One Size Does Not Fit All: Correlation-Aware Bundling with GED Augmentation for Arabic Multi-Trait Essay Scoring

Anonymous ACL submission

Abstract

In this paper, we propose an Arabic multi-trait AES model that explicitly addresses the heterogeneous nature of writing-quality traits through correlation-driven bundling. We group related traits into bundles in which each bundle receives tailored auxiliary signals: prompt-aware encoding for content-structure traits and BiLSTM-encoded sequential grammatical error detection (GED) signals for surface and language accuracy traits, supplemented by 549 hand-crafted features. Under Cross-prompt Leave-One-Prompt-Out (LOPO) evaluation, the proposed model achieves a mean QWK of 0.72 on the 6-content-traits subset and 0.62 on the 7-traits LAILA dataset, indicating substantial agreement with human annotators and outperforming the LAILA baseline (0.54). Our system achieves competitive performance with substantially lower deployment cost than LLM feature-based Qayyem(TRATES) model (0.65).A binary relevance classifier achieves macro-F1 of 0.90, completing a fully automated pipeline from topic detection to trait scoring. To the best of our knowledge, this is the first multi-trait Arabic AES system to integrate sequential GED signals and apply correlation-driven trait bundling.

1 Introduction

Automated Essay Scoring (AES) has been an active area of NLP research for decades; recently, AES systems have achieved near-human agreement on English datasets such as ASAP and ASAP++ (Yang et al., 2020; Ridley et al., 2021; Do et al., 2023). However, these systems primarily focus on holistic scoring, while multi-trait scoring, especially in cross-prompt settings[1], remains a challenging problem. The multi-trait scoring model evaluates the essay across multiple writing quality dimensions simultaneously, with different traits targeting different aspects of writing proficiency (Lagakis and Demetriadis, 2021; Ke and Ng, 2019).

The linguistic characteristics of Arabic, including its rich morphology, orthographic ambiguity, diglossia, and data scarcity, have slowed progress toward within-prompt holistic scoring on small datasets (Bashendy et al., 2024; Ghazawi and Simpson, 2024). The recent public dataset LAILA opens the possibility of cross-prompt, multi-trait scoring in Arabic (Bashendy et al., 2026). Yet, different approaches to LAILA either rely on hand-crafted features (Sayed et al., 2025) or prompt LLMs (Elbahnasawi et al., 2026). To the best of our knowledge, no prior work addresses the fundamental question of how to design an architecture that respects the heterogeneity of traits, where, for example, grammatical accuracy depends on local patterns, while organisation depends on discourse structure.

We present a comprehensive study of our proposed multi-trait Arabic AES pipeline, spanning data analysis, scoring strategy, model architecture, and relevance classification. Moreover, we propose a trait-bundled, GED-Augmented trait prediction model, where each bundle groups two to three traits and receives tailored auxiliary signals. Our proposed model achieves a mean QWK of 0.72 on the on-topic subset of LAILA (6 content traits) and 0.62 on the full dataset (7 traits). We complete the pipeline with a binary topic-relevance classifier that works as an off-topic essay filter (F1 = 0.90).

Beyond effectiveness, our pipeline is designed for practical deployment: scoring relies entirely on encoder-based models and deterministic feature extraction, avoiding the per-essay LLM forward passes that dominate the computational cost of recent LLM-based AES systems such as TRATES.

Our contributions are as follows:

- A comprehensive analysis of LAILA, identifying trait correlation clusters which motivate the bundle design

[1]The cross-prompt evaluation or (Leave-One-Prompt-Out / LOPO) is when the model scores essays responding to a prompt it has never seen during training.

- A traits-aware bundle AraBERT architecture with GED-Seq and hand-crafted features.
- A prompt-relevance classifier that completes the pipeline from on/off-topic detection to trait scoring.

The paper is organised as follow: §2 for related work. §3 presents a dataset analysis. §4 proposes the model, and §5 the topic relevance classifier. §6 discusses the results and we conclude in §7.

2 Related work

Multi-trait essay scoring has been widely addressed recently, which aims to assess the writing quality across multiple traits. Previous work has explored various approaches, from traditional feature-based methods to neural models, and LLM prompt-based approaches (Uto, 2021; Hussein et al., 2019; Li and Ng, 2024). Ridley et al. (2021) exploit cross-prompt multi-trait scoring by utilising a trait attention mechanism and prompt independent features on ASAP++. Do et al. (2023) improved on this with ProTACT, using prompt-aware representations and inter-trait similarity penalties. In subsequent work, Do et al. (2024) reframed multi-trait scoring as autoregressive generation, followed by self-explainable rationale-driven prediction (Do et al., 2026). A BERT encoder has been used by Cho et al. (2024) to predict both holistic and trait scores. Chu et al. (2025) utilise LLMs, which underperform fine-tuned models in multi-trait settings; however, they show higher performance on trait rationales. Cross-prompt generalisation is another major challenge in AES. Many works investigate cross-prompt AES-like feature-based methods that employ prompt-independent features (Ridley et al., 2020), contrastive learning for prompt-invariant representations (Chen and Li, 2023), and grammar-aware encoding with POS embeddings (Do et al., 2025). Most of them have been conducted on English datasets (ASAP, ASAP++, ELLIPSE).

2.1 Arabic Automated Essay Scoring

Arabic AES has been challenged by data scarcity. QAES is the first trait-specific annotated dataset for Arabic AES, comprising 195 Arabic essays (Bashendy et al., 2024) . In 2024, Ghazawi and Simpson (2024) introduced AR-AES (2046 essays) and proposed a within-prompt scoring model using AraBERT, which achieves 79.49% matching predictions. Recently Bashendy et al. (2026) present LAILA, a significant milestone: 7859 essays with seven trait scores across eight prompt settings. The TAQEEM 2025 shared task further benchmarked Arabic cross-prompt scoring, where the leading models used GPT-4o via few-shot prompting (Bashendy et al., 2025). Sayed et al. (2025) proposed 816 handcrafted features tailored for Arabic AES. Eltanbouly et al. (2025) proposed TRATES, a rubric-guided LLM scoring framework evaluated on English AES. Building on these works, Elbahnasawi et al. (2026) introduced Qayyem, a real-time Arabic AES platform that adapted the TRATES framework to Arabic and utilised the features to construct a multi-trait scoring system, which was evaluated on LAILA. Our study builds on these works, utilising LAILA, the 816 features as additional information, and evaluating against Qayyem (TRATES) as an external benchmark.

3 Datasets

LAILA dataset has 7859 Arabic essays written by high school students in Qatar. Each essay responds to one of 8 prompts covering two essay types: explanatory and persuasive. The essays are scored on 7 traits: relevance (REL), organization (ORG), development (DEV), vocabulary (VOC), style (STY), mechanics (MEC) and grammar (GRA). The holistic score is the sum of all traits (range 0–32) (Bashendy et al., 2026).

3.1 Analysis of LAILA

LAILA has seven traits, where all are content-related except REL and are scored on a scale of 0 to 5, while REL uses a three-point scale from 0 to 2. Table 1 presents the score distribution for the traits. The content traits have approximately a bell-curve distribution centred around 2-3[2], where the lower limit of $\approx$ 640 essays, score 0 across all traits, representing off-topic, including possible AI-generated or copied text that annotators judged as not addressing the prompt, hence all other traits are scored 0 if REL is 0. The REL 1-scored essays ($\approx$ 461 essays) represent partially relevant responses where some ideas relate to the prompt, while others do not. These essays received non-zero content trait scores, which creates ambiguity in the training signals. The highest score of 5 is relatively rare, reflecting the rubric's stringent criteria for assigning this score. Table 2 presents statistics per trait. All content traits show similar distributional properties: the mean ranges from 2.3 to 2.6, with standard deviations ranging from 1.1 to 1.3.

