5×10^{-5}, a batch size of 32, and 10 epochs. All experiments were conducted using 5-fold cross-validation.

Performance is measured using accuracy and Quadratic Weighted Kappa (QWK; Cohen, 1968). QWK is a variant of Cohen's Kappa that applies quadratic penalties to prediction errors proportional to their distance from the true label, making it particularly appropriate for ordinal classification where a prediction of grade 3 for a grade 4 text is less severe than a prediction of grade 1. QWK ranges from -1 (systematic disagreement) through 0 (chance agreement) to 1 (perfect agreement). We adopt QWK as our primary evaluation metric, supplemented by accuracy.

5 Results

5.1 Fine-Grained Results (12 Classes)

Table 2 presents the fine-grained results. In-domain, CAMeLBERT-MSA achieves the highest performance across all metrics (Acc = 0.640, QWK = 0.830), substantially outperforming XGBoost-Ling (Acc = 0.284, QWK = 0.460). The hybrid model CAMeLBERT-MSA+H1 provides only marginal improvements in accuracy, but is slightly worse than its base in QWK (0.821 vs. 0.830).

Model	Ling	CLS	Acc (In)	QWK (In)	Acc (Cross)	QWK (Cross)
XGBoost-Ling	✓		0.284	0.460	**0.130**	**0.240**
CAMeLBERT-MSA		✓	0.640	**0.830**	0.098	0.085
CAMeLBERT-MIX		✓	0.626	0.807	0.108	0.092
XGBoost-CLS		✓	0.380	0.643	0.091	0.161
CAMeLBERT-MSA+H1	✓	✓	**0.644**	0.821	0.101	0.089
CAMeLBERT-MIX+H3	✓	✓	0.630	0.810	0.110	0.077
XGBoost-Ling+CLS	✓	✓	0.390	0.670	0.080	0.170

Table 2: Fine-grained (12-class) readability results. Best in-domain scores in **bold** (top section); best cross-domain scores in **bold** (all rows). XGBoost with linguistic features alone achieves the highest cross-domain QWK.

The cross-domain results tell a strikingly different story. CAMeLBERT-MSA's QWK collapses from 0.830 to 0.085—a 90% relative decrease—and accuracy drops from 0.640 to 0.098, approaching the chance level of 0.083 for 12 classes. All models that rely on `[CLS]` representations show similarly catastrophic degradation. In contrast, **XGBoost-Ling achieves the highest cross-domain QWK at 0.240**, more than double that of any CLS-based model.

Notably, adding `[CLS]` embeddings to linguistic features (XGBoost-Ling+CLS) *reduces* cross-domain QWK from 0.240 to 0.170, indicating that contextual embeddings actively introduce domain-specific noise that harms transfer. The same pattern holds for accuracy: XGBoost-Ling (0.130) outperforms XGBoost-Ling+CLS (0.080) cross-domain, despite the latter's higher in-domain accuracy (0.390 vs. 0.284).

5.2 Coarse-Grained Results (4 Classes)

Table 3 shows the coarse-grained results, which reinforce the same pattern. In-domain, CAMeLBERT-MSA again dominates (Acc = 0.776, QWK = 0.802), with the hybrid H1 providing a slight improvement (QWK = 0.805).

Cross-domain, the pattern of transformer collapse persists. CAMeLBERT-MSA's QWK drops from 0.802 to 0.106, while XGBoost-Ling's QWK drops more modestly from 0.414 to 0.264. **XGBoost-Ling again achieves the highest cross-domain QWK (0.264)**, outperforming the best hybrid model by a factor of three (H1: 0.084).

Even more strikingly, XGBoost-Ling achieves the highest cross-domain accuracy (0.470), surpassing all transformer and hybrid models despite

Model	Ling	CLS	Acc (In)	QWK (In)	Acc (Cross)	QWK (Cross)
XGBoost-Ling	✓		0.550	0.414	**0.470**	**0.264**
CAMeLBERT-MSA		✓	0.776	0.802	0.373	0.106
CAMeLBERT-MIX		✓	0.765	0.785	0.352	0.051
XGBoost-CLS		✓	0.612	0.600	0.360	0.160
CAMeLBERT-MSA+H1	✓	✓	**0.780**	**0.805**	0.363	0.069
CAMeLBERT-MIX+H3	✓	✓	0.770	0.793	0.370	0.067
XGBoost-Ling+CLS	✓	✓	0.630	0.630	0.320	0.155

Table 3: Coarse-grained (4-class) readability results. Best in-domain scores in **bold** (top section); best cross-domain scores in **bold** (all rows). XGBoost with linguistic features alone achieves the highest cross-domain QWK.

its much lower in-domain accuracy. This reversal highlights the fundamental difference between in-domain pattern matching and cross-domain generalization.

