Factors Influence Cross-Prompt Scoring of Arabic Essays

Ahmed Ibrahim Suleiman
Department of Electrical and Computer Engineering
The Libyan Academy
Tripoli, Libya
ahmed.ibrahem@academy.edu.ly

Abduelbaset Mustafa Goweder
Department of Computer Science
The Libyan Academy
Tripoli, Libya
agoweder@academy.edu.ly

Abstract— Cross-prompt automated essay scoring (AES), which involves training on specific prompts and testing on unseen cases, represents a practical application scenario; however, this area remains underexplored for Arabic. Progress in AES is typically characterized by a succession of neural architectures that exhibit escalating levels of complexity. However, such progress is seldom evaluated against more basic choices such as tokenization, model-selection criteria, and text segmentation. We present the first controlled factorial study of cross-prompt Arabic AES on the LAILA dataset, crossing three architectures of increasing complexity (flat, hierarchical, multi-trait), three tokenizer–model configurations, and three random seeds, and benchmarking against a strong prompt-agnostic feature baseline. Our central finding is that tokenizer configuration drives performance far more than architecture: it outperforms each architectural step by approximately four times, covering a range of 0.16 to 0.19 in the overall Quadratic Weighted Kappa (QWK), the agreement metric used throughout. By contrast, the move from a flat to a hierarchical encoder yields only a small and seed-fragile gain (+0.045), multi-trait attention is mildly harmful (−0.02), and flat AraBERT matches the feature baseline (0.621 QWK) — which itself outperforms the hierarchical and multi-trait models in four of six configurations. We additionally provide evidence on how a seemingly straightforward "more-complex-is-better" result (M2 = 0.637 vs. M1 = 0.556 in our own early runs) is an artifact that dissolves once tokenization, selection criterion, and segmentation are each controlled. We release a corrected evaluation protocol — mean-per-prompt checkpoint selection and a sentence-based segmenter suited to Arabic — and argue that such controls are prerequisites for credible architecture claims in low-resource cross-prompt AES.

Keywords— Automated Essay Scoring (AES), Arabic NLP, Cross-Prompt Essay Scoring, Tokenization Effects, AraBERT, Hierarchical Models, Multi-Trait Learning, Domain Generalization, Controlled Factorial Study, Evaluation Protocol.

I. INTRODUCTION

Automated essay scoring (AES) assigns a quality score to a learner's written response. The practically important and harder variant is *cross-prompt* scoring, where no labeled essays exist for the target prompt and the model must generalize from essays written to other prompts [1], [2]. While AES has a substantial English literature, Arabic — a morphologically rich, lower-resource language — has received far less attention, and a recent survey reports no prior work applying hierarchical or multi-trait neural architectures, or studying tokenization effects, to Arabic cross-prompt AES.

A recurring narrative in AES research is one of ascending architectural complexity: flat encoders give way to hierarchical sentence–document models, which give way to multi-trait attention. Such comparisons, however, are frequently entangled with lower-level choices — which tokenizer is paired with the encoder, how the checkpoint is selected on the development set, and how essays are segmented into units — that are rarely held fixed across the architectures being compared. When these choices vary silently, an "architectural" improvement may in fact be a tokenization or evaluation artifact.

We make this concrete: in our own initial experiments, a hierarchical multi-trait model appeared to clearly beat a plain hierarchical one (0.637 vs. 0.556 QWK). Investigating the gap revealed that it was not architectural at all: the two systems had used different tokenizers, a development-set selection criterion that is provably unreliable under cross-prompt distribution shift, and a paragraph-based segmenter that degenerated unequally across models. Once each factor is controlled, the architectural advantage disappears.

Motivated by this, we run a **controlled factorial study** — three architectures × three tokenizer–model configurations × three seeds, against a prompt-agnostic feature baseline — on the LAILA Arabic AES dataset under a corrected protocol. Our findings are:

1. **Tokenization dominates architecture:** The tokenizer–model configuration moves holistic QWK by 0.16–0.19, with a stable ordering across architectures and seeds, whereas every architectural step moves it by at most 0.045 — a roughly four-fold gap.
2. **Architectural complexity rarely provides substantial benefits:** Flat AraBERT matches a simple feature baseline (0.621 QWK); the hierarchical step adds only +0.045 (and fragilely so), and multi-trait attention is mildly harmful — replicating, for Arabic and under control, the "simple-beats-complex" finding of cross-prompt AES.
3. **A documented confound.** We show how the apparent early "complexity wins" result is jointly produced by

DOI: **10.5281/zenodo.21350697**

tokenizer, selection-criterion, and segmentation choices, and dissolves under control.

4. **A corrected, reusable protocol.** A mean-per-prompt checkpoint-selection criterion (vs. the corrupted pooled criterion) and a sentence-based segmenter appropriate to Arabic, where paragraph markers are unreliable.

To the best of our knowledge, this is the first controlled factorial study of cross-prompt Arabic AES. Under identical leave-one-prompt-out splits on LAILA, our best configuration (M1-MIX, 0.666 holistic QWK) is the strongest published neural model — surpassing AraBERT (0.620), MOOSE (0.649), and ProTACT (0.578) — and is competitive with the best feature-based systems (RF 0.682, XGB 0.679) and few-shot LLMs (Fanar 0.669), without hand-crafted features or few-shot prompting.

II. RELATED WORK

Arabic AES. Arabic AES has been studied mostly in the prompt-specific setting and with classical or single-encoder neural models; AraBERT-based systems reach high agreement within prompt (e.g., QWK up to 0.88 on the AR-AES dataset [3]), while large language models reach ~0.58 on the cross-prompt LAILA dataset [4], [5]. On LAILA's own cross-prompt benchmark [4], the strongest published holistic-QWK results come from hand-crafted feature models (random forest and XGBoost, ≈0.68) and a few-shot LLM (Fanar, 0.669), while the best neural systems — AraBERT (0.620), the MOOSE cross-prompt model [6] (0.649), and ProTACT [7] (0.578) — trail them; notably, feature-based models match or exceed every neural system, foreshadowing the simplicity finding we confirm under control. A 2026 survey [5] reports no prior application of hierarchical or multi-trait architectures, or analysis of tokenization effects, to Arabic AES — the gap this paper addresses.