[2]see Appendix A for distribution plots

Score	REL	ORG	DEV	VOC	STY	MEC	GRA
0	636	642	644	641	645	641	642
1	461	920	672	1249	848	1040	1253
2	6762	1673	1817	2191	1723	2197	2127
3	–	2721	2961	2226	2443	2678	2513
4	–	1496	1597	1361	1690	1142	1112
5	–	407	168	191	510	161	212

Table 1: Score distribution across traits. REL uses a 0–2 scale; all other traits use 0–5.

	REL	ORG	DEV	VOC	STY	MEC	GRA	Hol.
Mean	1.78	2.60	2.60	2.38	2.66	2.40	2.36	16.78
Median	2	3	3	2	3	3	2	18
Mode	2	3	3	3	3	3	3	0
Std	0.58	1.29	1.20	1.24	1.32	1.19	1.22	7.43

Table 2: Descriptive statistics per trait (N = 7859).

3.1.1 Trait Correlation Analysis

Traits are not independent; they are connected through the cognitive and linguistic processes involved in writing. A student who can organise ideas clearly is most likely to develop them throughout the text. If the student has a rich vocabulary, then he can demonstrate style more effectively. Understanding the relationships among writing-quality traits is essential for building an accurate multi-trait essay-scoring model.

We explore Pearson correlations in three different data configurations: (1) the entire data set, including all 7 traits, where the other setting shows only the correlation of the 6 content traits, as follows: (2) essays where REL $\geq$ 1 (excluding off-topic), (3) essays with REL 2 only (fully on-topic).

The Trait Bundles The six content traits form three stable bundles in all data configurations: (MEC, GRA), (STY, VOC), and (ORG, DEV), as shown in Figure 1 which shows the correlation heatmap for the on-topic subsect, while Figure 4 in Appendix B represent the other data configuration. The strongest pair is MEC and GRA, where both assess the surface text without deep content comprehension. In the LAILA rubric, the MEC assesses spelling accuracy and punctuation placement, while GRA evaluates the diversity and correctness of syntactic structures. The second pair is highly correlated, STY and VOC. The two traits capture linguistic features: STY assesses discourse-level connectors and cohesive devices, and VOC measures lexical diversity and word choice. Finally, we can see the correlation between ORG and DEV, as they capture the relationship in content structure. ORG evaluates cohesive discourse organization, while DEV assesses the depth and development of the ideas throughout the text. In Figure 4.A in Appendix B, we can also see the correlation between REL and DEV, where REL focuses on alignment between the prompt and the text, and DEV assesses elaboration of ideas in the written text and engagement with the topic.[3]

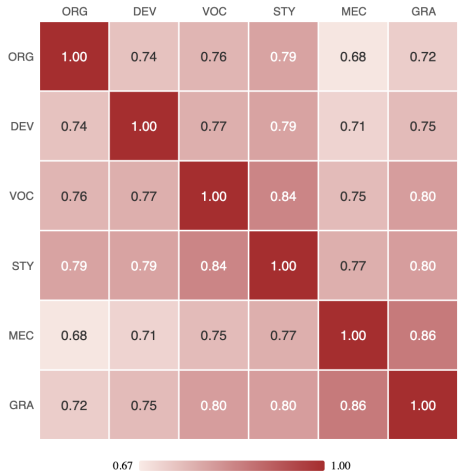

Figure 1: Pearson correlation between the 6 content traits for the on-topic subset (REL = 2, N = 6762)

Since these trait bundles are consistent across all data configurations (full dataset, REL $\geq$ 1 , REL = 2), we can be sure that the clusters are not due to dataset construction but are linguistic in nature.

These findings motivate our design for the 3-bundle framework (see §4). The shared features between the traits found in each bundle show that they need similar signals; MEC and GRA are aided by GED signals, STY and VOC need lexical sophistication signals, while ORG and DEV are helped by prompt-aware encoding.

3.2 The Relevance Problem

REL -unlike the other 6 content traits- functions as a categorical choice assessing topic adherence which creates systematic issues for multi-trait essay scoring models.

Compressed Scale and Class Imbalance. REL has a score range from 0 to 2, while other traits are evaluated on a scale of 0 to 5. Table 3 presents the distribution skew in the datasets: approximately 86.0% of the essays receive a score of 2, about 5.9% receive a score of 1, and 8.1% receive a score of 0. For most of the essays, REL lacks discriminative textual correlates; for example, although the essays with REL 0 are shorter than the essays with REL 2,

[3] See Appendix B for more correlation analysis.

REL	#Essays	Avg Words	Avg Sent
0	636	115.71	3.25
1	461	131.17	2.89
2	6762	183.20	4.20

Table 3: Essay statistics by relevance level.

the partially relevant essays (REL 1) overlap with REL 2 essays in length. We discuss the ambiguity of the annotation guidelines as well as the empirical impact of the REL on the model performance in Appendix C.

In order to develop a fully automated, accurate assessment model, building on the data quality analysis and REL observations, we treat REL as a filter and train the models to predict the six content traits only for the on-topic essays (REL 2). This separation allows us to avoid the corruption of trait-level training by off-topic documents and improves the learning of an accurate 6-trait scoring model.

4 GED-Augmented Multi-Trait Scoring Model

We propose a GED-augmented, multi-trait essay-scoring model. Figure 2, shows the proposed model architecture. The model splits the traits into 3 bundles, based on the trait correlations discussed in §3.1.1. Even though we finetune the same AraBERT model for each bundle, there are three different conditions: (i) encoding with or without prompts, (ii) presence of the GED sequence signals, and (iii) usage of the appropriate auxiliary features for each bundle.

4.1 The 3-Bundle Model

We partition the scoring traits into three bundles, each implemented as an independent AraBERT model with bundle-specific auxiliary signals. The proposed architecture has two operating modes, for the full dataset experiments , Bundle 0 includes REL, ORG, and DEV (7 traits total).While, for the production pipeline, REL is handled by the relevance classifier -as will discussed in §5-, and Bundle 0 scores only ORG and DEV (6 traits). For both experiments, VOC and STY form bundle 1, while MEC and GRA form bundle 2.

Prompt-Type Encoding The model receives a 2-dimensional one-hot vector that encodes the prompt type -provided as metadata in LAILA - as either explanatory or persuasive. The vector is projected through the Linear-ReLU model to present an 8-dimensional vector. The project from the binary signal is to interact with the other features during the concatenation, without overparameterising a single-bit distinction.

4.2 BERT Encoder

All bundles use the AraBERT v2 encoder (Antoun et al., 2020). For essays exceeding 512 tokens, we apply chunking with a window size of 512 and a stride of 256. The [CLS] representations from all chunks are mean-pooled to produce a single essay vector of 768 dimensions, then compressed to 128 through a 2-layer MLP with ReLU activations, layer normalisation, and intermediate dropout (0.1). This compression brings the BERT contribution into approximate balance with the auxiliary signals (GED, hand-crafted features, prompt-type) in the fused representation.

Since bundle 0 contains prompt-aware traits, the bundle reads both the prompt and the essay: the prompt is presented in the first segment of the input, followed by the essay chunk, as `[CLS] prompt [SEP] essay [SEP]`. Bundles 1 and 2 encode only the essay since these bundles assess the linguistic features, vocabulary choices, and language accuracy, without a direct relation to the prompt. Each bundle is an independent BERT model, fine-tuned separately with its own optimiser and bundle-specific loss function.

