QAES: First Publicly-Available Trait-Specific Annotations for Automated Scoring of Arabic Essays

May Bashendy
Qatar University
ma1403845@qu.edu.qa

Salam Albatarni
Qatar University
sa1800633@qu.edu.qa

Sohaila Eltanbouly
Qatar University
se1403101@qu.edu.qa

Eman Zahran
English Modern School
eman.zahran@emsdoha.net

Hamdo Elhuseyin
English Modern School
hamdo.huseyin@emsdoha.net

Tamer Elsayed
Qatar University
telsayed@qu.edu.qa

Walid Massoud
Qatar University
wmassoud@qu.edu.qa

Houda Bouamor
Carnegie Mellon University in Qatar
hbouamor@cmu.edu

Abstract

Automated Essay Scoring (AES) has emerged as a significant research problem within natural language processing, providing valuable support for educators in assessing student writing skills. In this paper, we introduce *QAES*, the first publicly available trait-specific annotations for Arabic AES, built on the Qatari Corpus of Argumentative Writing (QCAW). *QAES* includes a diverse collection of essays in Arabic, each of them annotated with holistic and trait-specific scores, including relevance, organization, vocabulary, style, development, mechanics, and grammar. In total, it comprises 195 Arabic essays (with lengths ranging from 239 to 806 words) across two distinct argumentative writing tasks. We benchmark our dataset against the state-of-the-art English baselines and a feature-based approach. In addition, we discuss the adopted guidelines and the challenges encountered during the annotation process. Finally, we provide insights into potential areas for improvement and future directions in Arabic AES research.

1 Introduction

Automated Essay Scoring (AES) is used to automatically evaluate essays, eliminating the need for human intervention. AES has gained significant importance in educational assessment, offering an efficient way to evaluate written essays promptly. Traditionally, AES focused on assigning a single holistic score to an essay, reflecting its overall quality (Xie et al., 2022; Yang et al., 2020). This approach simplifies the evaluation by summarizing the performance of the essay. However, relying solely on a uni-dimensional score is insufficient for guiding students on how to improve areas of weakness. Consequently, recent research has shifted towards trait-specific scoring, which assigns scores to distinct traits of the essay (Kumar et al., 2022; Ormerod, 2022). By assessing each trait separately, educators and students gain deeper insight into areas for improvement. Although significant progress has been made in AES for languages such as English (Klebanov and Madnani, 2022), the assessment of Arabic essays remains understudied. This is primarily due to the lack of publicly available annotated datasets tailored for Arabic essay scoring.

To address this challenge, we introduce the QCAW annotations for Automated Essay Scoring (*QAES*),[1] the first freely available annotated Arabic essay corpus for trait-specific scoring. We build *QAES* by annotating the Qatari Corpus of Argumentative Writing (QCAW) (Ahmed et al., 2024) across seven traits: Relevance, Organization, Vocabulary, Style, Development, Mechanics, and Grammar. We release these annotations to the research community, aiming to foster advancements in Arabic AES.

We also evaluate the performance of feature-based and state-of-the-art English baselines on the dataset. Our efforts help bridge the resource gap and empower educators and researchers with valuable insights into the nuances of Arabic writing.

The contributions of this paper are four-fold: (1) Annotating the Arabic essays of the QCAW dataset, providing both holistic and trait-specific scores, (2) Releasing *QAES*,[2] the **first** publicly-available trait-specific annotations for Arabic AES, (3) Providing insights into the annotation challenges and areas for improvement, and (4) Evaluating the performance of feature-based and state-of-the-art baselines on *QAES*.

[1] Pronounced in Arabic as "قيس" (as in "قيس بن الملوح").

[2] https://sites.google.com/view/bigir/datasets

Proceedings of The Second Arabic Natural Language Processing Conference, pages 337–351

The rest of the paper is organized as follows: Section 2 reviews related work on Arabic essay datasets. Section 3 presents the QCAW dataset considered for annotating. Section 4 outlines the annotation process. Section 5 provides a detailed analysis of the annotations. Section 6 discusses the experimental setup and results. Finally, Section 7 concludes with future directions.

2 Related Work

In Automated Essay Scoring (AES), a significant contrast exists between English and Arabic datasets. While English AES benefits from well-established large-scale benchmark datasets, such as the Automated Student Assessment Prize (ASAP)[3] dataset, there is a lack of publicly available annotated datasets for Arabic AES. This could be attributed to the limited research focus on the Arabic language, coupled with its high ambiguity, rich morphology, and complex morpho-syntactic rules. Despite these challenges, some initiatives have developed modest Arabic essay corpora for preliminary studies. These datasets, summarized in Table 1, although limited, are essential for advancing Arabic AES.

Publicly-Available Arabic Essay Datasets When exploring Arabic AES datasets, three stand out: (1) The Zayed Arabic English Bilingual Undergraduate Corpus (ZAEBUC) (Habash and Palfreyman, 2022), consisting of essays by first-year university students in the UAE, enriched with linguistic annotations such as POS tagging, grammar and spelling corrections, and lemmatization; (2) The Arabic Learner Corpus (ALC),[4] a collection of Arabic texts written by non-native learners of Arabic in Saudi Arabia also annotated with linguistic features, yet, they are not publicly disclosed; and (3) The Qatari Corpus of Argumentative Writing (QCAW) (Ahmed et al., 2024), containing long essays with publicly available POS tags but non-public holistic score annotations. While these datasets provide valuable linguistic insights, they lack publicly available holistic or trait-specific content quality annotations, limiting their usability in developing AES systems.

Non-public Arabic Essay Datasets Several datasets are not publicly available for research. For example the "Abbir" dataset (Alghamdi et al., 2014), comprising essays from college students in Saudi Arabia and including holistic scores ranging from 1 to 6 (best score). Another example is the "AAEE" dataset (Azmi et al., 2019), which scores essays written by students in grades 7 to 12, based on criteria such as semantic analysis, writing style, and spelling accuracy.

Dataset	Essays	Len	Tasks	Public	HOL	Traits
ZAEBUC	214	156	3	✓	×	×
ALC	1,585	178	2	✓	×	×
QCAW	195	499	2	✓[5]	✓	×
Abbir	640	150	2	×	✓	×
AAEE	350	-	8	×	✓	×
QAES	195	489	2	✓	✓	✓

Table 1: Comparing *QAES* with existing Arabic essay datasets. 'Len' denotes average essay length in words, and 'HOL' refers to holistic scoring.

Short Answer Datasets Several short answer datasets have been collected for Arabic AES. For instance, Abdeljaber (2021) introduced an accessible dataset with 330 answers for 10 questions. Similarly, Ouahrani and Bennouar (2020) presents the freely available AR-ASAG dataset, which includes 2,133 student answers to 48 questions from a cybercrimes course. Other datasets are not publicly available, such as the "eJaya-NN" dataset (Gaheen et al., 2020) with 240 answers to one quiz question, and the "Philosophy" dataset (Gomaa, 2014), an expansion of their initial Arabic short answer benchmark set (Gomaa and Fahmy, 2014) containing 600 responses to 50 questions scored from 0 to 10. Shehab et al. (2018) collected 210 short answers from secondary students in a sociology course. Additionally, Nael et al. (2022) translated the English ASAP Short Answer Scoring dataset into Arabic using Google Translate. Several other datasets exist, but they are typically close-domain and small-scale with limited applicability.