The coarse-grained results also reveal an interesting asymmetry in the hybrid models. While H1 achieves the highest in-domain QWK (0.805), its cross-domain QWK (0.084) is actually *lower* than that of standalone CAMeLBERT-MSA (0.103). This suggests that the projection-fusion mechanism, while effective at combining signals for in-domain classification, may amplify overfitting to domain-specific patterns when cross-domain transfer is required. The same trend holds for H3 (cross-domain QWK 0.067 vs. CAMeLBERT-MIX's 0.060), though the difference is smaller.

5.3 Cross-Domain Subdomain Analysis

Performance on the BAREC subdomains varies substantially by topic area. For the coarse-grained task, XGBoost-Ling achieves the best cross-domain results on STEM texts (Acc = 0.679, QWK = 0.293) and Social Sciences (Acc = 0.617, QWK = 0.287), with somewhat lower performance on Arts & Humanities (Acc = 0.427, QWK = 0.239). CAMeLBERT's cross-domain performance remains near chance across all subdomains, confirming that the transformer's domain-specific overfitting is not limited to a single target domain.

For the fine-grained task, subdomain variation follows a similar pattern. XGBoost-Ling achieves cross-domain QWK of 0.260 on STEM, 0.220 on Social Sciences, and 0.210 on Arts & Humanities. CAMeLBERT-MSA shows uniformly poor cross-domain performance across all subdomains (QWK < 0.10).

This pattern suggests that linguistic complexity features transfer more reliably to STEM and Social Science texts, which share certain structural properties with textbook prose (e.g., expository structure, technical vocabulary, formal register). Literary and humanities texts present a greater distributional shift due to their use of narrative structures, figurative language, and stylistic variation that differ markedly from instructional prose.

5.4 Feature Type Comparison

The XGBoost variants isolate the effect of input representation under a constant classifier. In-domain, the QWK ordering is Ling $<$ CLS $<$ Ling+CLS for both the fine-grained task (0.460, 0.643, 0.670) and the coarse-grained task (0.414, 0.600, 0.630). Cross-domain, this ordering reverses: Ling $>$ Ling+CLS $\geq$ CLS (fine-grained: 0.240, 0.170, 0.160; coarse-grained: 0.264, 0.160, 0.160). The near-equality of CLS and Ling+CLS cross-domain suggests that when both feature types are present, the classifier relies primarily on CLS features—which dominate in-domain—effectively discarding the linguistic features that would have provided better transfer. **The features most useful for in-domain classification are precisely those that hurt cross-domain transfer**, an inherent tension that cannot be resolved by naive feature combination.

5.5 Feature Importance

We performed permutation importance analysis for the XGBoost model on the test set. This method evaluates feature importance by randomly shuffling one feature at a time and measuring the resulting drop in model performance. Table 4 presents the top 6 ranked feature importances based on the observed performance degradation. The two rankings diverge sharply: in-domain predictions rely most heavily on surface lexical and TF-IDF statistics (`lemma_category_tfidf_mean`, `form_category_tfidf_mean`, `avg_word_length`, `Digit_Density`), whereas cross-domain predictions depend predominantly on syntactic complexity (`avg_parse_tree_height`, `avg_sentence_length_morphemes`, `avg_sentence_length_words`) and basic vocabulary coverage—features that capture domain-independent structural properties of text complexity rather than corpus-specific lexical distributions, reinforcing the pattern observed in §5.4.

5.6 Probing Analysis

Figure 1 presents the distribution of the R^2 correlations of linguistic features across layers. Specifi-

Feature	In-domain	Cross-domain
lemma_category_tfidf_mean	0.0414	—
n_sentences	0.0258	—
per_1	0.0248	—
form_category_tfidf_mean	0.0239	—
avg_word_length	0.0200	—
Digit_Density	0.0189	—
avg_parse_tree_height	—	0.0078
avg_sentence_length_morphemes	—	0.0049
avg_sentence_length_words	—	0.0041
per_3	—	0.0035
Basic_Vocab_Ratio	—	0.0028
pos_pron_dem	—	0.0028

Table 4: Top 6 permutation feature importance for in-domain and cross-domain test sets.