Cross-prompt AES and the simplicity finding. Cross-prompt AES emphasizes prompt-agnostic signal: PAES [1] learns from non-prompt-specific features (length, readability, syntactic/POS cues), and has since been extended to cross-prompt scoring of individual essay traits [8]. A growing body of work finds that simple, feature-based or low-capacity models match or beat complex neural models cross-prompt, because complex models overfit non-target-prompt essays and generalize worse; prompt-agnostic baselines are also more stable under score-distribution shift [2], [9]. We adopt this lens to frame our negative result and to motivate our feature baseline.

Hierarchical neural AES. The dominant neural design encodes essays hierarchically with the **sentence** as the base unit (sentence → document), e.g., attention-based recurrent–convolutional models [10] and, more recently, hierarchical BERT variants [11]. Our M1/M2 follow this design; crucially, the base unit is the sentence, not the paragraph.

Arabic segmentation and morphology. Arabic is morphologically rich; AraBERTv2 uses Farasa morphological pre-segmentation, whereas AraBERTv02 operates on raw text [12]. Reliable paragraph structure is often absent in Arabic documents: the AraSeg benchmark [13] explicitly categorizes documents by the availability of paragraph and punctuation markers, and sentence boundary detection in Modern Standard Arabic is reliable when punctuation is present [14]. These observations motivate our sentence-based segmentation and frame the tokenizer factor (morphological vs. raw pre-segmentation) which emerges as a key element.

Tokenization as a performance factor. Beyond Arabic, a growing literature treats the tokenizer as a first-order design decision rather than a fixed preprocessing detail. For Arabic specifically, Alrefaie et al. [15] report that BPE combined with Farasa morphological segmentation outperforms alternative strategies across several downstream tasks, attributing the gain to closer alignment with Arabic's rich morphology; morphology-aware tokenizers such as MorphBPE [16] likewise yield faster convergence and lower loss for morphologically rich languages. These findings motivate treating the tokenizer–model configuration as an explicit experimental factor and anticipate the morphological-segmentation effect (MM2 vs. MM02) we isolate here. To our knowledge, no prior Arabic AES work has measured tokenization effects under cross-prompt evaluation.

Cross-prompt scoring as domain generalization. Cross-prompt AES is an instance of domain generalization: each prompt is a domain, and the model must generalize to a held-out prompt whose score distribution is shifted [1]. Under this view, the prompt-agnostic feature baselines that resist overfitting and the implicit regularization we attribute to vocabulary mismatch are two routes to the same end — limiting the model's capacity to memorize seen-prompt idiosyncrasies. This framing links both the simplicity finding and our mismatch observation to the broader generalization literature, where constraining or perturbing the input representation is a recognized regularizer.

III. MODELS AND BASELINES

We compare three neural architectures of increasing complexity against a strong feature-based baseline, all evaluated under an identical cross-prompt protocol.

Feature-based baseline (PAES-style). Feature-based baseline (PAES-style). Following the prompt-agnostic feature paradigm [1], [2], we extract 19 prompt-independent features spanning four groups. Length: word, sentence, and paragraph counts, plus mean sentence length and mean word length. Lexical: type–token ratio and long-word ratio. Surface: punctuation and connective-word ratios. Part-of-speech: ten coarse POS ratios obtained from the CAMeL Tools MLE morphological disambiguator [17] (noun, verb, adjective, adverb, preposition, conjunction, pronoun, particle, numeral, and a residual "other" class); the ten ratios are compositional and sum to one, and Ridge's ℓ_2 penalty handles the resulting collinearity. The full feature definitions are given in Appendix A (Table A4). Features are z-standardized and scored with Ridge regression; the regularization strength is selected on the mean per-prompt development QWK — the same criterion used for the neural models. The pipeline is deterministic, so we report a single run with variance computed across the eight folds.

Notes. All features are z-standardized and scored with Ridge regression ($\alpha \in \{0.1, 1, 10, 100, 1000\}$, chosen on mean per-prompt development QWK; deterministic single run). n_w = word count; n_tokens = CAMeL word-tokenized units; each ratio is normalized by n_w or n_tokens as specified in Table A4. POS tags are the top analysis of the CAMeL Tools MLE disambiguator (calima-msa), with fine tags folded into ten coarse classes (e.g., noun

DOI: **10.5281/zenodo.21350697**

← noun/noun_prop/noun_quant; numeral ← noun_num/digit; other ← punc/abbrev/unanalyzed). † Punctuation set: . , ! ? 31 ‡ .- ( ) ' " ، : ؛ ؟ multi-character Modern Standard Arabic connectives (e.g., ثم، لكن، لأن، حيث، بينما، كما، بالإضافة، لذلك، وبالتالي).

M0 (flat). The essay is encoded by AraBERT with mean pooling, followed by a regression head.

M1 (hierarchical). The essay is split into sentence-based segments (Section 4); each segment is encoded by AraBERT and mean-pooled, segment vectors receive positional embeddings and pass through a 2-layer inter-segment Transformer (4 heads), and the pooled essay vector feeds a regression head.

M2 (multi-trait). M1 augmented with a cross-segment attention block and eight output heads (seven analytic traits plus the holistic score), trained with a trait-weighted Mean Square Error (MSE) objective.

Encoder and tokenizer configurations. All neural models use AraBERT-base [12] (≈136M parameters, pretrained with masked language modeling on ≈8.7B words). The key controlled factor is the "tokenizer–model configuration":

Table 1 Summary of Encoder and Tokenizer Configurations for Controlled Experimental Evaluation.