The absence of publicly available annotations for Arabic essay scoring obstructs the development and validation of AES models tailored to Arabic essays. Most existing Arabic essay datasets are small, often confined to specific domains, very limited in terms of prompts, primarily focus on short answers, and only consider holistic scoring, oversimplifying the multifaceted nature of writing assessment. Furthermore, these datasets are often proprietary or not publicly accessible, hindering reproducibility and collaboration. In response to

[3] https://www.kaggle.com/c/asap-aes

[4] https://www.arabiclearnercorpus.com

[5] The essays are publicly available, but not the annotations.

Task	Essays	Words/Essay	Tokens
1	115	500~[239-806]	57,486
2	80	473~[249-607]	37,856

Table 2: Descriptive statistics of the QCAW dataset.

this challenge, we extend the Qatari Corpus of Argumentative Writing (QCAW) to present *QAES*, the first public trait-specific annotations for Arabic AES. Augmenting the QCAW dataset with adequate annotations would unlock its full potential and drive the evolution of AES in Arabic, echoing the advancements seen in English AES.

3 The QCAW Corpus

The Qatari Corpus of Argumentative Writing (QCAW) (Ahmed et al., 2024) is a publicly available corpus comprising 195 *argumentative* essays written in both Arabic and English by first-year Arabic native university students as part of a compulsory university-level course. It provides a valuable resource for AES as its lengthy essays offer insights into Arabic linguistic styles and different writing proficiency levels. Furthermore, the essays were collected from a diverse student sample, considering factors such as gender, year of study, and major. In our work, we only consider the Arabic essays. These essays constitute responses to two distinct argumentative writing tasks, originally derived from TOEFL writing prompts: Task 1: *'Telephones and emails have made communication between people less personal"*; and Task 2: *'Technology has enabled students nowadays to learn more information quickly"*. Table 2 provides a breakdown of the tasks featured in the QCAW dataset, showing a skew towards Task 1 in terms of both the number of essays and their length. On average, after tokenizing with the NLTK tokenizer and removing punctuation, essays are approximately 489 words long, ranging from 239 to 806 words. The total number of tokens across all essays is 95,342.

4 Constructing *QAES*

To create *QAES*, annotators were hired to annotate the Arabic essays of the QCAW dataset across seven traits: Relevance (REL), Organization (ORG), Vocabulary (VOC), Style (STY), Development (DEV), Mechanics (MEC), and Grammar (GRA) in addition to a Holistic (HOL) score of overall quality. The process entails the deployment of trusted assessment guidelines/rubrics, the selection of specialist annotators, and the delivery of in-depth training sessions to ensure consistency and understanding throughout the annotation process.

4.1 Annotation Guidelines

To score student responses, we provided the annotators with the rubrics used in the Core Academic Skills Test (CAST) developed by the Qatar University Testing Center (QUTC).[6] These rubrics were designed to assess students' ability to write persuasive/argumentative essays, which matches perfectly with the tasks covered in the QCAW corpus. Additionally, CAST's rubrics underwent extensive measures to ensure that they were aligned with the guidelines established by subject matter experts and reviewed by independent specialists.

The rubrics evaluate the seven traits considered in our work. The REL trait is assessed on three-level scale: 0 (not relevant), 1 (partially relevant), and 2 (completely relevant). The other six traits are evaluated on a five-level scale: 1 (lowest) to 5 (highest). Each level includes a detailed description summarizing the characteristics of the text classified at that level, as shown in Table 3. For example, a score of 1 in the ORG trait is assigned if the introduction and conclusion are absent, and there is no organization or sequence between paragraphs. In contrast, a score of 5 indicates a well-organized text with a clear introduction, two to three coherent body paragraphs, and a strong conclusion. Similarly, the STY trait is assigned a score of 1 if the text employs very basic linear connecting words such as "and" and "then". The highest score (5) is given when the discourse is well-developed, with good inclusion of subtopics and details, a strong conclusion, appropriate use of a variety of organizational patterns, and a wide range of structural cohesion devices. Overall, a score of zero is given to all traits if the response was solely memorized or copied from the Internet, or if the student failed to attempt the task. In addition to the trait scores, an overall holistic score is computed by summing the seven trait scores.

4.2 Annotation Process

Two Arabic language specialists (main annotators) were selected to evaluate student essays. Both possess teaching and assessment experience with essay questions for a similar age group. One of the annotators is a CAST-certified annotator with prior

[6] https://www.qu.edu.qa/sites/en_US/testing-center/TestDevelopment/cast

السمة	1	2	3	4	5
الهيكل العام **Organization**	المقدمة والخاتمة قد تغيب عن النص، وعرض يفتقر إلى التنظيم والتسلسل بين الفقرات.	المقدمة أو الخاتمة قد تغيب عن النص، وعرض يفتقر إلى التنظيم والتسلسل بين الفقرات.	النص مقبول التنظيم يحتوي على مقدمة وخاتمة، وعرض يحتوي على فقرة واحدة (أو فقرتين) يموزهما حسن الترابط.	النص جيد التنظيم يحتوي على مقدمة تمهد للموضوع وخاتمة مناسبة، وعرض يحتوي على فقرتين إلى ثلاث تتسمان بالتسلسل والترابط.	النص جيد التنظيم يحتوي مقدمة تمهد للموضوع وخاتمة تقود إلى خلاصات فعالة، وعرض يحتوي فقرتين إلى ثلاث تتسم بالتسلسل والترابط الجيد.
الأسلوب والتماسك البنائي **Style**	ربط الكلمات أو مجموعات الكلمات بروابط خطية أساسية جدًا فقط مثل: آو أو ثم إلخ.	تطور الخطاب في شكل قائمة بسيطة من النقاط باستخدام الروابط الأكثر شيوعًا فقط.	تطور الخطاب بشكل مباشر كتسلسل خطي من النقاط باستخدام أدوات تماسك بنائي شائعة.	تطور الخطاب بشكل واضح مع النقاط الرئيسة المدعمة بالتفاصيل ذات الصلة، واستخدام مناسب دائمًا للأنماط التنظيمية المختلفة ومجموعة من أدوات التماسك البنائي مع القفز العرضي في الجمل الطويلة.	تطور الخطاب بشكل متقن مع تضمين الموضوعات الفرعية والتفاصيل بشكل جيد واستنتاج جيد، واستخدام مناسب دائمًا لمجموعة متنوعة من الأنماط التنظيمية ومجموعة واسعة من أدوات التماسك البنائي.

Table 3: A sample grading rubric for the organization and style traits in Arabic. A full detailed English-translated version is in Appendix A.

experience employing the CAST rubrics and international standardized writing assessments. Additionally, a third annotator, a language assessment expert with extensive teaching and scoring expertise for CAST and other language evaluations, was employed to help resolve disagreements between the two primary annotators.

Before starting the scoring process, the two main annotators received training, which included moderation and norming sessions to ensure they fully understand the assessment rubric and maintain consistent annotation procedures. The third annotator led the initial moderation session, providing an overview of the process and QCAW background, and presenting examples of QCAW essays. The rubrics were then thoroughly discussed with an emphasis on the differences between each level within each trait. Next, the norming process took place, where the two main annotators *independently* scored four essays of varying levels. After submitting their scores, a discussion followed about the rationale behind each score, culminating in a group discussion to address any discrepancies. After revising some essays, a second moderation session was held between the two main annotators to reach agreements on specific decisions. The third annotator was subsequently briefed to ensure all decisions were mutually agreed upon by all three annotators. In cases where there is a difference of 11 points or more out of a total of 32 points in the overall holistic score between the main annotators, the third annotator reviews the responses and the scores provided by the main annotators to determine the final score for each trait.