4.3 GED Sequence Encoder

We proposed the GED augmentation method in the multi-trait scoring model for Bundles 1 and 2. For the detection of grammatical errors, we used the CAMeL-Lab GED model proposed by Alhafni et al. (2023). Instead of encoding the aggregated feature vector that contains only the error analysis results, we encode the full error-label sequence using a Bi-LSTM to preserve the distribution of errors across the essay. The GED model generates a sequence of error tags (one per token) from 13 error categories based on ARETA error annotation (Belkebir and Habash, 2021). In particular, the GED encoder transforms the per-token error label into a fixed-dimensional vector. Each per-token label is first mapped to a learned 16-dimensional embedding vector, then passed through a bidirectional LSTM (hidden size 64) to read the sequence of vectors and contextualise the error based on its position relative to the full error pattern across the essay. The LSTM outputs a 128-dimensional vector, which is pooled using a length-based binary mask. The

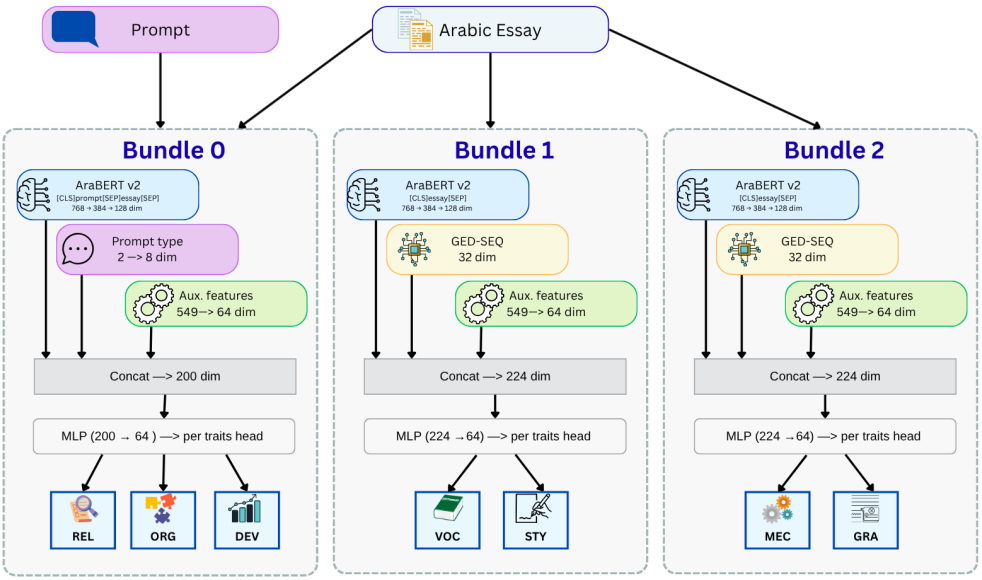

Figure 2: The proposed correlation-aware 3-bundle AraBERT architecture for multi-trait Arabic essay scoring.

final linear projection produces a 32-dimensional GED sequence representation that preserves positional and distributional patterns, such as error clustering, error-type co-occurrence, and the trajectory of writing quality throughout the essay. The GED encoder is trained end-to-end jointly with the BERT encoder, feature projection, and per-trait output heads within each bundle, using a single MSE loss computed jointly over the bundle's traits.

4.4 Feature Engineering

We employ the publicly available feature engineering model released by Sayed et al. (2025). They produced a comprehensive set of 816 engineered features for Arabic AES, covering surface-level, readability, lexical, syntactic, and semantic features. We ran the model in Laila and extracted all the features for each essay. We excluded three categories of features: (1) metadata columns, (2) target trait scores and holistic scores, and (3) prompt-specific word frequency features (top_n_* columns), which are critical for cross-prompt generalisation. The remaining 549 features are exclusively prompt-independent, such as readability indices, lexical diversity measures, and others. All three bundles receive the same 549 features. The feature vector is projected using a two-layer MLP. The first layer compresses the features into a 128-dimensional vector, while the second refines it to 64 dim.

Feature Fusion Finally, all features -the prompt, the essay, the GED seq, and the 549 features- are concatenated into a single fused vector. Specifically, the 128-dimensional BERT representation h_{BERT} is concatenated with the prompt-type embedding (8-d, Bundle 0 only), the GED sequence representation (32-d, Bundles 1 and 2 only), and the 64-dimensional hand-crafted feature vector (all bundles), as Equation 1.

$$h_{\text{fused}} = [\, h_{\text{BERT}} \,;\, h_{\text{PT}} \,;\, h_{\text{GED}} \,;\, h_{\text{Feat}} \,] \qquad (1)$$

The compression of the BERT representation to 128 dimensions brings BERT into approximate balance with the auxiliary signals in the fused vector, at a ratio of approximately 4:2:1 for BERT, hand-crafted features, and GED respectively. A dropout layer (0.1) is then applied to the fused representation, followed by a shared 64-dimensional mixing layer (with ReLU and Dropout 0.1) and a separate projection per trait. The complete fusion configuration per bundle is summarised in Table 8.

5 Prompt Topic Relevance Classification

For any AES model, and during production, any essay must pass the topic-relevance gate before being scored using a scoring framework. We build a binary classification model, following the methods proposed by Qwaider et al. (2025); however, we employ the AraBERT cross-encoder instead of the

SBERT encoder. The model is tested in two different settings (seen-prompt and unseen-prompt).

5.1 Dataset Construction

Based on the quality analysis and the subjective annotation of the REL in the LAILA, we construct a topic-relevance dataset. We generate off-topic examples by replicating the on-topic essays in LAILA and re-pairing them with wrong prompts, thereby producing a balanced dataset.

In the seen-prompt setting, the dataset follows the same cross-prompt structure as LAILA, with (5:2:3) prompts for the (train:dev:test) splits. To construct the dataset, we filter the on-topic essays, then duplicate them.After that, we follow different steps for each set as follows:

- Train set: duplicated essays are paired with a random incorrect prompt from the remaining training prompts.
- Dev set: off-topic essays use a mixture: 70% receive a wrong prompt ID from the training pool, and 30% use a within-split prompt swap. Then the DEV set has 3 distinct prompt IDs.
- Test set: since the test set only has one prompt, off-topic essays are paired with the wrong prompt ID from the training set, distinct from the development one.

This configuration ensures that some of the prompts are revealed (seen) in the dev and test sets.

In the cross-prompt setting (un-seen) for model generalisation, we use a (4:2:2) prompt distribution across the train, dev, and test sets, respectively. The off-topic essays in both dev and test sets are created by swapping prompts within each pair. This is to ensure that neither the test prompts nor the specific prompt pairings are revealed during training. The strict configuration here will be used to see the model's generalisation ability across prompt settings.

5.2 Prompt Relevance Model

We propose an AraBERT cross-encoder architecture where both the prompt and the essay form a single input ([CLS] prompt [SEP] essay_chunk [SEP]), which enables full cross attention between the prompt and each token of the essay. For the long essays, we enabled sliding-window segmentation (stride 256), so that each segment is concatenated with the full prompt. The embedding is then pooled and passed through an MLP to generate a relevance score [0,1].

The classification model will function as a filter in the assessment pipeline. If the essay receives a 0 score (off-topic), it will only receive feedback on topic adherence and will not be passed to the 6-trait AES system. If an essay is assessed as prompt-relevant, it will be checked by the AES system to assess the quality of the writing across different traits.

6 Results and Discussion

6.1 Evaluation Metrics

We used the same evaluation settings for all experiments to preserve consistency. All multi-trait models are evaluated using Leave-One-Prompt-Out (LOPO) cross-prompt validation, with the same structure as LAILA. Quadratic weighted kappa (QWK) is used as the primary metric, following most recent AES research (Cohen, 1968). We computed QWK per trait and averaged across traits, using the full range of scores [0,2] for relevance and [0,5] for other traits for the model applied to the full data set. For 6-content trait settings, we computed QWK using the score range [1, 5], excluding REL. All results across all 8 LOPO folds are reported as the mean QWK, unless otherwise stated. For prompt relevance classification in §6.3, we report macro-F1 as the primary evaluation metric, alongside QWK for consistency.