Code	Encoder	Tokenizer	Description
MM02	AraBERTv02	AraBERTv02	matched, no morphological pre-segmentation
MIX	AraBERTv02	AraBERTv2	mismatched (model and tokenizer vocabularies differ)
MM2	AraBERTv2	AraBERTv2	matched, Farasa morphological pre-segmentation

MM02 and MM2 are "properly matched" encoders that differ only in morphological pre-segmentation; MIX is a "vocabulary-mismatched" pairing that arose in early experiments and which we retain as an explicit factor because, as we show, it behaves as an implicit regularizer. The three configurations evaluated in this study are summarized in Table 1.

IV. EXPERIMENTAL SETUP

Dataset: We use the LAILA dataset [4], the largest publicly available Arabic Automated Essay Scoring (AES) benchmark, comprising 7,859 essays written in response to eight prompts. Each essay is annotated with a holistic score on a 0–32 scale and seven analytic trait scores: Relevance (rel), Organization (org), Vocabulary (voca), Style (sty), Development (deve), Mechanics (mech), and Grammar (gram). Relevance is scored on a 0–2 scale, whereas the remaining traits are scored on a 0–5 scale. Figure 1 illustrates the distribution of the analytic trait scores using boxplots. Overall, the score distributions are relatively consistent across traits, with median values centered around 3. Vocabulary exhibits slightly greater variability than the other traits, while the remaining traits show comparable dispersion. A small number of outliers appear at both extremes of the scale for most traits, indicating essays with exceptionally low or high trait scores. The horizontal orange line represents the overall mean score across all analytic traits, which lies close to the median of most distributions, suggesting that the dataset is reasonably balanced without pronounced skewness. Figure 2 shows the number of essays for each prompt. Six prompts contain 1,062–1,181 essays and two prompts (P3, P4) are smaller, containing 521 and 500 essays, respectively. Figure 3 shows the holistic-score distribution. The holistic distribution is zero-inflated: a large mass of essays score 0 (off-topic/blank), followed by a gap and an approximately unimodal distribution over 7–32 peaking around 18–21 (mean 16.8 ± 7.4); this zero-inflation is itself a challenge for QWK. Essays are short — mean 175 words (median 164, range 7–767), most falling *below* the nominal 300-word target (Figure 4). We adopt the eight official cross-prompt folds: each fold trains on five prompts, validates on two, and tests on one held-out prompt. This is a low-resource fine-tuning regime — the fine-tuning corpus (≈1.37M words; ≈1.66M AraBERTv02 sub-tokens) is only ~0.02% of AraBERT's pretraining data.

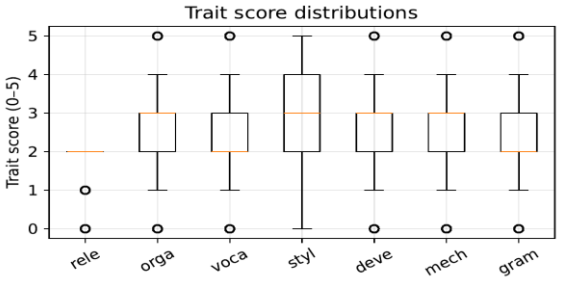

Figure 1. Analytic trait-score distributions (relevance on 0–2; others 0–5).

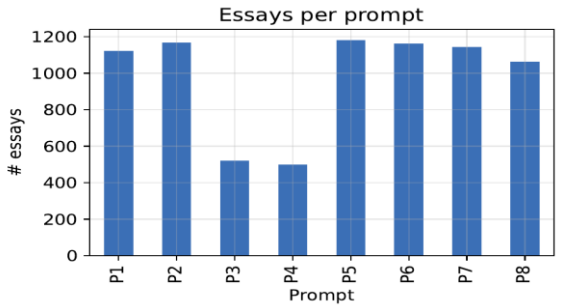

Figure 2. Essays per prompt in LAILA dataset (Six large prompts; P3 and P4 smaller).

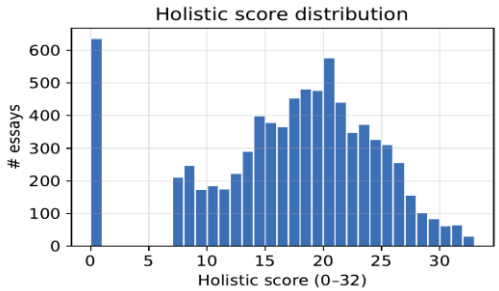

Figure 3. Holistic-score distribution: zero-inflated, then a unimodal mass over 7–32.

DOI: **10.5281/zenodo.21350697**

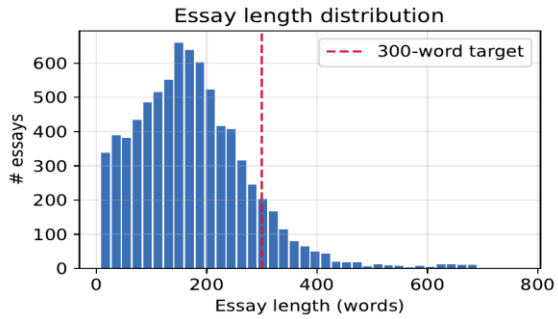

Figure 4. Essay-length distribution; most essays fall below the 300-word target.

Segmentation: Hierarchical models require a base unit. Following the standard hierarchical-AES design, in which the base unit is the *sentence* [10], and motivated by the fact that Arabic paragraph structure is unreliable — only 60.8% of LAILA essays contain blank-line paragraph breaks, consistent with the document categories of the AraSeg benchmark [13] — we segment every essay uniformly: we split on blank lines and on sentence punctuation (`.!,؛؟`), then group the resulting sentences into at most eight segments of up to 256 tokens. Under this scheme ~19% of essays form a single segment (very short essays) while the rest span 2–8 segments (Figure 5). We found that an earlier, paragraph-marker-based segmenter left ~14% more essays as a single (degenerate) segment for M2 than for M1, silently confounding the architecture comparison; the unified sentence segmenter removes this confound.