Each writing task was then completely assessed independently by the main annotators. The final score per trait was calculated as the rounded integer mean of the two assessments unless the response was reviewed by the third annotator in the case of score discrepancy as described earlier.

5 Annotation Analysis

In this section, we analyze the annotations via two essay examples explaining score assignments, examine the inter-annotator agreement, and study trait distributions across tasks.

5.1 Example Analysis of Graded Essays

We analyzed two sample essay responses, one from each task, to clarify the scoring rationale employed by the annotators. A detailed description of the two examples is provided in Appendix B.

The first essay, 1-16A, deviated from the given prompt, which called for reflections on the impact of e-mail and telephones on human relations. Instead, the response briefly discussed the pros and cons of social media without substantively engaging with a specified topic or advocating for a particular point of view, leading to a REL score of 1. Despite including an introduction, three paragraphs, and a conclusion, the ORG received a score of 3 due to inadequate coherence. The VOC, though commendable in range, suffered from numerous lexical errors and inappropriate word choices, as they were significantly off-topic and clearly deviated from the main subject, such as "الحوادث المرورية" (traffic accidents), "الأسلاك و الكابلات" (wires and cables), and "التنمر و التهديد" (bullying and threatening), leading to a score of 2. The STY scored 3 points for using various connecting words, albeit with limited use of cause-and-effect indicators and occasional lapses in coherence, as evidenced by simplistic linking words such as "وذلك" (and that) and sequential sentences lacking cohesion. The DEV of ideas, which received a score of 2, only superficially addressed the question's theme without employing persuasive elements, such as citations or examples, resulting in a disjointed narrative lack-

ing coherence. For MEC, the student demonstrated proficiency in punctuation and spelling, earning a score of 3, with no significant errors that detract from readability. The GRA category also received a score of 3 despite pervasive inaccuracies, including errors in derivation and parsing, which could potentially lead to comprehension difficulties upon initial reading. Despite these issues, the essay remained readable, though compromised by lexical and grammatical errors such as "شي" and "أج محمولة".

A higher-level essay example, 2-54A, demonstrated strong REL to the prompt and earned a score of 2, as the ideas directly and accurately addressed the topic without deviation. Its cohesive ORG was commended with a score of 4, showcasing an introductory, followed by four interconnected paragraphs and a concluding summary of the writer's stance. The VOC trait, which gained a score of 5, highlighted the student's adept use of a wide range of conventional vocabulary related to the topic, such as "تطوير و تنمية" (development and growth), "للتطور و الحداثة" (for development and modernity), and "جودة التعليم" (quality of education). In terms of STY, the essay got a score of 4 for effectively employing appropriate linking devices to convey explanations, interpretations, and presentations, ensuring seamless transitions between ideas throughout the paragraphs, such as "وذلك من خلال" (and that is through), "مثالا علي ذلك" (as an example of that), "ولكن" (but), "حيث" (where), and "ولاسيما" (especially). The DEV trait scored 4 for the comprehensive exploration of the impact of technology on learning speed and information accessibility, supported by well-reasoned arguments and examples. The MEC got a score of 4, as the essay proficiently used punctuation and spelling, with minor errors attributable to carelessness, such as "أنها قد" (it is) and "إنماء" (but). The GRA achieved the highest possible score of 5 due to correct and versatile sentence structures, with few observed grammatical errors, showing the student's exceptional writing proficiency. Overall, the essay's strong coherence, vocabulary richness, stylistic finesse, and grammatical accuracy underscore its effectiveness in addressing the given prompt.

5.2 Inter-Annotator Agreement

To analyze the quality of the annotations, we measure the Inter-Annotator Agreement (IAA) using Quadratic Weighted Kappa (QWK), which is widely recognized as the standard metric for AES.

Trait	Task 1	Task 2
REL	0.788	0.817
ORG	0.631	0.807
VOC	0.705	0.766
STY	0.628	0.743
DEV	0.759	0.846
MEC	0.639	0.787
GRA	0.588	0.679
Average	0.677	0.778

Table 4: Inter-Annotator Agreement (IAA) assessed via QWK for Tasks 1 & 2. Colors indicate strength of agreement as Moderate, Substantial, and Almost Perfect.

QWK measures the agreement between human and automated scores (Williamson et al., 2012). Particularly, it is well-suited for dealing with *ordinal* scales, as it takes into account the magnitude of the differences between the ratings, making it a suitable metric for computing the inter-annotator reliability in the context of essay scoring (Doewes et al., 2023b). Table 4 presents the IAA for all traits of Task 1 and 2, along with their corresponding strength of agreement according to the scale outlined by Landis and Koch (1977). In particular, Task 1 poses a greater challenge for grading compared to Task 2, as evidenced by its relatively lower QWK values. This disparity in performance could be attributed to the misinterpretation of Task 1 among some students, which was a challenge encountered during the annotation process. Specifically, Task 1 prompt was about mobile phones and emails and their impact on interpersonal relationships. However, some students instead discussed the effects of social media use on communication. Furthermore, a significant portion of the students did not grasp the nature of an argumentative essay, resulting in them transforming the topic into a discussion essay. These issues likely introduce inconsistencies in the grading process.

For the traits, GRA exhibited the lowest agreement for both tasks, indicating a potential source of ambiguity or subjectivity in the scoring. One possible explanation is the difficulty of establishing a consistent quantifiable grading scale of mistakes in essays of varying lengths and qualities. This variability allows annotators considerable discretion in determining the appropriate number of mistakes associated with each score level, complicating the task of consistently identifying and assessing grammatical errors. In contrast, the REL and DEV traits show higher IAA scores for both tasks, with QWK

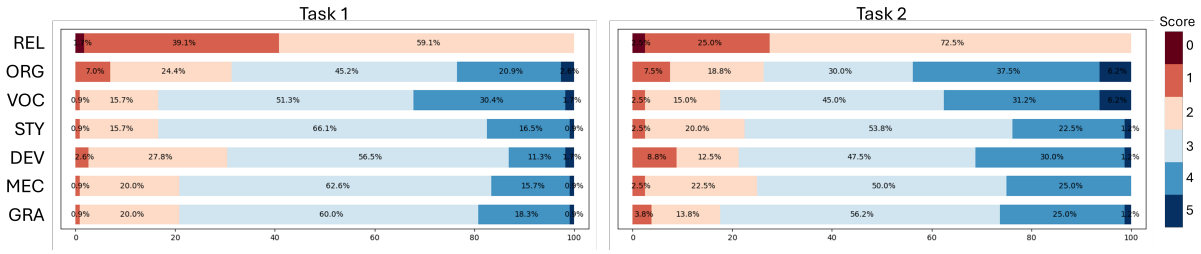

Figure 1: Trait score distribution in *QAES* for the two tasks.

values exceeding 0.75 for Task 1 and 0.8 for Task 2. The higher agreement in REL can be attributed to its three-level scoring rubric, which simplifies the evaluation process. Similarly, the DEV trait benefits from the most detailed rubric among all traits, minimizing the influence of annotator interpretation, and ensuring more consistent scoring. This difference in IAA scores highlights the importance of clear and detailed rubrics in reducing subjectivity and improving inter-annotator reliability.