6.2 GED-Augmented Multi-Trait Scoring

To show the contribution of our proposed architecture, comparing to the benchmarks, we evaluate a series of systematic ablation configurations on LAILA, including all the traits, under LOPO cross-prompt settings, as following:

a) Baseline: a single AraBERT model with prompt-aware encoding. Predicts all 7 traits from a single shared representation and heads of multi-attention without any auxiliary signals.
b) +GED-Seq: a full BiLSTM-encoded GED-sequence (32d) is used. The model retains prompt-aware encoding and 4 bundle traits. The traits are grouped into 4 bundles: REL, ORG+DEV, STY+VOC, and MEC+GRA.
c) 3-Bundle + GED-Seq: same as before, but we add the REL to the ORG+DEV bundle, as it correlates with DEV. Hence, we end with a three-bundle configuration.
d) Full system (+ Features). This is the complete architecture proposed in §4, with all signals active.

e) Gold Relevance (+oracle). The proposed architecture with gold REL scores as input, showing the upper bound of the model when the REL prediction noise is removed.

Table 4 reports the overall QWK per trait for all configurations in addition to the two external models: LAILA (the benchmark) and TRATES (Qayyem). The results show that introducing the BiLSTM-encoded GED sequence yields the largest single-step improvement (b). The improvements on the surface and linguistic traits directly receiving the GED signal —MEC, GRA , STY— confirm that error-pattern information benefits the traits assessing language accuracy and lexical choice. The concurrent gain on REL is not attributable to GED, since bundle 0 does not receive the GED signal; rather, it reflects the architectural change of separating REL into its own prediction head, removing the interference from co-predicting REL with content traits on a mismatched score scale.

Grouping REL into the ORG+DEV bundle in (c) also shows further improvement on both REL and DEV. This confirms that co-locating REL, ORG, and DEV in a prompt-aware bundle, while assigning GED-Seq to the surface and linguistic bundles, allows each bundle to specialise without interference from dissimilar traits. The final improvement is achieved by adding the 549 engineered features to the full model. The features improved all traits, with ORG showing the largest improvement from 0.67 to 0.7, consistent with the feature set's strength in readability and coherence measures.

Our full system (d) achieves 0.62 on the 7-traits, compared to TRATES, which achieves 0.65.[4] The per-trait comparison is nuanced; our model matches TRATES in 3 traits (ORG, STY, GRA), while TRATES retains minor advantages on VOC and DEV. The biggest gap lies mainly in REL, where TRATES is better calibrated to the compressed 0–2 scale, further supporting our argument that REL requires dedicated handling rather than co-prediction with content traits. This is also seen in the oracle experiment, showing that REL prediction error not only reduces relevance QWK but also corrupts the other 6 traits. The oracle improvement confirms that corruption operates through the training data itself — off-topic essays with zero scores degrade the learned score distributions across all bundles, rather than through shared parameters. Compared to baseline LAILA, our proposed model shows an improvement in all 7-traits (0.54→0.62), demonstrating the benefit of correlation-aware bundles with task-specific auxiliary signals. The results from all the experiments conducted support the rationale for the relevance filter approach we used, in which topic relevance evaluation is dissociated from multi-trait evaluation to prevent distortion across all traits.

Relevance Filtered Subset (6 Traits) We repeat the ablation experiments on the on-topic subset, which represents the production evaluation condition. All configurations employ prompt-aware encoding for bundle 0 and implement the 3-bundle architecture, since REL is already handled by the prompt relevance classifier §5. We address the following ablation configurations:

a) Baseline: a 3-bundle architecture with prompt-aware encoding.
b) + GED-Seq. Adding the BiLSTM-encoded GED sequence to Bundles 1 and 2.
c) Full system (+ Features): the complete proposed model will activate all signals.

Table 5 reports the results per trait for the subset ablation study. The proposed complete architecture confirms that both GED and features contribute significant information. The GED-Seq (a→b) improves the mean QWK from 0.70 to 0.71, with MEC showing the expected improvement, where both MEC and GRA are grouped in bundle 2, which receives the GED signal. Adding essay features (b→c) raises the mean QWK to 0.72, where the improvement focuses on VOC and ORG. This is supported by evidence that GED can positively affect the attributes it aims to improve (MEC, GRA), hand-crafted features benefit all bundles of traits, especially (VOC and ORG), and that their combination can boost the mean QWK more than any individual method. This is evidence that each of the three sources of signal captures a different aspect of writing quality.

6.3 Topic Relevance Classification

We evaluated the topic relevance model under two settings, seen and unseen test sets. Table 6 reports the results in both settings, detailed results per fold are presented in Appendix F in Tables 11 and 12. It is clear that the model demonstrates high overall performance in both settings, with macro-F1 of 0.904 (seen) and 0.902 (unseen). The minor gap between them indicates that the classifier gener-

[4] Table 7 in Appendix D shows the model configuration parameters.

Config	REL	ORG	VOC	STY	DEV	MEC	GRA	Mean (7t)
LAILA	0.32	0.60	0.60	0.59	0.53	0.55	0.59	0.54
TRATES	**0.56**	**0.70**	**0.66**	0.66	**0.65**	0.61	0.64	**0.65**
(a) Baseline	0.20	0.66	0.59	0.63	0.56	0.58	0.61	0.55
(b) + GED-Seq	0.38	0.68	0.63	**0.67**	0.59	0.62	0.64	0.60
(c) 3-Bundles + GED-Seq	0.44	0.67	0.63	0.66	0.62	0.62	**0.65**	0.61
(d) Full system +Features	0.46	**0.70**	0.64	0.66	0.64	**0.62**	0.64	0.62
(d+oracle) Gold Rel	1.00	0.85	0.78	0.83	0.79	0.81	0.81	0.84

Table 4: GED ablation on the full dataset (7 traits, LOPO). LAILA and TRATES are external baselines.

Config	ORG	DEV	VOC	STY	MEC	GRA	Mean (6t)
(a) 3-Bundle Baseline	0.76	0.62	0.70	0.75	0.68	0.72	0.70
(b) + GED-Seq	0.74	0.64	0.70	0.75	0.71	0.72	0.71
(c) Full system (+Features)	**0.76**	**0.64**	**0.74**	**0.75**	**0.72**	**0.72**	**0.72**

Table 5: Ablation on the relevance-filtered subset (REL = 2, 6 traits).

Metric	Seen	Unseen
QWK	0.85	0.81
F1 (macro)	0.90	0.90
Accuracy	0.92	0.91
Precision (off-topic)	0.96	0.96
Recall (off-topic)	0.86	0.85
F1 (off-topic)	0.86	0.89
Precision (on-topic)	0.92	0.89
Recall (on-topic)	0.99	0.96
F1 (on-topic)	0.94	0.92

Table 6: Topic relevance classification: seen vs. unseen settings (mean across folds).

alises well across prompts rather than memorising prompt-specific representations. We observe that off-topic precision is high across both classes and both settings. The off-topic class precision is 0.96 in both, indicating that the model correctly identifies off-topic essays. The on-topic precision is 0.92 (seen) and 0.89 (unseen), showing that most unfiltered essays are on-topic.