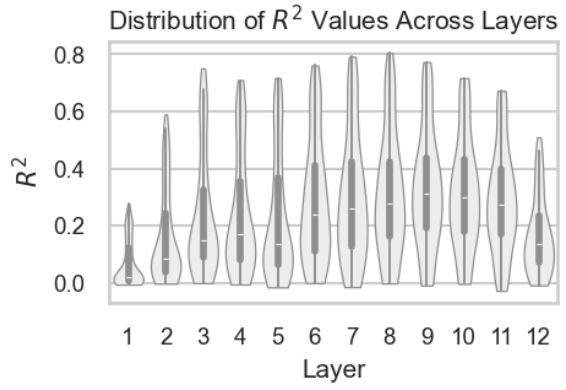

Figure 1: Distribution of probe R^2 values across CAMeLBERT layers for linguistic feature groups.

cally, it shows that the lower layers (1–3) encode a limited range of features and only partially capture them, mainly focusing on lexical cues and surface-level information. In contrast, the middle layers show stronger correlations across most linguistic features, indicating more effective encoding in these layers. This suggests that linguistic features are best represented in the middle layers. In the upper layers (10–12), the R^2 values stabilize or decrease slightly. This pattern indicates that, while linguistic information is still present, the representations become more specialized toward semantic abstraction and downstream tasks. See Alzubi (2026) for the full list of feature correlations with each layer.

5.7 Captum Attribution Analysis

We use the Captum attribution analysis to identify the textual elements CAMeLBERT-MSA relied on most for readability prediction. The most influential POS categories are nouns and adjectives, with morphological dominance driven by singular number and masculine gender, and clitic importance led by definiteness and coordination markers. The model assigns the highest importance to content words, especially nouns and proper nouns, which show positive excess attribution across grades, indicating a consistent signal for readability. Verbs, adjectives, and adverbs vary across grades, while function words generally show weaker or negative contributions, with grammatical words increasing at higher grades, suggesting stronger syntactic encoding. For clitics, the definite article *al-* has the strongest attribution across grades, while *wa-* becomes more prominent in higher grades, likely reflecting increased coordination and syntactic complexity.

6 Discussion

6.1 Why Transformers Fail Cross-Domain

The dramatic cross-domain collapse of fine-tuned CAMeLBERT models suggests that these models learn to exploit domain-specific topical and stylistic cues present in textbook data rather than genuinely capturing text complexity. The `[CLS]` representation, while powerful for in-domain classification, appears to encode distributional patterns specific to the training domain—patterns that become misleading when the text distribution shifts.

This interpretation is supported by two additional observations. First, XGBoost-CLS (which uses the same `[CLS]` embeddings but in a different classifier) also degrades substantially cross-domain, confirming that the problem lies in the representation rather than the classification architecture. Second, adding linguistic features to the transformer (hybrid models H1 and H3) fails to rescue cross-domain performance (QWK ≤ 0.089 for fine-grained), suggesting that the `[CLS]` signal dominates the combined representation and overwhelms the more transferable linguistic features.

We note, however, that BAREC differs from the training data not only in topic and register but also in format: BAREC items are single sentences, whereas JSTC and DARES segments are typically multi-sentence passages. Part of the observed collapse may therefore reflect this format shift rather than pure domain shift, particularly for `[CLS]` representations that may encode passage-level distributional cues unavailable in short sentence inputs.

6.2 Why Linguistic Features Transfer Better

Linguistic features such as average sentence length, lexical diversity (MATTR, MTLD), clause density, morphological complexity, and POS distributions measure domain-independent properties of text complexity. A syntactically complex sentence remains complex regardless of whether it discusses photosynthesis or Ottoman history. These features thus provide a noisier but more robust signal that maintains predictive value under domain shift.

The finding that XGBoost-Ling+CLS performs *worse* than XGBoost-Ling cross-domain is particularly telling. It demonstrates that `[CLS]` embeddings do not merely add complementary information—they actively introduce domain-specific noise that degrades the classifier's ability to exploit the transferable linguistic signal.

The probing results further confirm a layered encoding structure. Lower layers capture surface-level and lexical features, while middle layers are more strongly associated with syntactic information. In contrast, higher layers show reduced sensitivity to explicit linguistic features and instead emphasize task- and domain-specific contextual representations, aligning with the behavior of the classification head.

Several feature categories identified as important for cross-domain transfer are only weakly encoded in the `[CLS]` representation. Fine-grained POS distributions, rare morphological markers, and lexical frequency statistics show consistently low probe performance across layers. These features capture domain-independent aspects of text complexity, yet remain underrepresented in the learned embedding. This explains strong in-domain results, where topical signals are sufficient, and weaker cross-domain performance, where such signals do not transfer reliably. It also accounts for the limited benefit of hybrid models, since well-encoded features are largely redundant, while weakly encoded ones contribute too little to improve performance.