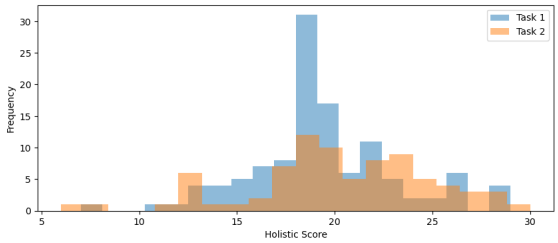

Figure 2: Holistic score distribution in *QAES*.

5.3 Annotation Statistics

Figure 1 depicts the distribution of trait scores across *QAES* tasks. Overall, the distributions resemble a normal distribution, indicating a positive trend. Notably, MEC and GRA traits exhibit consistency across tasks. This suggests that these traits are task-independent, showing the student's language proficiency rather than knowledge of a specific topic. Also, there is a similarity between the distributions of MEC and GRA traits, which suggests a strong correlation between those traits.

An important observation drawn from the distribution of REL scores reveals that Task 2 was notably easier to address than Task 1, as evidenced by the skew towards score 2. This is further supported by Figure 2, which reveals a distribution of holistic scores skewed toward higher values for Task 2. Interestingly, we also observed that the rest of the traits are task-dependent. Consequently, there are more students receiving scores of 4 and 5 for Task 2 than for Task 1, in these traits.

This analysis indicates a clear distinction in writing skills across the two tasks, which highlights the importance of using diverse traits to comprehensively evaluate essay writing skills, offering valuable insights for educators to enhance student performance in essay writing.

6 Arabic AES: Preliminary Experiments

To explore the potential of *QAES* to develop Arabic AES models, we conducted several preliminary experiments. This section details the baseline models and discusses their performance.

6.1 Experimental Setup

In this section, we review the baseline models covering their implementation details, the evaluation measures, and the dataset splits used for analysis.

6.1.1 Baselines

We consider two types of baselines: feature-based and English state-of-the-art (SOTA). The feature-based baseline involves feature engineering to train a regression model. For the English SOTA baseline, we selected models that performed best on the ASAP dataset, including one achieving the best performance at the holistic level (Xie et al., 2022) and another at the traits level (Kumar et al., 2022).

Feature-based (LR) It uses traditional features to capture different aspects of writing proficiency and trains a scoring model with Linear Regression (LR) for its interpretability and simplicity. The features include (i) surface features (e.g., text length in words and characters), (ii) syntactic features (e.g., sentence structure, POS, spelling errors), (iii) lexical features (e.g., vocabulary richness, lexical density), (iv) semantic features (e.g., coherence, perplexity scores), and (v) N-gram features (e.g., frequency of word sequences). Table 9 presents the full list of features used in our model.

Holistic SOTA (NPCR) We implemented the model introduced by Xie et al. (2022) for English

holistic scoring, optimizing regression and ranking. It predicts the difference between representations of a reference essay and the input essay, then adds the reference essay's score to the final prediction. BERT embeddings (Devlin et al., 2019) were originally used for essay representations; we replaced BERT with AraBERT (Antoun et al., 2020) and followed the same setup.

Traits SOTA (STL & MTL) We implemented the multi-task approach by Kumar et al. (2022), which achieved impressive results in English trait assessment by treating individual traits as the primary task. Their model combines CNNs and RNNs, with initial essay representations obtained from GloVe embeddings (Pennington et al., 2014). We tested both their single-task learning (STL) and multi-task learning (MTL) models for both holistic and trait scoring using AraBERT (Antoun et al., 2020) embeddings (instead of GloVe) while retaining their original approach.

6.1.2 Implementation Details

To implement the LR model, we used Scikit-learn's regression models and preprocessing tasks and NLTK for text tokenization and stop-word removal. We used Farasa toolkit (Abdelali et al., 2016) for POS tagging and Spellchecker for spelling error detection. For semantic analysis, we leveraged AraBERT embeddings from the publicly available checkpoints on Hugging Face.[7]

To implement the English SOTA models, we used the publicly available code by Xie et al. (2022).[8] We built the Kumar et al. (2022) model using PyTorch, adopting the hyperparameters in Kumar et al. (2022).

6.1.3 Evaluation Measures

To evaluate our models, we use QWK, the most common metric for assessing agreement between human annotators and systems. However, it has limitations, including a need for a large sample size, and its sensitivity to the score scale (Doewes et al., 2023a). Hence, for all traits except REL, we need at least 50 predictions to calculate QWK and obtain reliable results (Cicchetti, 1981). To address these limitations, we also use the Root Mean Square Error (RMSE), a metric commonly used for ordinal classification tasks (Esuli et al., 2009) similar to predicting score levels in AES tasks. Reporting both QWK and RMSE aims to more reliably evaluate the quality of the tested models.

[7] https://huggingface.co/aubmindlab/bert-base-arabertv02

[8] https://github.com/CarryCKW/AES-NPCR

		LR	STL	MTL	NPCR
Task 1	**QWK**	0.16	0.12	0.11	-0.09
	RMSE	6.08	4.07	4.36	4.74
Task 2	**QWK**	0.26	0.12	0.04	0.04
	RMSE	7.20	5.24	5.18	5.29

Table 5: Evaluation of holistic scoring.

6.1.4 Dataset Splits

As the dataset is small, cross-validation was essential. However, QWK's reliability depends on the sample size, requiring a minimum of 50 essays per trait (except for relevance). Hence, we used a 2-fold cross-validation split: 40% for training, 10% for development, and 50% for testing. This unconventional split maximizes test examples for more reliable results. We have made the splits available for reproducibility.

6.2 Results and Discussions

In this section, we evaluate the baseline performance across various scoring dimensions using QWK and RMSE. We compare the 4 baseline models: LR, STL, MTL, and NPCR, across the two tasks. In addition, we conducted an ablation study on the LR model to assess the impact and determine the significance of each feature category for various traits.

Holistic Scoring Table 5 presents the holistic scoring results. The LR model achieves the highest QWK scores in both tasks, indicating better agreement, while the NPCR model records the lowest QWK scores for both tasks. STL and MTL models show similar agreement levels, though MTL's QWK score drops notably in Task 2. For RMSE, STL and MTL models score lower in Task 1, reflecting more precise predictions. In Task 2, MTL has the lowest RMSE, while LR consistently has the highest RMSE in both tasks. Overall, the performance is weak, indicated by a very low correlation and an error of about 4-7 points.