Model selection: We select the checkpoint with the highest "mean per-prompt development QWK". We show (Section 5.4) that the commonly used "pooled" development QWK is a corrupted proxy under cross-prompt evaluation: because the two development prompts have different score distributions, per-prompt offsets can drive the pooled QWK to near zero even when within-prompt ranking is strong, making checkpoint selection effectively arbitrary.

Training and evaluation. The following settings were used: batch size 16; learning rates of 2e-5 for the encoder and 1e-4 for the regression heads; and early stopping based on the model selection criterion. Following recent methodological recommendations in Automated Essay Scoring (AES) that emphasize reproducibility and robustness to stochastic optimization [18], [19], all neural models were trained using three fixed random seeds (42, 1337, and 2026). Varying the random seed controls for the stochastic effects of weight initialization, data shuffling, and dropout, enabling us to distinguish genuine architectural improvements from random optimization variance. Accordingly, we evaluate the full factorial design {M0, M1, M2} × {MM02, MIX, MM2} × {42, 1337, 2026}, in addition to the feature-based baseline. We report holistic QWK as the primary evaluation metric, together with Pearson correlation and the mean ± standard deviation of QWK across the three random seeds as an explicit measure of training stability. Throughout the study, all comparisons change only one experimental factor at a time. For statistical significance (Section 5), we employ a linear mixed-effects model with random intercepts for prompt and seed, paired Wilcoxon signed-rank tests over the eight prompts, and cluster-bootstrap confidence intervals with Holm correction.

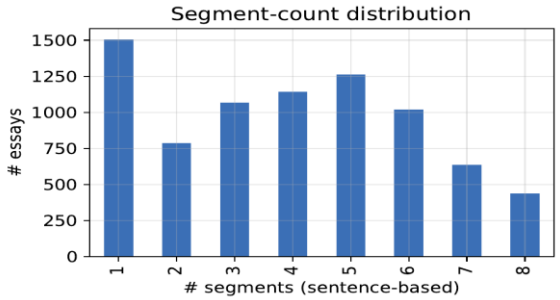

Figure 5. Segment-count distribution under the sentence-based segmenter (~19% single-segment).

V. RESULTS

Table 2 reports the full factorial (mean QWK over three seeds). Two headline patterns emerge: the tokenizer configuration moves performance far more than architecture, and added architectural complexity does not pay.

Table 2. The Holistic QWK (Holistic Working Matrix)

Model	MIX (mismatched)	MM2 (Farasa)	MM02 (matched-v02)
Feature baseline	—	—	0.621 [0.12]
M0 (flat)	0.621 ± 0.036 [0.12]	—	0.419 ± 0.011 [0.21]
M1 (hierarchical)	**0.666 ± 0.015 [0.10]**	0.614 ± 0.010 [0.15]	0.505 ± 0.007 [0.22]
M2 (multi-trait)	0.650 ± 0.014 [0.11]	0.593 ± 0.077 [0.15]	0.457 ± 0.029 [0.25]

Holistic QWK, mean over seeds {42, 1337, 2026}, unified sentence segmenter, mean-per-prompt selection.

Cell format: mean ± between-seed SD [mean across-fold SD]. The baseline is deterministic (single run; bracket is its across-fold SD).

A. *The tokenizer's configuration is the influential factor*

The ordering MIX > MM2 > MM02 is identical for M1 and M2 and stable across all three seeds (between-seed SD ≤ 0.015 for M1). The configuration (Figure 6) spans a range of **+0.161** for M1 (0.505 → 0.666) and **+0.193** for M2 (0.457 → 0.650). Decomposing the effect: morphological pre-segmentation helps over the matched baseline (MM2 − MM02 = +0.109 for M1, +0.136 for M2), and the mismatched pairing helps further still (MIX − MM2 = +0.052 for M1, +0.057 for M2). The matched configurations are also markedly less stable: MM02 and MM2 exhibit per-fold collapses (individual prompts falling below 0.10) and seed sensitivity, whereas MIX is consistently the most stable (per-fold SD ≈ 0.09).

DOI: **10.5281/zenodo.21350697**

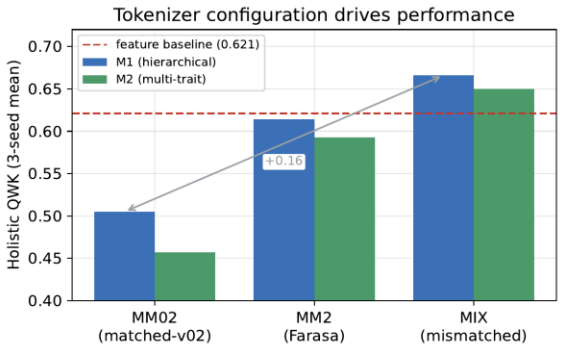

Figure 6. Holistic QWK by tokenizer's configuration for M1/M2 (3-seed mean); dashed line = feature baseline.