6.3 Comparison with Prior Work

Our in-domain CAMeLBERT results (QWK = 0.830 on 12 grades) are competitive with recent Arabic readability benchmarks. Elmadani et al. (2025) report QWK of 0.810 on BAREC's in-domain evaluation, and Almujaiwel et al. (2025) report Macro-F1 of 0.420 on concept-based readability estimation. Our linguistic XGBoost model's in-domain QWK (0.460) is lower, reflecting the well-documented advantage of contextual representations for in-domain classification. However, the cross-domain comparison—which no prior Arabic readability study has systematically conducted—reveals that this in-domain advantage is misleading.

Our findings are consistent with Van Der Sluis and Van Den Broek (2025), who demonstrated across multiple languages and tasks that interpretable models generalize better than neural models under domain shift. They extend the English-language findings of Deutsch et al. (2020), who showed that linguistic features remain competitive with neural approaches for readability—though that study did not include cross-domain evaluation. Critically, our results contrast with work that has found hybrid approaches superior to either feature type alone (Lee et al., 2021); we show that this finding does not hold under domain shift, where the hybrid's neural component becomes a liability rather than an asset.

6.4 Implications for Educational NLP

A readability tool deployed in real educational settings must handle diverse text types: textbooks, supplementary readings, news articles, and online content. Our results demonstrate that current transformer-based approaches, despite their impressive in-domain performance, are not ready for such deployment without explicit domain-adaptation strategies. Linguistic-feature models, while exhibiting lower in-domain performance, provide a substantially safer baseline for cross-domain use.

This finding has particular relevance for Arabic educational NLP, where the diversity of text types, genres, and registers encountered by learners is wide, and where the cost of misclassifying a text's difficulty level can directly affect curriculum alignment and learner outcomes. A model that correctly classifies textbook passages but assigns university-level STEM texts to the elementary band would be worse than useless in a mixed-resource educational setting.

More broadly, cross-domain evaluation should be a standard component of readability assessment research, particularly when the intended application involves diverse input texts. Although XGBoost-Ling achieved a modest 0.240 QWK, this result is qualitatively superior to the near-chance performance of the neural models we tested.

7 Conclusion

We introduced JSTC, a new Arabic readability corpus consisting of approximately 82,512 textbook segments with 127 linguistic features, which was further combined with the DARES dataset comprising 10,755 samples, and conducted the first systematic cross-domain evaluation of Arabic readability models. Our central finding is that fine-tuned transformers, despite achieving strong in-domain QWK (0.830), collapse to near-chance under domain/format shift (QWK 0.085), while XGBoost with linguistic features alone maintains the most robust cross-domain performance (QWK 0.240–0.264). Adding `[CLS]` embeddings to linguistic features actively degrades cross-domain transfer.

These results argue for prioritizing linguistically-grounded, interpretable feature sets over opaque contextual representations in educational readability applications, particularly when cross-domain robustness is required. The finding that contextual embeddings actively harm cross-domain transfer—even when combined with linguistic features—suggests that the two feature types encode fundamentally different types of information, and that naive combination strategies are insufficient.

Future work should explore domain adaptation techniques (e.g., adversarial training, domain-invariant representations) that could improve transformer robustness without sacrificing in-domain performance, as well as lighter-weight regularizers such as label smoothing and early stopping that may reduce overconfidence in domain-specific patterns. Evaluation on additional cross-domain benchmarks beyond BAREC—such as the Arabic portion of ReadMe++ (Naous et al., 2024), which provides sentence-level readability labels across multiple domains—would help confirm the generality of our findings and disentangle domain shift from format shift; rotating which of JSTC, DARES, and BAREC serves as the held-out corpus would further isolate corpus-specific effects. More balanced training data, particularly with expanded early-grade coverage, would also help mitigate the imbalance-induced biases observed here. Finally, investigating whether recent larger Arabic language models or instruction-tuned models exhibit improved cross-domain transfer would clarify whether our findings reflect a fundamental limitation of the fine-tuning paradigm or a limitation of the specific model scale studied here.

Limitations

BAREC was selected as our cross-domain test corpus because it represents the most extreme distributional shift available, but it differs from the JSTC and DARES training data along multiple dimensions simultaneously—topic, register, and format. In particular, BAREC items are single sentences while JSTC and DARES segments are typically multi-sentence passages, so part of the observed cross-domain collapse may reflect this format shift rather than pure domain shift.