We note that a single significant disagreement between raters impacts QWK more than RMSE. This implies that LR model, despite having higher QWK values in both tasks, has a higher RMSE because it captures overall patterns but lacks precise scores. Conversely, other models might more ac-

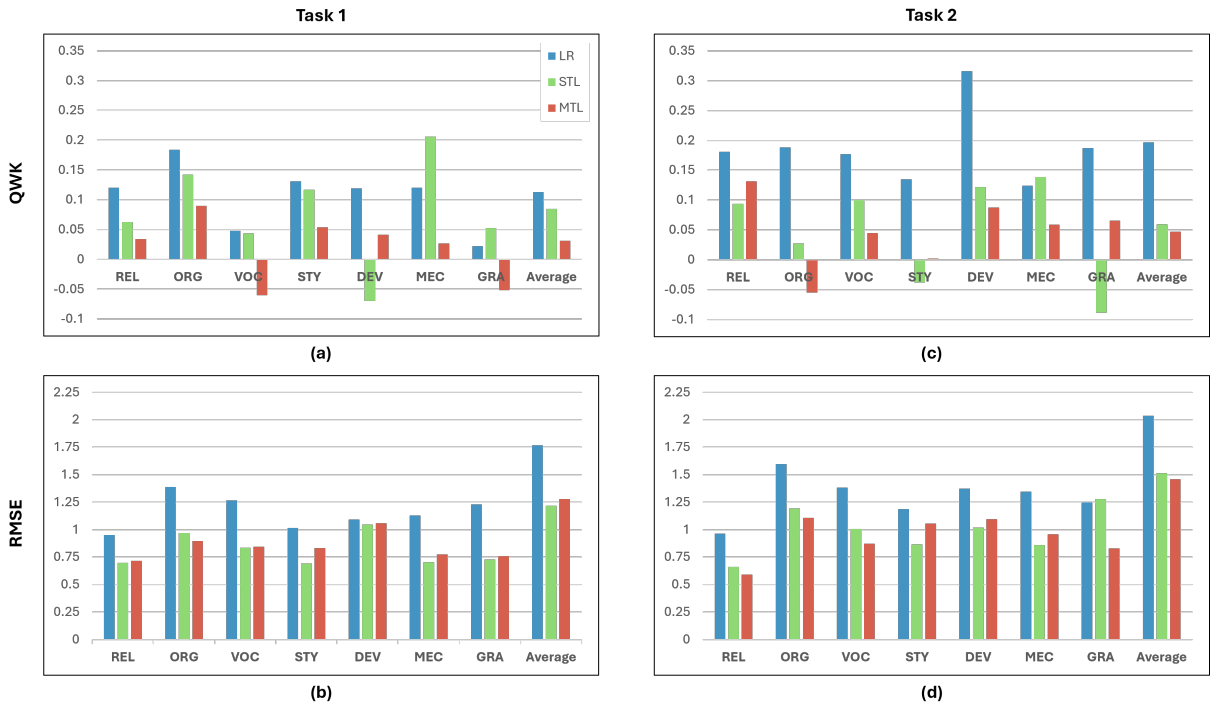

Figure 3: Trait-based performance of LR, STL, and MTL models on Tasks 1 & 2 measured in QWK and RMSE.

curately predict most scores but severely misjudge some, leading to lower RMSE and QWK scores.

Trait Scoring Subfigures (a) and (b) in Figure 3 display the QWK and RMSE performance for Task 1, respectively, while subfigures (c) and (d) show the QWK and RMSE performance for Task 2, respectively. This provides a comparative view of the model's performance across various scoring traits. For both tasks, LR consistently achieved the highest QWK across nearly all dimensions, indicating better agreement with human annotators. For example, in Task 1, LR's QWK score for ORG was 0.184, outperforming STL's 0.142 and MTL's 0.090. Conversely, MTL's struggle is evident, with negative scores in several traits, such as -0.052 for GRA in Task 1 and -0.055 for ORG in Task 2. STL exhibited intermediate performance, generally outperforming MTL but not reaching the efficacy of LR. Task 2 showed higher QWK measures across all models and traits, aligning with our prior observation that it is easier for students to comprehend. The superior performance of LR can be attributed to the dataset's small size, as LR is less dependent on large training examples for effective training. In contrast, STL and MTL, with 439,211 and 411,266,6 parameters respectively, being neural-based methods, typically require more substantial training data to achieve optimal performance.

Regarding RMSE, for Task 1, STL consistently achieved the lowest RMSE values across most dimensions, indicating higher accuracy. For instance, STL recorded an RMSE of 0.698 for REL, compared to LR's 0.951 and MTL's 0.716. However, for Task 2, the trend is not consistent across all traits. While STL still outperforms LR in many cases, MTL tends to have the lowest average RMSE values. This analysis indicates that when considering RMSE measure, both MTL and STL are more effective methods than LR, showcasing their potential as robust models for AES.

Typically, a lower RMSE indicates a higher QWK, reflecting a good fit and agreement. However, the LR model showed the highest QWK but also the highest RMSE, suggesting lower accuracy. This discrepancy may arise from QWK's sensitivity to sample range and quantity.

Conversely, RMSE showed minimal error suggesting good model performance. Despite possible underfitting in STL and MTL neural models, due to their large parameter spaces, their superior RMSE performance indicates that RMSE might be a better measure of model accuracy than QWK for small-scale datasets like *QAES*.

Ablation study We conducted an ablation study to analyze the impact of different feature categories on LR model performance for various traits. A

Task	Excluded Feature Category	REL	ORG	VOC	STY	DEV	MEC	GRA
1	N-gram	-0.020	**0.134**	**0.046**	0.061	**0.101**	**0.109**	**0.010**
	Semantic	0.028	0.011	-0.002	**0.071**	0.052	-0.015	-0.054
	Syntactic	-0.038	0.035	0.010	-0.057	0.010	-0.057	-0.039
	Lexical	-0.027	0.056	0.013	-0.019	-0.008	0.048	0.007
	Surface	**0.163**	0.046	-0.027	-0.077	-0.047	0.039	-0.054
2	N-gram	**0.035**	-0.011	-0.016	-0.037	0.089	0.034	-0.043
	Semantic	0.008	0.024	-0.019	-0.013	0.066	**0.062**	0.012
	Syntactic	0.032	**0.053**	**0.085**	0.039	**0.111**	0.021	**0.116**
	Lexical	-0.025	-0.051	-0.033	-0.004	0.018	0.040	0.060
	Surface	-0.010	-0.153	-0.026	**0.093**	0.032	0.038	-0.008

Table 6: QWK drop when **excluding** one feature category. Highest drop (best) and highest boast (lowest drop, worst) per trait per task when excluding a feature category are boldfaced and underlined respectively.

detailed representation of the drop in QWK when training LR model with all feature categories but one is provided in Table 6. Results show varying effects on QWK scores across traits and tasks. In Task 1, excluding N-gram features significantly reduces performance across all traits except REL, showing their crucial role. Across both tasks, omitting surface features boosts VOC and GRA scores, while removing semantic features diminishes DEV performance as expected, given their role in measuring text coherence and complexity. While lexical and syntactic features are vital for text understanding, their impact varies for each trait and task. This study shows the critical need for a tailored evaluation of feature importance, taking into account the unique characteristics of each trait and the specific requirements of each task.

7 Conclusion and Future work

In conclusion, *QAES* represents a pioneering step in the field of Arabic AES by introducing the first publicly available trait-specific annotations for Arabic essays. Our work addresses a critical gap in Arabic AES research, providing a valuable resource for exploring and developing advanced essay scoring systems based on multiple traits including relevance, organization, vocabulary, style, development, mechanics, and grammar.

Benchmarking against English state-of-the-art baselines and a feature-based approach has demonstrated *QAES*'s potential to serve as a platform for future studies. Despite the challenges encountered during the annotation process, the insights gained pave the way for enhancing the reliability of Arabic AES systems. Moving forward, it is imperative to continue refining the annotation guidelines, expanding the dataset, and exploring innovative methodologies to further improve the progress of Arabic AES. We believe that the journey towards improving AES for Arabic is just beginning, and *QAES* is a promising step forward in this ongoing quest for automated essay scoring proficiency.

8 Limitations

Although this work represents a significant step forward in the application of AES for Arabic, several limitations should be acknowledged. First, the dataset size remains relatively small compared to AES datasets available for other languages. This presents a challenge for the application of machine learning, particularly deep learning models, as evidenced by our preliminary results.