B. *Architectural complexity barely helps*

Reading the stable MIX column top to bottom isolates the architecture effect on a clean footing (Figure 7):

Table 3. Stable MIX column (Structural Effect Isolation)

Model	QWK (MIX)	Δ
Feature baseline	0.621	—
M0 (flat BERT)	0.621	+0.000
M1 (hierarchical)	0.666	+0.045
M2 (multi-trait)	0.650	−0.016

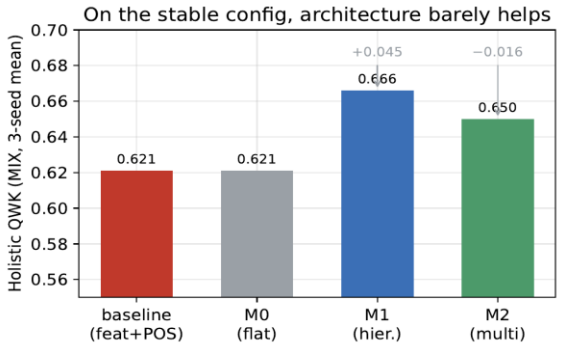

Figure 7. Architecture effect on the stable MIX configuration: baseline = M0; M1 +0.045; M2 −0.016.

A review of Table 3 and Figure 7 reveals three observations: (i) Flat BERT (M0) **matches** the feature baseline (0.621), and the baseline outperforms the neural models in four of the six neural configurations (it is beaten only in the MIX column). (ii) The hierarchical step yields only +0.045, and this gain is fragile — it is driven by a single seed (M1−M0 = +0.114 at seed 42 but +0.014 and +0.008 at the other two seeds). (iii) The multi-trait step is negative. In other words, every architectural step moves QWK by ≤0.045 and one of them is harmful, while the tokenizer factor moves it by up to 0.19 — tokenization dominates architecture by roughly 4×.

C. *Multi-trait attention does not add over the hierarchy (negative result)*

M2 is below M1 in **every** configuration: −0.016 (MIX), −0.021 (MM2), −0.048 (MM02), and M2 is also less stable. This replicates, in Arabic and under controlled conditions, the finding that added neural complexity does not help cross-prompt AES [2], [9]; the accepted mechanism — complex models overfit non-target-prompt essays — also accounts for the instability of the matched-encoder configurations here.

D. *Anatomy of an architectural illusion*

Our early, uncontrolled experiments suggested a clean ascending story (M2 = 0.637 ≫ M1 = 0.556). Under control these collapses (Figure 8): that comparison confounded **three** factors at once — M1 used the matched tokenizer while M2 used the mismatched one; the corrupted pooled-dev selection criterion; and a paragraph-based segmenter that degenerated more often for M2. Holding the tokenizer fixed and fixing both the selection criterion and the segmenter, the architectural difference vanishes (M2 − M1 ≈ −0.02). This is the paper's cautionary core: in low-resource cross-prompt AES, plausible **architectural gains** can be artifacts of tokenization, evaluation, and preprocessing choices unless each is controlled explicitly.

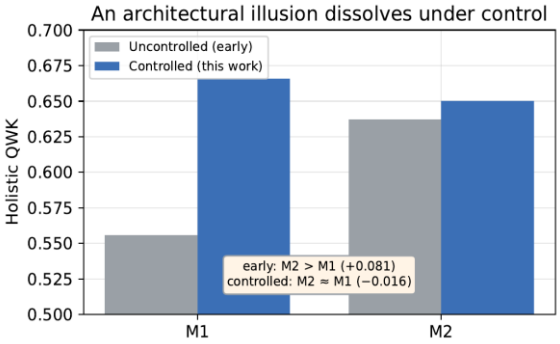

Figure 8. The early, uncontrolled M2>M1 result reverses to M2≈M1 once tokenizer, criterion, and segmenter are fixed.

E. *Statistical Significance*

We assess significance with a linear mixed-effects model (LMM) as the primary test — QWK ~ architecture × tokenizer with random intercepts for prompt and seed (192 observations) — complemented by paired Wilcoxon signed-rank tests over the eight prompts and cluster-bootstrap 95% confidence intervals (10,000 resamples over prompts), with Holm correction across the pairwise family. In the LMM, the **tokenizer main effect is large and highly significant** (likelihood-ratio test $p < 10^{-7}$; MIX vs. matched-v02 estimate +0.161, 95% CI [0.089, 0.233]; Farasa vs. matched-v02 +0.108 [0.036, 0.181]), whereas the **architecture effect is not significant** (M2 vs. M1 −0.048, 95% CI [−0.121, 0.024]), and neither is the interaction. Prompt identity accounts for roughly half the variance (random-intercept variance 0.016 ≈ residual 0.016). The bootstrap CIs concur: all tokenizer contrasts and M1-MIX vs. the feature baseline (+0.045, [0.016, 0.073]) exclude zero, while every M2−M1 contrast includes zero within a tight interval (e.g., MIX: −0.016, [−0.044, 0.012]) — evidence of equivalence, not mere non-rejection. With only eight prompts, the per-prompt Wilcoxon tests are conservative and do not survive Holm correction even for the large tokenizer effects; we therefore rely on the LMM and report the Wilcoxon results for transparency. In short: the tokenizer effect is statistically robust, whereas the hierarchical gain and the multi-trait difference are not

DOI: **10.5281/zenodo.21350697**

statistically distinguishable from zero — quantitatively reinforcing the thesis.

Table 4 provides a detailed summary of the principal pairwise comparisons underlying the statistical analysis. For each comparison, we report the estimated change in holistic QWK (Δ), the corresponding 95% confidence interval obtained from cluster bootstrap, and the paired Wilcoxon p-values before and after Holm correction. Together with the linear mixed-effects model, these results quantify both the magnitude and the statistical reliability of the tokenizer and architectural effects.

Table 4. Pairwise comparisons (Δ = mean per-prompt difference; 95% bootstrap CI; Wilcoxon p; Holm-adjusted p; n = 8).