The set of cross-domain regularization strategies explored in this study was limited, which may have affected the absolute performance reported for neural and hybrid models across datasets and text sources.

Both the in-domain and cross-domain datasets are imbalanced, particularly across subdomains and early grade levels, which may have influenced model training and evaluation.

Ethics Statement

The cross-domain performance degradation observed in this study highlights the potential risks of deploying models that appear accurate in controlled settings but fail on diverse real-world texts, which could lead to mismatched educational materials and negatively affect learners.

Acknowledgments

Generative AI tools were used to reduce the length of a much longer version of this article. Other than the insight in 5.4, all technical content, experimental design, results, and analyses are the authors' own work.

References

Abdel-Karim Al-Tamimi, Manar Jaradat, Nuha Aljarrah, and Sahar Ghanim. 2014. AARI: Automatic Arabic readability index. *Information Technology Journal*, 13:1350–1357.

Sultan Almujaiwel, Damith Premasiri, Tharindu Ranasinghe, Mo El-Haj, and Ruslan Mitkov. 2025. Complex concept-based readability estimation from Arabic curriculum. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24(11):1–21.

Sarah Alzubi. 2026. Robust and interpretable cross-domain arabic readability prediction using hybrid

modeling: Evidence on the limitations of transformer-based classifiers. Master's thesis, Brigham Young University, April.

Mohammed Attia, Younes Samih, and Yo Ehara. 2023. Statistical measures for readability assessment. In *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*, pages 153–161, Tokyo, Japan. Association for Computational Linguistics.

Violetta Cavalli-Sforza, Hind Saddiki, and Naoual Nassiri. 2018. Arabic readability research: Current state and future directions. *Procedia Computer Science*, 142:38–49.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

Michael A. Covington and Joe D. McFall. 2010. Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100.

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. Association for Computational Linguistics.

Mahmoud El-Haj and Paul Rayson. 2016. OSMAN – a novel Arabic readability metric. Proceedings of LREC 2016.

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 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. European Language Resources Association.

Khalid N. Elmadani, Nizar Habash, and Hanada Taha-Thomure. 2025. A large and balanced corpus for fine-grained Arabic readability assessment. *arXiv preprint*.

Ahmed Elshabrawy, Muhammed AbuOdeh, Go Inoue, and Nizar Habash. 2023. CamelParser2.0: A state-of-the-art dependency parser for Arabic. In *Proceedings of ArabicNLP 2023*, pages 170–180. Association for Computational Linguistics.

Jonathan Neil Forsyth. 2014. Automatic readability detection for Modern Standard Arabic. Master's thesis, Brigham Young University.

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. INCOMA Ltd.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104. Association for Computational Linguistics.

Nouran Khallaf and Serge Sharoff. 2021. Automatic difficulty classification of Arabic sentences. *arXiv preprint*.

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. Association for Computational Linguistics.

Ho Hung Lim, Tianyuan Cai, John S. Y. Lee, and Meichun Liu. 2024. Robustness of hybrid models in cross-domain readability assessment. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. European Language Resources Association.

Philip M. McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.

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*. Association for Computational Linguistics.

Naoual Nassiri, Violetta Cavalli-Sforza, and Abdelhak Lakhouaja. 2019. MoSAR: Modern Standard Arabic readability corpus for L1 learners. In *Proceedings of the 4th International Conference on Big Data and Internet of Things*, pages 1–7. ACM.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source Python toolkit for Arabic natural language processing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7022–7032. European Language Resources Association.

Mohamed Amine Ouassil, Mohammed Jebbari, Rabia Rachidi, Mouaad Errami, Bouchaib Cherradi, and Abdelhadi Raihani. 2024. Enhancing Arabic text readability assessment: A combined BERT and BiLSTM approach. In *2024 International Conference on Circuit, Systems and Communication (ICCSC)*, pages 1–7. IEEE.

Sowmya Vajjala. 2022. Trends, limitations and open challenges in automatic readability assessment research. arXiv preprint arXiv:2105.00973.

Frans Van Der Sluis and Egon L. Van Den Broek. 2025. Model interpretability enhances domain generalization in the case of textual complexity modeling. *Patterns*, 6(2):101177.

Ahmed Wasfy, Omer Nacar, Abdelakreem Elkhateb, Mahmoud Reda, Omar Elshehy, Adel Ammar, and Wadii Boulila. 2025. QARI-OCR: High-fidelity Arabic text recognition through multimodal large language model adaptation. *arXiv preprint arXiv:2506.02295*.