Furthermore, while the reported average agreement between the two annotators is substantial across the two tasks, additional moderation sessions could have potentially resolved more disagreements and clarified ambiguous points.

Lastly, this study did not develop any model specifically tailored for Arabic language assessment. Instead, it focused on adapting existing AES methodologies to the Arabic context. We encourage further research to explore the development of specialized models for Arabic AES, considering the unique linguistic characteristics and challenges of the language.

Acknowledgment

This work was supported by NPRP grant# NPRP14S-0402-210127 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa: A fast and furious segmenter for arabic. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Demonstrations*, pages 11–16.

Hikmat A. Abdeljaber. 2021. Automatic arabic short answers scoring using longest common subsequence and arabic wordnet. *IEEE Access*, 9:76433–76445.

Abdelhamid M. Ahmed, Xiao Zhang, Lameya M. Rezk, and Wajdi Zaghouani. 2024. Building an Annotated L1 Arabic/L2 English Bilingual Writer Corpus: The Qatari Corpus of Argumentative Writing (QCAW). *Corpus-based Studies across Humanities*, 1(1):183–215.

Mansour Alghamdi, Mohamed Alkanhal, Mohamed Al-Badrashiny, Abdulaziz Al-Qabbany, Ali Areshey, and Abdulaziz Alharbi. 2014. A hybrid automatic scoring system for Arabic essays. *AI Communications*, 27(2):103–111.

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 Resource Association.

Aqil M. Azmi, Maram F. Al-Jouie, and Muhammad Hussain. 2019. Aaee – automated evaluation of students' essays in arabic language. *Information Processing Management*, 56(5):1736–1752.

Domenic V Cicchetti. 1981. Testing the normal approximation and minimal sample size requirements of weighted kappa when the number of categories is large. *Applied psychological measurement*, 5(1):101–104.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Afrizal Doewes, Nughthoh Kurdhi, and Akrati Saxena. 2023a. Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring. In *16th International Conference on Educational Data Mining, EDM 2023*, pages 103–113. International Educational Data Mining Society (IEDMS).

Afrizal Doewes, Nughthoh Arfawi Kurdhi, and Akrati Saxena. 2023b. Evaluating Quadratic Weighted Kappa as the Standard Performance Metric for Automated Essay Scoring. In *Proceedings of the 16th International Conference on Educational Data Mining*, pages 103–113. International Educational Data Mining Society.

A. Esuli, S. Baccianella, and F. Sebastiani. 2009. Evaluation measures for ordinal regression. In *Intelligent Systems Design and Applications, International Conference on*, pages 283–287, Los Alamitos, CA, USA. IEEE Computer Society.

Marwa M. Gaheen, Rania M. ElEraky, and Ahmed A. Ewees. 2020. Automated students arabic essay scoring using trained neural network by e-jaya optimization to support personalized system of instruction. *Education and Information Technologies*, 26(1):1165–1181.

Wael Hassan Gomaa. 2014. Arabic Short Answer Scoring with Effective Feedback for Students Arabic Short Answer Scoring with Effective Feedback for Students. (January).

Wael Hassan Gomaa and Aly Aly Fahmy. 2014. Automatic scoring for answers to arabic test questions. *Computer Speech Language*, 28(4):833–857.

Nizar Habash and David Palfreyman. 2022. ZAEBUC: An annotated Arabic-English bilingual writer corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 79–88, Marseille, France. European Language Resources Association.

Beata Beigman Klebanov and Nitin Madnani. 2022. *Automated essay scoring*. Springer Nature.

Rahul Kumar, Sandeep Mathias, Sriparna Saha, and Pushpak Bhattacharyya. 2022. Many hands make light work: Using essay traits to automatically score essays. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1485–1495, Seattle, United States. Association for Computational Linguistics.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Omar Nael, Youssef ELmanyalawy, and Nada Sharaf. 2022. Arascore: A deep learning-based system for arabic short answer scoring. *Array*, 13:100109.

Christopher Michael Ormerod. 2022. Mapping between hidden states and features to validate automated essay scoring using deberta models. *Psychological Test and Assessment Modeling*, 64(4):495–526.

Leila Ouahrani and Djamal Bennouar. 2020. AR-ASAG an ARabic dataset for automatic short answer grading evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2634–2643, Marseille, France. European Language Resources Association.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Abdulaziz Shehab, Mahmoud Faroun, and Magdi Rashad. 2018. An automatic arabic essay grading system based on text similarity algorithms. *International Journal of Advanced Computer Science and Applications*, 9.

David M. Williamson, Xiaoming Xi, and F. Jay Breyer. 2012. A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1):2–13.

Jiayi Xie, Kaiwei Cai, Li Kong, Junsheng Zhou, and Weiguang Qu. 2022. Automated essay scoring via pairwise contrastive regression. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2724–2733, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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, Online. Association for Computational Linguistics.

A Grading Rubric

For annotating *QAES*, we employed the rubric from the Core Academic Skills Test (CAST) developed by the Qatar University Testing Center (QUTC). This rubric was used to score 7 traits: REL, ORG, VOC, STY, DEV, MEC, and GRA. Table 7 presents a translated version of CAST grading rubric for each considered trait.

B Detailed Example Analysis of Graded Essays

In this section, we closely examine two comprehensive examples from *QAES*, one from each task, and thoroughly discuss the rationale behind the scores assigned to the responses by the annotators for the seven different traits: relevance (REL), organization (ORG), vocabulary (VOC), style (STY), development (DEV), mechanics (MEC), and grammar (GRA). Table 8 displays the complete text for the two selected examples, using three different colors to highlight various traits: blue for MEC and GRA, red for STY, and green for VOC. This color-coding provides clear examples for each trait, facilitating a better understanding of how the responses were evaluated.

Example (16A) of Task 1 had 500 words. Unfortunately, the provided response was not directly related to the prompt. Specifically, the prompt requested insight into the effects of email and telephones on human relationships, but the response briefly touched upon the advantages and disadvantages of social media without delving into the primary topic or attempting to persuade readers to adopt a particular stance. Accordingly, a score of 1 was assigned to the REL trait of this response.

The ORG trait received 3 points, as the response had an introduction, three paragraphs, and a conclusion, although coherence could have been better. The VOC category was awarded 2 points, as the writing contained a good range of words; however, many lexical errors and inappropriate word choices that could impede comprehension were found. Phrases and words were used that do not fit the topic and do not convey the intended meaning. Refer to Table 8, where such instances are colored in green. The STY was awarded 3 points for employing a range of standard transitional phrases, though it demonstrated a limited use of cause-and-effect connectors. Additionally, several sentences were missing transitional elements entirely, which are essential for linking them to preceding content and demonstrating the coherence of the paragraph's ideas. For clarity, examples of these linking words are highlighted in red in Table 8.

The Dev of ideas was awarded only 2 points because the response barely addressed the question, providing only a superficial treatment of the relevant ideas. The expected means of persuasion, such as examples, were notably absent. The text began by introducing the concept of social networking sites and proceeded to discuss their benefits in the second paragraph. However, it then shifted to their drawbacks in the following paragraph, and subsequently to issues related to mobile phones. This approach was fragmented and lacked coherence, resulting in an incomplete and disjointed discussion.