Comparison	Δ	95% CI	p	p (Holm)
M1: MIX − MM02 (tokenizer)	+0.161	[+0.055, +0.278]	0.055	0.66
M1: Farasa − MM02	+0.108	[+0.018, +0.199]	0.109	0.98
M2: MIX − MM02 (tokenizer)	+0.193	[+0.056, +0.342]	0.039	0.51
MIX: M2 − M1 (negative result)	−0.016	[−0.044, +0.012]	0.383	1.00
MM02: M2 − M1	−0.048	[−0.144, +0.029]	0.641	1.00
MIX: M1 − M0 (C1)	+0.045	[+0.007, +0.090]	0.148	1.00
M1-MIX − baseline	+0.045	[+0.016, +0.073]	0.078	0.86

VI. DISCUSSION

A. Why does the tokenizer configuration dominate?

Two distinct effects compose the +0.16–0.19 range. First, the AraBERTv2 vocabulary — trained over Farasa pre-segmented text but applied here to raw, unsegmented input — breaks words into markedly finer sub-units than the matched AraBERTv02 vocabulary (MM2 vs. MM02, +0.11–0.14). Second, and more surprisingly, the "mismatched" pairing (MIX vs. MM2, +0.05–0.06) is the single best and most stable configuration. We interpret the mismatch as a source of implicit regularization: a tokenizer whose vocabulary does not exactly match the encoder's pre-training perturbs the input representation in a way that slows memorization of the seen prompts and preserves cross-prompt generalization longer — consistent with the accepted mechanism that high-capacity models overfit non-target-prompt essays. The matched configurations, by contrast, fit the seen prompts faster and collapse on individual held-out prompts (per-fold QWK below 0.10), producing the high variance we observe.

B. Anatomy of the mismatch: fragmentation, not unknown tokens.

To characterize the mismatched configuration concretely, we tokenized the full LAILA corpus (≈1.37M words) with both vocabularies (Table 5). The AraBERTv2 vocabulary — built over Farasa-segmented text but here applied to raw, unsegmented essays — fragments words far more finely than the matched AraBERTv02 vocabulary: 1.77 versus 1.21 sub-tokens per word (2.43M vs. 1.66M sub-tokens), a 46% increase. Critically, this is not an out-of-vocabulary effect: the proportion of words containing an unknown ([UNK]) token is essentially identical across the two vocabularies (1.83% vs. 1.80%), and at the sub-token level [UNK] is no higher — indeed slightly lower — under the mismatch (1.04% vs. 1.49%). Those ≈1.8% [UNK]-bearing words are out-of-script tokens (Latin strings, digits) common to both settings, not a product of the mismatch. Instead, the mismatch acts through finer segmentation of ordinary Arabic words: the frequent form للرياضة("for sport") is a single token under AraBERTv02 but shatters into five near-character pieces under AraBERTv2 (##ل ##ر ##ي ##ا ##ضة), and the same one-to-five pattern holds across clitic-prefixed forms (للأجيال, للاستفادة, للأجهزة). Because the v2 vocabulary presupposes that such proclitics have already been split off by Farasa — a step not applied to the input at inference — fused surface forms fall off the vocabulary's frequent-token manifold and are decomposed into short, high-frequency sub-units. We read this higher-granularity input as the source of the implicit regularization: it slows memorization of prompt-specific surface forms and preserves cross-prompt generalization, consistent with the lower variance and higher held-out QWK of the mismatched configuration.

Table 5. Tokenization of the LAILA corpus

Vocabulary	Sub-tokens / word	Total sub-tokens	[UNK] (sub-token)	Words with [UNK]
AraBERTv02 (matched)	1.21	1.66M	1.49%	1.80%
AraBERTv2 on raw text (mismatch)	1.77	2.43M	1.04%	1.83%

C. Why does complexity not help?

Cross-prompt generalization rewards prompt-agnostic signal. The hierarchical encoder offers a small, fragile benefit over the flat one, but multi-trait attention — extra capacity tied to trait-specific structure — does not transfer to unseen prompts and slightly hurts, mirroring English cross-prompt findings. That a flat encoder merely matches, and the deeper models often *under*-perform, a 19-feature prompt-agnostic baseline underscores how little the neural machinery contributes once prompt-specific cues are unavailable.

D. Implications:

For practitioners, the highest-leverage decisions in low-resource cross-prompt Arabic AES are tokenization/pre-segmentation and a sound evaluation protocol, not encoder depth or trait heads. For researchers, the result is a cautionary template: architecture comparisons must hold tokenizer, selection criterion, and segmentation fixed, and report multiple seeds with variance, or risk reporting

DOI: 10.5281/zenodo.21350697

confounded gains. Finally, we frame these conclusions within our testbed: LAILA is a low-resource, zero-inflated corpus, and both its large mass of zero (off-topic) scores and its small fine-tuning set favour prompt-agnostic robustness over high-capacity architectures; the balance between simple and complex models could shift on larger or less skewed Arabic corpora.

E. Comparison with published LAILA benchmarks

Table 6 places our systems against the published LAILA cross-prompt benchmark, evaluated on the identical eight-fold leave-one-prompt-out splits and the same holistic QWK metric. our systems (M0/M1/M2-MIX, in bold) are three-seed means, while LR/RF/XGB/NN are hand-crafted-feature models and Fanar is a few-shot LLM. Our best configuration, M1-MIX (0.666), is the strongest neural system — surpassing AraBERT (0.620), MOOSE (0.649), and ProTACT (0.578) — and is competitive with the strongest hand-crafted feature models (RF 0.682, XGB 0.679) and the best few-shot LLM (Fanar, 0.669), without feature engineering or few-shot prompting. That feature-based models match or exceed every neural system in the original benchmark independently corroborates our finding that architectural complexity does not pay cross-prompt. In contrast to the single-run published baselines, our figures are three-seed means with reported variance under the corrected mean-per-prompt selection criterion.

Table 6. Holistic QWK in the cross-prompt setting on LAILA Dataset.