On the other hand, recall is more nuanced, and is more critical for any deployed system, as it determines which essays should pass the gate and receive the multi-trait score and which should be rejected. In the on-topic classes, recall is near perfect (0.99, 0.96) in seen and unseen settings, respectively. This indicates that relevant essays are always detected and passed to the scoring model. However, off-topic recall is lower (0.86 seen, 0.85 unseen), showing that roughly 15% of off-topic essays pass through the filter and get scored. The problematic cases are discussed in Appendix F. Excluding them, the model achieves recall of 0.98 (seen) and 0.92 (unseen), confirming that the AraBERT architecture detects mismatches in topic relevance when the off-topic boundary is discriminable.

7 Conclusion and Future work

We conduct a comprehensive investigation of multi-trait Arabic AES, including data quality, scoring strategy, correlation-aware bundles, model architecture, and prompt relevance classification. From our analysis of LAILA, we identify REL as a distinct trait that corrupts content-trait learning. Moreover, we reveal three stable correlation bundles among the six content traits that motivate our architectural design. Our proposed GED-augmented, correlation-aware bundle model achieves 0.72 QWK on the REL-filtered subset (6 traits), supporting the AES system's pipeline. The model also reaches 0.62 on the full dataset (7 traits) under cross-prompt evaluation. The proposed model used tailored auxiliary features for each bundle to support the heterogeneous nature of the writing traits. The binary relevance classifier (F1 = 0.90) completes a fully automated pipeline from relevance prediction to trait scoring.

In Future, we will work in several directions. First, the 3-bundle architecture requires three AraBERT instances We plan to investigate other parameter-efficient alternatives with bundle-specific adapters that share a single backbone to reduce computational cost while preserving specialisation. Second, enhancing the multi-trait scoring system by providing rationales for each score and feedback to improve writing quality. Third, investigating confidence scores for the topic relevance model that can be applied to on-topic, off-topic and partially relevant essays.

Limitations

Our work has several limitations. First, the computational cost behind the 3-bundle architecture, which requires three separate AraBERT instances. Second, the production experiments are conducted on a subset of the LAILA dataset (on-topic essays with six traits), which limits direct comparability with systems evaluated on the full dataset. Third, the GED-Seq signal depends on the CAMeL-Lab GED model, meaning that its errors may propagate into the scoring architecture. Finally, the topic relevance classifier is trained and evaluated on synthetic off-topic essays generated by prompt swapping, which may not fully simulate the real-world off-topic scenarios.

Ethical Considerations

Our work uses the publicly available LAILA dataset, which contains anonymised essays with no personally identifiable information. The essays cover standard educational prompts and are unlikely to contain offensive content; we did not encounter any during data analysis. In any deployment settings where AI authorship detection is required, a separate module would be needed.

Generally speaking, this work uses the LAILA dataset (CC BY-NC-ND 4.0), AraBERT v2 (custom permissive license, AUB MIND Lab), the CAMeL-Lab Arabic GED model (MIT License), and the publicly available feature set. All artifacts are used for non-commercial academic research, consistent with their intended use.

References

Bashar Alhafni, Go Inoue, Christian Khairallah, and Nizar Habash. 2023. Advancements in arabic grammatical error detection and correction: An empirical investigation. In *Proceedings of the 2023 Conference on empirical Methods in natural language processing*, pages 6430–6448.

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.

May Bashendy, Salam Albatarni, Sohaila Eltanbouly, Walid Massoud, Houda Bouamor, and Tamer Elsayed. 2025. Taqeem 2025: Overview of the first shared task for arabic quality evaluation of essays in multi-dimensions. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 966–976.

May Bashendy, Salam Albatarni, Sohaila Eltanbouly, Eman Zahran, Hamdo Elhuseyin, Tamer Elsayed, Walid Massoud, and Houda Bouamor. 2024. Qaes: First publicly-available trait-specific annotations for automated scoring of arabic essays. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 337–351.

May Bashendy, Walid Massoud, Sohaila Eltanbouly, Salam Albatarni, Marwan Sayed, Abrar Abir, Houda Bouamor, and Tamer Elsayed. 2026. Laila: A large trait-based dataset for arabic automated essay scoring. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3074–3091.

Riadh Belkebir and Nizar Habash. 2021. Automatic error type annotation for arabic. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 596–606.

Yuan Chen and Xia Li. 2023. PMAES: Prompt-mapping contrastive learning for cross-prompt automated essay scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1503, Toronto, Canada. Association for Computational Linguistics.

Minsoo Cho, Jin-Xia Huang, and Oh-Woog Kwon. 2024. Dual-scale bert using multi-trait representations for holistic and trait-specific essay grading. *ETRI Journal*, 46(1):82–95.

SeongYeub Chu, Jong Woo Kim, Bryan Wong, and Mun Yong Yi. 2025. Rationale behind essay scores: Enhancing s-llm's multi-trait essay scoring with rationale generated by llms. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5796–5814.

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

Heejin Do, Yunsu Kim, and Gary Lee. 2024. Autoregressive score generation for multi-trait essay scoring. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1659–1666.

Heejin Do, Yunsu Kim, and Gary Geunbae Lee. 2023. Prompt-and trait relation-aware cross-prompt essay trait scoring. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1538–1551.

Heejin Do, Taehee Park, Sangwon Ryu, and Gary Lee. 2025. Towards prompt generalization: Grammar-aware cross-prompt automated essay scoring. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2818–2824, Albuquerque, New Mexico. Association for Computational Linguistics.

Heejin Do, Sangwon Ryu, and Gary Geunbae Lee. 2026. Teach-to-reason with scoring: Self-explainable rationale-driven multi-trait essay scoring. *Expert Systems with Applications*, page 132119.

Hoor Elbahnasawi, Marwan Sayed, Sohaila Eltanbouly, Fatima Brahamia, and Tamer Elsayed. 2026. Qayyem: A real-time platform for scoring proficiency of arabic essays. *arXiv preprint arXiv:2603.01009*.

Sohaila Eltanbouly, Salam Albatarni, and Tamer Elsayed. 2025. Trates: Trait-specific rubric-assisted cross-prompt essay scoring. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20528–20543.

Rayed Ghazawi and Edwin Simpson. 2024. Automated essay scoring in arabic: a dataset and analysis of a bert-based system. *arXiv preprint arXiv:2407.11212*.

Mohamed Abdellatif Hussein, Hesham Hassan, and Mohammad Nassef. 2019. Automated language essay scoring systems: A literature review. *PeerJ Computer Science*, 5:e208.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. In *IJCAI*, volume 19, pages 6300–6308.

Paraskevas Lagakis and Stavros Demetriadis. 2021. Automated essay scoring: A review of the field. In *2021 International Conference on Computer, Information and Telecommunication Systems (CITS)*, pages 1–6. IEEE.

Shengjie Li and Vincent Ng. 2024. Automated essay scoring: A reflection on the state of the art. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17876–17888.

Chatrine Qwaider, Kirill Chirkunov, Bashar Alhafni, Nizar Habash, and Ted Briscoe. 2025. Evaluating prompt relevance in Arabic automatic essay scoring: Insights from synthetic and real-world data. In *Proceedings of The Third Arabic Natural Language Processing Conference*, pages 162–178, Suzhou, China. Association for Computational Linguistics.

Robert Ridley, Liang He, Xin-yu Dai, Shujian Huang, and Jiajun Chen. 2021. Automated cross-prompt scoring of essay traits. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13745–13753.

Robert Ridley, Liang He, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2020. Prompt agnostic essay scorer: a domain generalization approach to cross-prompt automated essay scoring. *arXiv preprint arXiv:2008.01441*.

Marwan Sayed, Sohaila Eltanbouly, May Bashendy, and Tamer Elsayed. 2025. Feature engineering is not dead: A step towards state of the art for Arabic automated essay scoring. In *Proceedings of The Third Arabic Natural Language Processing Conference*, pages 231–245, Suzhou, China. Association for Computational Linguistics.