The MEC category received 3 points as the student used punctuation and spelling correctly throughout the response, without any frequent or gross errors. Finally, the GRA received 3 points, as there were various grammatical errors, such as derivation and parsing marks, which indicated inaccuracies in sentence construction. The text is full of many spelling and grammatical errors in most paragraphs, which may sometimes cause misun-

Trait	1	2	3	4	5
REL	Partially relevant to the topic	Completely relevant to the topic			
ORG	The introduction and conclusion are absent. There is no organization or sequence between paragraphs.	Either the introduction or conclusion is absent. There is no organization or sequence between paragraphs.	The text is well-organized and contains an introduction and conclusion, but the body has one paragraph (or two paragraphs) that lacks good coherence.	The text is well-organized, contains an appropriate introduction and conclusion, and has two to three body paragraphs that are sequential and coherent.	The text is well-organized and contains an introduction that introduces the topic, a conclusion that effectively concludes the text, and two to three body paragraphs that are sequential and well-connected.
VOC	Use of a limited range of vocabulary and phrases that do not make sense together, with repetition and lexical errors are common, and a generally inappropriate choice of vocabulary that obscures meaning.	Use of a basic range of vocabulary, with repetition, lexical errors, and many inappropriate vocabulary choices that may obscure meaning.	Use a sufficient range of vocabulary, with some repetition and lexical errors, with a small number of inappropriate vocabulary that may obscure the meaning.	Use of a good and appropriate range of vocabulary with few lexical errors, inappropriate choice of vocabulary without affecting meaning, and occasional use of idiomatic expressions.	Use of a broad, correct, and appropriate range of vocabulary with few occasional errors, showing good knowledge of idiomatic expressions and awareness of implicit levels of meaning.
STY	The text employs very basic linear connecting words such as "and" and "then."	Discourse develops as a simple list of points using only the most common connections.	Discourse develops directly as a linear sequence of points using common structural cohesion devices.	Discourse is clearly developed with main points supported by relevant details, always appropriate use of different organizational patterns, and a range of structural cohesion devices with occasional 'jumping' in long sentences.	Discourse is well developed, with good inclusion of subtopics and details and a good conclusion, always appropriate use of a variety of organizational patterns, and a wide range of structural cohesion devices.
DEV	The content is not related to the subject of the question; the ideas are characterized by randomness; most of them lack coherence, sequence, and succession, and the main idea disappears with the use of general structures that are not related to the persuasive text, and a clear lack of evidence and evidence in the text.	The content is relatively related to the topic of the question, the ideas are sequential, and the main idea clearly disappears during writing, with limited coverage of all opinions and not being fully presented, with the use of many methods that do not support the persuasive text, and the use of structures that do not express the meaning.	The content is completely related to the subject of the question. The ideas are characterized by succession in most of the text, with the main idea gradually disappearing and the presence of an implicit, unclear adoption of a specific opinion or position, with the use of several evidence and proofs that need organization, sequence, and coherence in their presentation, and the employment of some well-known persuasive methods that It does not emphasize the idea and importance of the topic.	The content is completely related to the topic of the question. The ideas are characterized by clarity, organization, sequence, and coherence, with a clear appearance of the main idea in the text with its relationship to the sub-ideas that maintain their connection to the main idea, in addition to adopting a specific, clear position towards an issue, while presenting some arguments, evidence, and evidence coherently. A comprehensive presentation of the different opinions related to the topic and the use of three persuasive methods such as examples, conclusions, sayings, and others.	The content is completely related to the subject of the text. The ideas are characterized by clarity, organization, sequence, and coherence, with a clear appearance of the main idea in the text with its relationship to the sub-ideas that maintain its connection to the main idea, in addition to adopting a specific, clear position towards an issue, while presenting arguments, evidence, and evidence coherently. A comprehensive presentation of the different opinions related to the topic, and the presence of a variety of persuasive methods such as examples, conclusions, sayings, etc., as well as the correct use of structures that support persuasion and influence (rhetorical questions, intentional ambiguity, interjection sentences, etc.)
MEC	Limited application of spelling rules.	Frequent spelling and punctuation errors.	Effectively applies standard formatting, paragraphing, spelling, and punctuation most of the time.	Effectively applies standard formatting, paragraphing, spelling, and punctuation with few errors.	Completely accurate paragraph organization, punctuation, and spelling, except for a few occasional pen slips.
GRA	Use a limited set of simple grammatical structures and sentence patterns with little flexibility and/or precision.	Correct use of some simple grammatical structures with frequent, systematic errors that may obscure the meaning.	The use of a variety of grammatical structures, with notable errors and imprecisions, can sometimes obscure the meaning.	Good use of variety of grammatical structures with rare errors and minor imperfections in sentence structure that do not affect the meaning.	Always correct and flexible use of a wide variety of grammatical constructions with occasional minor slips.

Table 7: CAST Persuasive/Argumentative Writing Rubric - English Translation.

derstanding when reading the sentence for the first time. Examples of these errors are colored in blue in Table 8.

The chosen sample (54A) for Task 2 comprised 500 words. It earned 2 marks for REL. Specifically, the prompt requested insights into how technology has enabled students nowadays to learn more information quickly and effectively. The student's response was entirely pertinent to the subject matter of the question, with all the ideas presented being directly related to the topic and no ideas deviating from the content of the required question. The essay began with an introduction, followed by four successive and well-connected paragraphs, culminating in a conclusion that summarizes the writer's stance on the issue at hand. This ORG was awarded 4 marks. For VOC, which got 5 marks, the student utilized an extensive range of conventional vocabulary with connotations, with the majority of their vocabulary stemming from the semantic field of the topic. Examples of such word choices are colored in green in Table 8.

The use of appropriate linking devices that indicated explanation, interpretation, and presentation according to the ideas presented earned the essay 4 points in the STY category. Examples of such devices are shown in red in Table 8. The topic was conveyed through sequential and interconnected paragraphs, with each idea seamlessly leading to the next one. The DEV trait was awarded 4 marks due to the student's well-reasoned response on the topic of technology and its impact on the speed of learning and access to information. The writer's stance was clearly articulated and supported by comprehensive examples and explanations that demonstrated the validity of the content presented. All the ideas, both main and subsidiary, were related to the content of the question. Examples of ideas presented to enrich the topic: in the first paragraph, "facilitating the teaching method" and "developed traditional methods into modern ones"; in the second paragraph "Facilitating the delivery of information to students", "facilitating students' understanding", and "taking into account learning styles"; in the third paragraph "delivering information to students", and "helps take into account different learning styles". In the fourth paragraph, "Help take care of people with special needs" and in the fifth paragraph, "Qatar's role in developing the technology factor", and "Qatar Vision 2030".

The student demonstrated strong proficiency in punctuation and spelling throughout the essay, with only minor errors resulting from carelessness or haste. 4 marks were awarded to the student in this MEC area. The GRA score was the highest possible (5 points) because of the students' correct and versatile use of sentence structures and writing forms. The writing contained various grammatical structures with rare errors. Generally speaking, the errors in mechanics and grammar did not exceed 10 errors in 500 words. This is excellent - especially when the errors are lapses, such as the words colored in blue shown in Table 8.

C Extracted Features

We utilized five categories of features in the development of our feature-based model. Table 9 presents all the extracted features across these categories, accompanied by brief descriptions. The feature list comprises 23 Surface features, 12 Syntactic features, 5 Lexical features, 2 Semantic features, and 500 unigram and bigram features.