System	Type	Holistic QWK
RF (816 features)	Feature	0.682
XGB (816 features)	Feature	0.679
Fanar (5-shot)	LLM	0.669
M1-MIX (ours)	**Neural**	**0.666**
LR (816 features)	Feature	0.661
NN (816 features)	Feature	0.651
M2-MIX (ours)	**Neural**	**0.650**
MOOSE	Neural	0.649
M0-MIX (ours)	**Neural**	**0.621**
AraBERT	Neural	0.620
ProTACT	Neural	0.578

F. Robustness to seed count

To address the concern that three seeds may be too few, we re-ran the stable MIX column — on which our architectural conclusions rest — at two further seeds (five in total: 42, 1337, 2026, 7, 123; Table A2 the final column is the five-seed mean). The holistic-QWK means shift by at most 0.010 (M1-MIX 0.666 → 0.661, M2-MIX 0.650 → 0.650, M0-MIX 0.621 → 0.631), the between-seed SD stays small (≤0.034), and the architectural picture is unchanged: the hierarchical gain remains modest (M1 − M0 = +0.031) and the multi-trait step remains non-positive (M2 − M1 = −0.011), reinforcing our central finding. The matched-encoder columns, which serve only to establish the large tokenizer effect, retain three seeds.

VII. LIMITATIONS

The best configuration is mismatched: Counterintuitively, the mismatched configuration (MIX), which couples the Farasa tokenizer with the AraBERT-v02 encoder despite their divergent vocabularies, consistently achieves the strongest and most stable performance across all eight cross-prompt folds. We attribute this to an incidental regularization effect induced by vocabulary mismatch — an empirically reproducible yet mechanistically unexplained phenomenon. Accordingly, we document it as a noteworthy finding rather than a deployment recommendation, and leave its mechanistic account to future investigation.

Instability of matched configurations: The matched encoders exhibit per-fold collapses; while this is reproducible across three seeds, it complicates point estimates and we therefore emphasize the stable MIX column for architectural conclusions.

Seeds, dataset, encoder family, and language: The full factorial uses three seeds; we additionally validated the stable MIX column — on which our architectural conclusions rest — at five seeds (42, 1337, 2026, 7, 123; Table A2), where the means shift by at most 0.010 and the conclusions are unchanged (between-seed SD is small, 0.007–0.034). We use a single dataset (LAILA), a single encoder family (AraBERT), and one language; generalization to other Arabic corpora, encoder families, and languages is untested.

Holistic focus: We analyze holistic QWK; a full multi-trait evaluation is left to future work.

VIII. CONCLUSION

In this paper, the first controlled factorial study of cross-prompt Arabic AES is presented. Holding tokenizer, model-selection criterion, and segmentation fixed and varying one factor at a time across three seeds, it is found that **the tokenizer–model configuration is the dominant lever**: it shifts holistic QWK several-fold more than any architectural step, while the flat-to-hierarchical-to-multi-trait progression contributes little and one step is harmful. Flat AraBERT only matches a simple prompt-agnostic feature baseline, which in turn beats the deeper neural models in most configurations. We further showed how an apparent "complexity wins" result is an artifact of uncontrolled tokenization, evaluation, and preprocessing, and we release a corrected protocol — mean-per-prompt checkpoint selection and Arabic-appropriate sentence segmentation. In low-resource cross-prompt AES, getting tokenization and evaluation right matters more than adding architectural complexity.

DOI: **10.5281/zenodo.21350697**

APPENDIX A — DETAILED TABLES

Table A1. Full factorial (Holistic QWK, 3-seed mean ± between-seed SD [mean across-fold SD]).

Model	MIX	MM2 (Farasa)	MM02 (matched-v02)
Feature baseline	—	—	0.621 [0.12]
M0 (flat)	0.621 ± 0.036 [0.12]	—	0.419 ± 0.011 [0.21]
M1 (hierarchical)	**0.666 ± 0.015 [0.10]**	0.614 ± 0.010 [0.15]	0.505 ± 0.007 [0.22]
M2 (multi-trait)	0.650 ± 0.014 [0.11]	0.593 ± 0.077 [0.15]	0.457 ± 0.029 [0.25]

Table A2. Per-seed holistic QWK, MIX column, across five seeds.

System	seed 42	seed 1337	seed 2026	seed 7	seed 123	mean (5)
baseline	0.621	—	—	—	—	0.621
M0-MIX	0.572	0.635	0.656	0.620	0.671	0.631
M1-MIX	0.686	0.649	0.664	0.637	0.671	0.661
M2-MIX	0.665	0.632	0.654	0.654	0.646	0.650

Table A3. Per-prompt QWK (3-seed mean), MIX column + baseline.

cell	p1	p2	p3	p4	p5	p6	p7	p8
M0-MIX	0.45	0.67	0.77	0.70	0.61	0.64	0.63	0.49
M1-MIX	0.46	0.73	0.81	0.70	0.61	0.63	0.75	0.65
M2-MIX	0.44	0.75	0.79	0.69	0.59	0.68	0.68	0.57
baseline	0.37	0.63	0.82	0.65	0.53	0.62	0.69	0.66

Table A4. The 19 prompt-agnostic baseline features (PAES-style), grouped by type.