Masaki Uto. 2021. A review of deep-neural automated essay scoring models. *Behaviormetrika*, 48(2):459–484.

Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1560–1569.

A Trait Distributions

Figure 3 plots the score distributions across the 7 traits, as well as the holistic score.

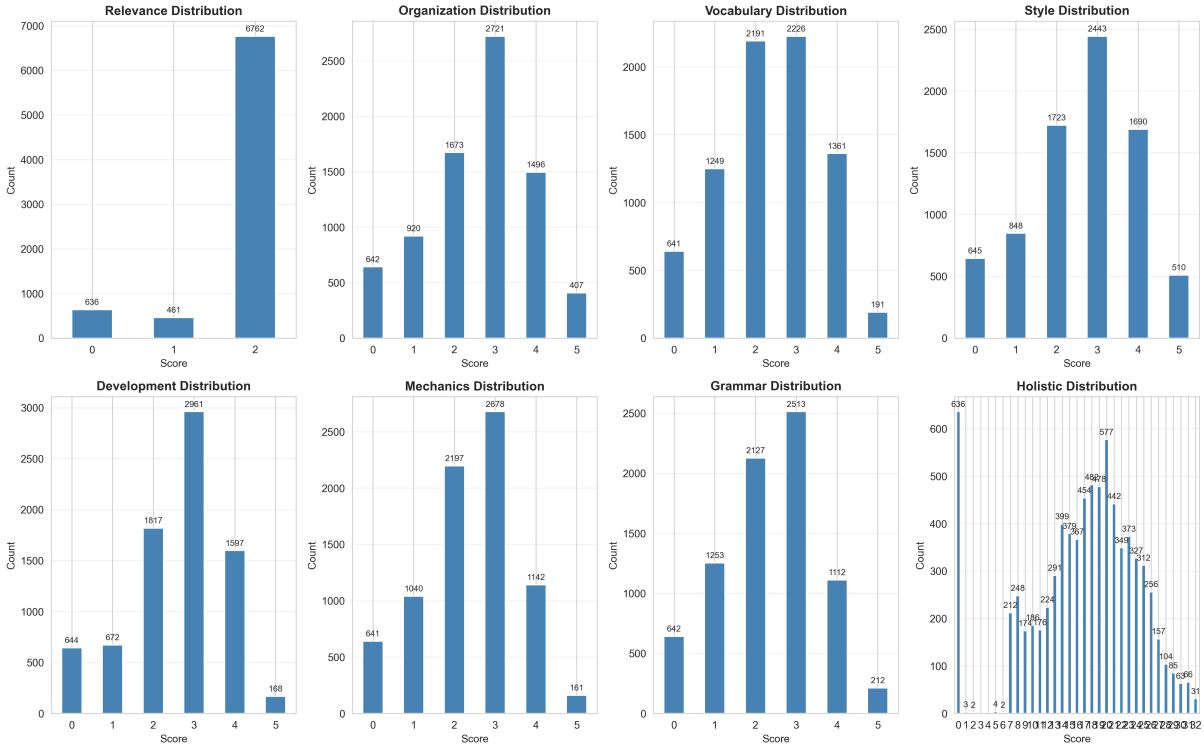

Figure 3: Trait distributions

B Traits Correlation Analysis

B.1 The holistic score presents a near-perfect correlation.

It correlates with each content trait at r = 0.93–0.96. These numbers align with the annotation protocol explained in the LAILA guidelines. The holistic correlation indicates that higher trait scores generally lead to higher holistic scores and lower trait scores lead to lower holistic scores. Although the holistic score is calculated as the sum of the seven traits, it still reflects the same pattern of dependency among them. In other words, a student who performs well in one trait is likely to perform well in other traits, resulting in a higher overall holistic score, and vice versa.

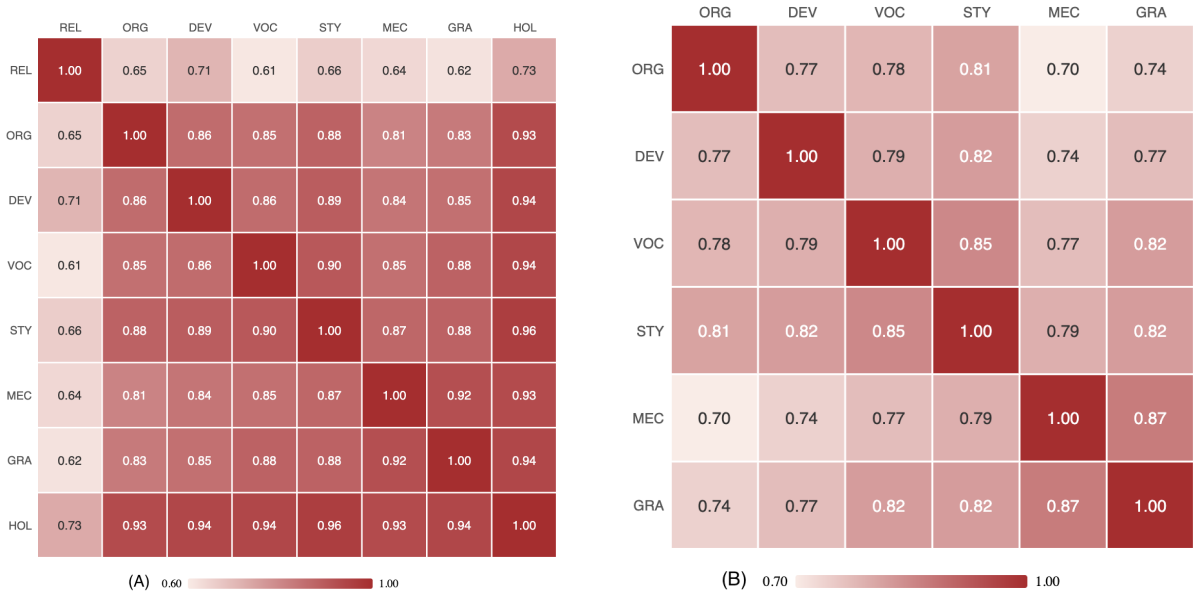

Figure 4: A: Pearson correlation matrix across all 7 traits and the holistic score (N = 7859). B: Pearson correlation between the 6 content traits for essays with relevance ≥ 1 (N = 7223).

B.2 Relevance as the weakest link.

Figure 4 clearly shows that the REL trait has the weakest correlation with any other trait, with Pearson's r ranging from 0.6 to 0.7. However, all six content traits are strongly correlated with one another, ranging from 0.83 to 0.92. These correlation results confirm that relevance assesses a different dimension of linguistic quality than the other 6 content traits. REL is a categorical score, not a ranking score. Including relevance in a joint model with other traits will degrade the model's overall performance, as we show in Section §4.

C Ambiguous Annotation Boundaries

The annotation guidebook instructs annotators to score REL based on prompt\topic adherence and surface quality, and to score copied or AI-generated text "as written". However, we observe some inconsistencies in the REL annotation, where the majority of the essays score 0 if the essays are copied or generated by AI. This inconsistency creates confusion, where a well-written text could get 0's in all other traits based on the 0 REL. In addition, the guidebook acknowledges that the 1-vs-2 boundary is the most subjective. An essay that addresses the prompt but discusses a different aspect of a topic may be scored 1 or 2, depending on the annotator.

The annotation guidelines in Figure 5 state the following:

Dealing with Copied or AI Generated Texts

During the scoring process, the evaluator may encounter texts that appear to be partially or entirely copied from external sources such as the internet or AI tools.