#	Feature	Group	Definition
1	Word count	Length	Number of whitespace-delimited tokens (n_w)
2	Sentence count	Length	Sentences after splitting on . ! ? ؟ ۔and newlines
3	Paragraph count	Length	Blank-line–separated blocks (minimum 1)
4	Mean sentence length	Length	n_w ÷ number of sentences
5	Mean word length	Length	Mean number of characters per word
6	Type–token ratio	Lexical	(unique words) ÷ n_w
7	Long-word ratio	Lexical	(words longer than 6 characters) ÷ n_w
8	Punctuation ratio	Surface	(punctuation characters) ÷ n_w †
9	Connective ratio	Surface	(discourse-connective tokens) ÷ n_w ‡
10	Noun ratio	POS	(noun-tagged tokens) ÷ n_tokens
11	Verb ratio	POS	(verb-tagged tokens) ÷ n_tokens
12	Adjective ratio	POS	(adjective-tagged tokens) ÷ n_tokens
13	Adverb ratio	POS	(adverb-tagged tokens) ÷ n_tokens
14	Preposition ratio	POS	(preposition-tagged tokens) ÷ n_tokens
15	Conjunction ratio	POS	(conjunction-tagged tokens) ÷ n_tokens
16	Pronoun ratio	POS	(pronoun-tagged tokens) ÷ n_tokens
17	Particle ratio	POS	(particle-tagged tokens) ÷ n_tokens
18	Numeral ratio	POS	(numeral/digit-tagged tokens) ÷ n_tokens
19	Other ratio	POS	(residual: punctuation, abbrev., unanalyzed) ÷ n_tokens

Dataset statistics: 7,859 essays; length mean 175 words (median 164, range 7–767); holistic mean 16.8 ± 7.4 (range 0–32, zero-inflated); segments/essay mean 4.0 (19.2% single-segment). Per-prompt counts: P1 1122, P2 1168, P3 521, P4 500, P5 1181, P6 1162, P7 1143, P8 1062. Corpus: 1.37M words → 1.66M v02 sub-tokens (1.21/word) vs 2.43M v2 sub-tokens (1.77/word).

ACKNOWLEDGMENT

We begin by acknowledging our profound appreciation to Allah for facilitating the completion of this research study.

Furthermore, we would like to express our heartfelt appreciation to the Libyan Academy and the Libyan Ministry of Higher Education & Scientific Research for their logistical support throughout this research study.

REFERENCES

[1] R. Ridley, L. He, X. Dai, S. Huang, and J. Chen, "Prompt Agnostic Essay Scorer: A Domain Generalization Approach to Cross-Prompt Automated Essay Scoring," arXiv:2008.01441, 2020.

[2] S. Li and V. Ng, "Conundrums in Cross-Prompt Automated Essay Scoring: Making Sense of the State of the Art," in *Proc. 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024, pp. 7661–7681.

[3] R. Ghazawi and E. Simpson, "Automated Essay Scoring in Arabic: A Dataset and Analysis of a BERT-Based System," arXiv:2407.11212, 2024.

[4] M. Bashendy, W. Massoud, S. Eltanbouly, S. Albatarni, M. Sayed, A. Abir, H. Bouamor, and T. Elsayed, "LAILA: A Large Trait-Based Dataset for Arabic Automated Essay Scoring," in *Proc. 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2026, pp. 3074–3091.

[5] K. Dahimi, H. Cherroun, and A. Belabbaci, "Automated Scoring of Arabic Text Using Large Language Models: A Literature Review," arXiv:2606.09830, 2026, to appear at NCMAI 2026.

[6] P.-K. Chen, B.-W. Tsai, S. K. Wei, C.-Y. Wang, J.-C. Wang, and Y.-T. Huang, "Mixture of Ordered Scoring Experts for Cross-prompt Essay Trait Scoring," in *Proc. 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025, pp. 18071–18084.

[7] H. Do, Y. Kim, and G. G. Lee, "Prompt- and Trait Relation-aware Cross-prompt Essay Trait Scoring," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

[8] R. Ridley, L. He, X. Dai, S. Huang, and J. Chen, "Automated Cross-Prompt Scoring of Essay Traits," in *Proc. AAAI Conference on Artificial Intelligence*, vol. 35, no. 15, 2021, pp. 13745–13753.

DOI: **10.5281/zenodo.21350697**

[9] K. Yang, M. Raković, Y. Li, Q. Guan, D. Gašević, and G. Chen, "Unveiling the Tapestry of Automated Essay Scoring," arXiv:2401.05655, 2024.

[10] F. Dong, Y. Zhang, and J. Yang, "Attention-Based Recurrent Convolutional Neural Network for Automatic Essay Scoring," in *Proc. CoNLL 2017*, 2017, pp. 153–162.

[11] J. Xue, X. Tang, and L. Zheng, "A Hierarchical BERT-Based Transfer Learning Approach for Multi-Dimensional Essay Scoring," *IEEE Access*, vol. 9, pp. 125403–125415, 2021.

[12] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-Based Model for Arabic Language Understanding," in *Proc. Fourth Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT4), LREC*, 2020.

[13] M. Elkholy, K. Elmadani, N. Habash, and B. Alhafni, "Arabic Sentence Segmentation Across Genres and Punctuation Conditions (AraSeg)," arXiv preprint, 2026.

[14] C.-E. González-Gallardo and J.-M. Torres-Moreno, "Automated Sentence Boundary Detection in Modern Standard Arabic Transcripts Using Deep Neural Networks," *Procedia Computer Science*, 2018.

[15] M. T. Alrefaie, N. E. Morsy, and N. Samir, "Exploring Tokenization Strategies and Vocabulary Sizes for Enhanced Arabic Language Models," arXiv:2403.11130, 2024.

[16] E. Asgari, Y. El Kheir, and M. A. Sadraei Javaheri, "MorphBPE: A Morpho-Aware Tokenizer Bridging Linguistic Complexity for Efficient LLM Training Across Morphologies," arXiv:2502.00894, 2025.

[17] O. Obeid et al., "CAMeL Tools: An Open-Source Python Toolkit for Arabic NLP," in *Proc. Language Resources and Evaluation Conference (LREC)*, 2020.

[18] H. F. Doewes, "Rethinking Automated Essay Scoring," Ph.D. dissertation, Eindhoven Univ. of Technology, Eindhoven, The Netherlands, 2026.

[19] N. Gorbachev, "Automated Essay Scoring for Russian as a Second Language: A Deep Ordinal Learning Approach," M.S. thesis, Univ. of Bologna, Bologna, Italy, 2026.

DOI: **10.5281/zenodo.21350697**