Here, we emphasize an important point:

In the context of this project, and for the purpose of training the automated scoring system only, the evaluator should score the text exactly as it is presented, according to the scoring criteria only, without considering whether it is copied or suspicious in terms of its source.

This guidance is based on the fact that verifying the originality of texts or detecting cases of plagiarism is not part of the evaluator's role at this stage. Instead, such cases are handled later through specialised automated tools (such as plagiarism detection software or AI-supported linguistic analysis). Therefore, the evaluator is expected to:

- *Not to reject any text merely because it is suspected to be copied.*
- *Assess the performance as if the student were the actual author.*
- *Focus only on the quality of the text in terms of ideas, organisation, vocabulary, structure, and style, and assign the appropriate score according to the approved criteria.*

This approach ensures fairness in scoring and helps train the automated model to process all types of texts as they are actually written in real educational settings.

5. التعامل مع النصوص المنقولة أو المولدة

قد يواجه المصحح أثناء عملية التقدير بعض النصوص التي تبدو منقولة جزئيًا أو كليًا من مصادر خارجية مثل الإنترنت أو أدوات الذكاء الاصطناعي.

وهنا نؤكد على نقطة مهمة:

في سياق هذا المشروع، ولأغراض تدريب نظام التصحيح الآلي فقط، ينبغي تصحيح النص كما هو مكتوب أمامك، ووفقًا لمعايير التقدير فقط، دون الالتفات إلى كونه منقولًا أو مشكوكًا في مصدره.

ويُعزى هذا التوجيه إلى أن التحقق من أصل النصوص أو الكشف عن حالات النسخ واللصق ليس من مهام المصحح في هذه المرحلة، بل يتم التعامل معها لاحقًا من خلال **آليات إلكترونية متخصصة** (مثل برامج الكشف عن الانتحال أو التحليل اللغوي المدعوم بالذكاء الاصطناعي). لذا، المطلوب من المصحح هو:

- **عدم رفض أي نص** بسبب الشك في كونه منقولًا.
- **تقدير الأداء كما لو أن الطالب هو كاتبه الحقيقي.**
- **التركيز فقط على جودة النص من حيث الأفكار، التنظيم، المفردات، التراكيب، والأسلوب**، ومنح الدرجة المناسبة وفق المعايير المعتمدة.

هذا النهج يضمن عدالة التقدير، ويُساعد على تدريب النموذج الآلي على معالجة النصوص بجميع أنواعها كما تُكتب فعليًا في البيئات التعليمية الواقعية.

Figure 5: Annotatioan guidleines

C.1 Empirical Impact on Model Performance

The relevance problem has direct consequences for multi-trait scoring model. When an essay gets REL 0, all the other traits must be assigned 0 score and the essay should not go through the assessment process, since the essay is off-topic. Any multi-trait model trained on such data learns a relation between low relevance and low trait score which contaminates the model ability fo predict the content traits independently. Due to the scale asymetry of the traits, using the relevance in calculating the mean QWK which is not statistically comparable to the content traits, increases the difficulty of the task and unfairly degrade the systems.

D Model Configuration

Hyperparameter	Value
Model	AraBERT v2 (aubmindlab/bert-base-arabertv2)
Optimizer	AdamW
Learning rate	2e-5
Weight decay	0.01
LR schedule	Linear warmup (10%) + decay
Max gradient norm	1.0
Batch size	8
Max epochs	10
Max token length	512
Sliding window stride	256
Dropout	0.1
Frozen BERT layers	0 (all layers fine-tuned)
GPU	NVIDIA RTX 6000 Ada (48GB)
Total parameters	$\sim$330M (3 $\times$ $\sim$110M AraBERT)
GED-Seq model	camelbert-msa-qalb14-ged-13

Table 7: Training configuration for the 3-bundle architecture.

Bundle	Traits	Components	Encoding	Dim
0	REL, ORG, DEV	BERT (128) + P-type (8) + Feat (64)	Prompt-aware	200
1	STY, VOC	BERT (128) + GED-Seq (32) + Feat (64)	Essay-only	224
2	MEC, GRA	BERT (128) + GED-Seq (32) + Feat (64)	Essay-only	224

Table 8: Bundle configuration: trait assignment, signal components, encoding type, and fused dimension.

D.1 Implementation Details

All experiments are implemented in PyTorch using the HuggingFace Transformers library for AraBERT v2 (aubmindlab/bert-base-arabertv2). The CAMeL-Lab Arabic GED model (camelbert-msa-qalb14-ged-13) is used for grammatical error tagging. The BiLSTM GED-sequence encoder is implemented using `torch.nn.LSTM`. Evaluation metrics are computed using `scikit-learn`: QWK with `sklearn.metrics.cohen_kappa_score(weights='quadratic')` and macro-F1 with `sklearn.metrics.f1_score(average='macro')`. Pearson correlations are computed using `scipy.stats.pearsonr`.

E GED-Augmented Multi-Trait Scoring

E.1 Full Dataset

Experiment	Prompt	REL	ORG	DEV	VOC	STY	MEC	GRA	Mean
(a) Baseline	P1	0.23	0.53	0.50	0.49	0.52	0.47	0.47	0.46
	P2	0.25	0.71	0.60	0.62	0.67	0.61	0.63	0.59
	P3	0.09	0.78	0.67	0.69	0.80	0.75	0.75	0.65
	P4	0.46	0.71	0.60	0.55	0.67	0.66	0.70	0.62
	P5	0.03	0.61	0.51	0.53	0.55	0.44	0.52	0.46
	P6	−0.01	0.50	0.44	0.48	0.50	0.42	0.48	0.40
	P7	0.56	0.76	0.60	0.72	0.70	0.66	0.70	0.67
	P8	0.01	0.68	0.57	0.66	0.64	0.60	0.62	0.54
	AVG	**0.20**	**0.66**	**0.56**	**0.59**	**0.63**	**0.58**	**0.61**	**0.55**
(b) + GED-Seq	P1	0.23	0.50	0.39	0.45	0.46	0.41	0.44	0.41
	P2	0.31	0.74	0.64	0.64	0.72	0.66	0.68	0.63
	P3	0.40	0.73	0.70	0.66	0.78	0.75	0.72	0.68
	P4	0.61	0.71	0.58	0.62	0.71	0.64	0.64	0.65
	P5	0.13	0.64	0.57	0.56	0.63	0.55	0.59	0.52
	P6	0.37	0.71	0.65	0.69	0.72	0.67	0.73	0.65
	P7	0.70	0.75	0.64	0.72	0.66	0.67	0.67	0.69
	P8	0.29	0.68	0.55	0.74	0.69	0.64	0.66	0.61
	AVG	**0.38**	**0.68**	**0.59**	**0.63**	**0.67**	**0.62**	**0.64**	**0.60**
(c) 3-Bundles + GED-Seq	P1	0.29	0.57	0.52	0.46	0.45	0.42	0.43	0.45
	P2	0.36	0.72	0.68	0.59	0.70	0.63	0.69	0.63
	P3	0.32	0.68	0.67	0.68	0.76	0.76	0.76	0.66
	P4	0.60	0.82	0.73	0.59	0.67	0.67	0.65	0.68
	P5	0.23	0.50	0.53	0.58	0.58	0.54	0.58	0.51
	P6	0.25	0.54	0.47	0.70	0.71	0.65	0.73	0.58
	P7	0.63	0.74	0.61	0.72	0.70	0.66	0.69	0.68
	P8	0.81	0.81	0.73	0.74	0.67	0.65	0.66	0.72
	AVG	**0.44**	**0.67**	**0.62**	**0.63**	**0.66**	**0.62**	**0.65**	**0.61**
(d) Full system	P1	0.37	0.58	0.51	0.44	0.44	0.42	0.42	0.45
	P2	0.26	0.66	0.60	0.61	0.72	0.69	0.70	0.60
	P3	0.58	0.75	0.69	0.68	0.76	0.75	0.72	0.70
	P4	0.70	0.82	0.76	0.59	0.64	0.63	0.64	0.68
	P5	0.33	0.60	0.48	0.55	0.61	0.55	0.57	0.53
	P6	0.27	0.57	0.59	0.71	0.71	0.66	0.73	0.61
	P7	0.79	0.84	0.79	0.75	0.71	0.67	0.69	0.75
	P8	0.37	0.79	0.70	0.74	0.68	0.65	0.64	0.65
	AVG	**0.46**	**0.70**	**0.64**	**0.64**	**0.66**	**0.62**	**0.64**	**0.62**
(d+oracle) Gold Rel	P1	1.00	0.85	0.81	0.80	0.82	0.80	0.80	0.84
	P2	1.00	0.85	0.81	0.75	0.84	0.80	0.80	0.84
	P3	1.00	0.85	0.77	0.74	0.86	0.81	0.80	0.83
	P4	1.00	0.88	0.81	0.74	0.86	0.86	0.86	0.86
	P5	1.00	0.83	0.79	0.74	0.79	0.78	0.77	0.82
	P6	1.00	0.80	0.81	0.82	0.85	0.79	0.84	0.84
	P7	1.00	0.84	0.72	0.81	0.79	0.75	0.79	0.81
	P8	1.00	0.87	0.81	0.86	0.86	0.84	0.83	0.87
	AVG	**1.00**	**0.85**	**0.79**	**0.78**	**0.83**	**0.81**	**0.81**	**0.84**

Table 9: GED ablation: per-prompt QWK on the full dataset (7 traits). (a) Baseline: single AraBERT. (b) + GED-Seq. (c) 3-bundle restructuring. (d) Full system with features. (d+oracle) Gold relevance upper bound.

E.2 On-topic Dataset

Experiment	Prompt	ORG	DEV	VOC	STY	MEC	GRA	Mean
	P1	0.69	0.50	0.64	0.72	0.58	0.59	0.62
	P2	0.72	0.59	0.52	0.73	0.65	0.70	0.65
	P3	0.78	0.70	0.68	0.82	0.75	0.81	0.76
	P4	0.82	0.65	0.72	0.73	0.77	0.73	0.74
(a) 3-Bundle Baseline	P5	0.74	0.62	0.70	0.71	0.60	0.68	0.68
	P6	0.77	0.66	0.78	0.78	0.67	0.79	0.74
	P7	0.76	0.64	0.76	0.72	0.70	0.71	0.71
	P8	0.79	0.57	0.78	0.76	0.74	0.72	0.73
	AVG	**0.76**	**0.62**	**0.70**	**0.75**	**0.68**	**0.72**	**0.70**
	P1	0.70	0.60	0.69	0.72	0.59	0.60	0.65
	P2	0.65	0.45	0.58	0.74	0.67	0.71	0.63
	P3	0.78	0.69	0.65	0.80	0.80	0.82	0.76
	P4	0.82	0.72	0.68	0.75	0.79	0.79	0.76
(b) + GED-Seq	P5	0.71	0.69	0.62	0.70	0.62	0.65	0.67
	P6	0.77	0.71	0.78	0.75	0.73	0.77	0.75
	P7	0.76	0.63	0.75	0.74	0.70	0.73	0.72
	P8	0.75	0.62	0.84	0.79	0.74	0.72	0.74
	AVG	**0.74**	**0.64**	**0.70**	**0.75**	**0.71**	**0.72**	**0.71**
	P1	0.72	0.62	0.66	0.72	0.61	0.57	0.65
	P2	0.71	0.48	0.73	0.72	0.71	0.74	0.68
	P3	0.78	0.69	0.72	0.81	0.77	0.75	0.75
	P4	0.84	0.72	0.70	0.77	0.81	0.79	0.77
(c) Full model	P5	0.69	0.62	0.69	0.68	0.65	0.69	0.67
	P6	0.78	0.72	0.76	0.77	0.73	0.79	0.76
	P7	0.76	0.66	0.78	0.74	0.71	0.73	0.73
	P8	0.77	0.62	0.84	0.77	0.74	0.72	0.74
	AVG	**0.76**	**0.64**	**0.74**	**0.75**	**0.72**	**0.72**	**0.72**

Table 10: Ablation on the relevance-filtered subset (REL = 2, 6 traits). (a) 3-Bundle baseline: no GED, no features. (b) + GED-Seq to Bundles 1–2. (c) Full model with 549 features.

F Topic Relevance Classification

The two problematic cases, Fold 2 and Fold 8, in seen and unseen settings, respectively, share the same root cause: the thematic overlap between the prompt pairs used to generate the off-topic examples. Fold 2, has two prompts (P2: "social media as a threat to family relationships" and P4: "phones reducing face-to-face communication"), where both address technology' impact on relationships. Moreover, in Fold 8 we can see that prompt 6 and prompt 7 talk about ("reducing homework", and "young people staying up late, that affect their academic performance"), respectively, so both prompts also shared the theme of study life and academic impact. In both cases, the model classifies all essays as on-topic since the test set contains a related prompt-essay pair. The low performance lies in the evaluation rather than the model's understanding of relevance. Excluding these two folds, the model achieves an R of 0.98 (seen) and 0.92 (unseen), confirming that the AraBERT architecture detects mismatches in topic relevance when the off-topic boundary is unambiguous.

Fold	QWK	F1 (macro)	P (off)	R (off)	P (on)	R (on)	F1 (off)	F1 (on)	Acc
F1	0.92	0.96	0.99	0.93	0.93	0.99	0.96	0.96	0.96
F2	0.01	0.35	0.77	0.02	0.50	1.00	0.03	0.67	0.51
F3	0.91	0.95	0.98	0.93	0.94	0.98	0.95	0.96	0.95
F4	0.98	0.99	0.99	1.00	1.00	0.99	0.99	0.99	0.99
F5	0.99	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00
F6	0.99	1.00	0.99	1.00	1.00	0.99	1.00	1.00	1.00
F7	0.99	1.00	0.99	1.00	1.00	0.99	1.00	0.99	1.00
F8	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Mean	**0.85**	**0.90**	**0.96**	**0.86**	**0.92**	**0.99**	**0.86**	**0.94**	**0.92**

Table 11: Topic relevance classification — **Seen** setting. P = Precision, R = Recall, off = off-topic, on = on-topic.

Fold	QWK	F1 (macro)	P (off)	R (off)	P (on)	R (on)	F1 (off)	F1 (on)	Acc
F1	0.71	0.85	0.99	0.72	0.78	0.99	0.83	0.87	0.86
F2	0.71	0.85	0.84	0.88	0.87	0.83	0.86	0.85	0.85
F3	0.95	0.98	0.96	0.99	0.99	0.96	0.98	0.98	0.98
F4	0.94	0.97	0.98	0.97	0.97	0.98	0.97	0.97	0.97
F5	0.87	0.93	0.99	0.88	0.89	0.99	0.93	0.94	0.93
F6	0.95	0.97	0.95	1.00	1.00	0.95	0.97	0.97	0.97
F7	0.98	0.99	0.98	1.00	1.00	0.98	0.99	0.99	0.99
F8	0.38	0.66	0.99	0.38	0.62	1.00	0.55	0.76	0.69
Mean	**0.81**	**0.90**	**0.96**	**0.85**	**0.89**	**0.96**	**0.89**	**0.92**	**0.91**

Table 12: Topic relevance classification — **Unseen** setting. P = Precision, R = Recall, off = off-topic, on = on-topic.