Task 1 - Communication (16A)	Task 2 - Technology (54A)
حديثا، طغت وسائل التواصل الاجتماعي على حياتنا وأصبحت الوسيلة الأولى في تواصل الأمة البشرية بين بعضها البعض، فهذا التطور قد اختصر الوقت والجهد، وساهم بطريقة فعالة في تغيير نمط حياة الإنسان، فقد سهل عملية التواصل بين الناس في مختلف أنحاء العالم ، ولكن، هل يؤثر ذلك على مجتمعاتنا سلبيا؟ يرى الفريق المؤيد أن وسائل التواصل والبريد الإلكتروني نهضة كبيرة وبدون هذا الاختراع لم نكن لنصل إلى كل هذا التطور الكبير في وقت قصير، وذلك لأن التواصل هو أساس بناء كل شيء. بالإضافة إلى ذلك، هو موفرا للمال والوقت، لأن الرسائل قديما كانت ترسل من مكتب البريد وكان يكلف المال وتستغرق عملية وصول الرسالة إلى الطرف الآخر شهورا طويلة، فكان ذلك عائق بين الناس لأن محادثة واحدة بين طرفين قد تستغرق سنين طويلة. زيادة على ذلك، إن الهواتف النقالة لا تحتوي على الأسلاك والكابلات مما يجعلها أخ محمولة وسهلة التنقل. أصبحت إمكانية استخدام الهاتف النقال أكثر من الهواتف الأرضية. من هنا، يوفر الهاتف النقال خدمه الاتصال بالطوارئ في أي وقت ومكان حتى وإن لم تشمل المنطقة تغطية هاتفية أو إرسال. بالإضافة إلى كل ما سبق، تسهل مواقع التواصل سهولة الوصول إلى الأخبار والأحداث حول العالم في حدود ثواني ودقائق معدودة، ليكون الناس على وعي بأخبار العالم الاقتصادية والسياسية والبيئية وغيرها.كما أنها وسيلة فعالة للترويج للسلع التجارية عبر المواقع المتاحة للبيع والشراء. في الجهة الثانية، يرى الفريق المعارض أن كثرة استخدام والجلوس على مواقع التوصل الاجتماعي والهواتف النقالة بشكل يومي للتوصل تفقد الفرد مهارة التواصل الشخصي ويصبح من الصعب عليه التواصل اللموس مع الناس وإنشاء روابط اجتماعية في أماكن العمل والمدرسة وغيرها. ولأن التواصل عبر الهاتف النقال ليس تواصل مباشرا، فإن سوء تفاهم إحدى الطرفين للآخر أكثر، لعدم توفر لغة الجسد ونبرة الصوت في هذا النوع من التواصل. تقلل الهواتف النقالة من الزيارات العائلية وتزيد من نسبة قطع الأرحام، وذلك لأن الناس أصبحت تتواصل بشكل دائم بالتكنولوجيا الحديثة. وأصبحت مباركة الناس لبعضها في الأفراح ومواساتهم في الآخر أن تقتصر فقط على الرسالات النصية. علاوة على ذلك، إن الهواتف النقالة تزيد من نسبة الإصابة بالحوادث المرورية مما يبين خطورة استخدامها أثناء القيادة، فهي تهدد حياة الفرد وكل من حوله على الطريق، وقد كشفت دراسة أعدتها جامعة ايوأ الأمريكية أن انشغال قائد السيارة باستخدام الهاتف المحمول في أثناء القيادة يؤثر سلبا على قدرة الدماغ. أيضا، أصبحت ممارسة التنمر والتهديد والتخويف ومضايقة الناس غير مقتصرة فقط وجها لوجه، فالكثير من الأشخاص يتنمرون على الناس عبر مواقع التواصل بدون الكشف عن هويتهم الحقيقية، خاصة على فئة الأطفال والمراهقين، مما يسبب الكثير من المشاكل النفسية وفي بعض الأحيان يؤدي إلى الانتحار. وبلا شك، إن استخدام الهواتف النقالة تطور بشكل سلبي كبير عند الكثير من المراهقين لتصبح مثل الإدمان، فتستخدم بإفراط ولا مبالاة مما يؤثر سلبا على المهام اليومية والواجبات المنزلية لدي. في الختام، إن استخدام أحدث الوسائل لتسهيل التواصل بيننا كمجتمعات ليس أمرا خاطئا، حيث أننا نعيش في جيل يعتمد كامل الاعتماد على استخدام الهواتف والتواصل عبره، ولكن عندما يتعدى الفرد المستوى الصحيح والأمثل لاستخدامه، يصبح خطرا في جوانب كثيرة، فلابد منا أن نوازن بين استخدام الهاتف والتواصل عبره مع مهامنا اليومية وتواصلنا مع الناس حولنا، لأن ذلك يساعدنا على الاحتفاظ وتطوير مهاراتنا التواصلية في كلتا الجهتين	نحن وفي أواخر القرن العشرين، أصبح التعليم رمزا للتطور والحداثة بين الشعوب، فكثيرا ما نرى في الآونة الأخيرة استخدام التكنولوجيا أثناء عملية التعليم، حيث يوقن البعض أن التكنولوجيا تزيد قدرات الإنسان وإمكانياته، فعندما يصبح التعليم ممزوجا بالتكنولوجيا، سيتحقق التطور حتما، ولكن البعض يرى أن ذلك التطور يقتصر على نهضه البلاد. بصفتي كطالبة تشهد اهتمام دولة قطر بالمجال التعليمي، أرى أن التكنولوجيا قد كانت عاملا مهما في تطوير طريقه التدريس للتلاميذ وتسهيل عملية استيعاب وفهم التلاميذ للمعلومات، أي أنها عملت على تطوير وتنمية العملية التعليمية. عندما يتم إدخال التكنولوجيا في مجال التعليم ، يشكل ذلك تغييرا جذريا في المستوى التعليمي ، فيشمل ذلك التغيير المعلم والطلاب و وسيلة التدريس، حيث أنها تجعل طريقة التدريس أسهل على المعلمين من خلال تطوير الوسائل التقليدية للتدريس إلى وسائل حديثة، ومثالا على ذلك استخدام المعلمين لتقنية العرض المسماه بـ (الداتا شو) وهي آلية مستخدمة لعرض الدروس على التلاميذ . فقد سهلت تلك التقنية عملية إيصال المعلومات للطلاب، فنلاحظ أن المعلم لا يواجه صعوبات بالغة أثناء محولاته في إيصال المعلومات.كما أن جميع التقنيات المستخدمة في التدريس قد ساهمت في إرضاء جميع العقليات ، فهي قد راعت أنماط التعلم المختلفة لدى الطلاب، وذلك من خلال توفير آليات التدريس الحديثة بجميع الأنشطة أو الفعاليات التعليمية، أي أن الطالب الذي يمتلك النمط البصري، ستتاح له الفرصة في التعلم من خلال الصور والفيديوهات التي تعرض فسيتمكنه ذلك من تلقي المعلومات بشكل أسرع وأوضح. لاشك أن ذلك لا يقتصر فقط على الطالب البصري لاسيما أيضا الطالب السمعي والحسي ، وكل ذلك يساهم في تنمية القدرات الذهنية للطالب فتساعده على تلقي المعلومات بشكل ممتع وسريع ، وفي المقابل أيضا يطور ذلك قدرات المعلم في الشرح فتزيد من إتقانه لمهنته مما ينعكس على الجيل القادم. إذا تلقينا المسألة من منظور آخر، فليكن من منظور ذوي الاحتياجات الخاصة فهم جزء لا يتجزأ من المجتمع. لقد حظي الطلاب من ذوي الاحتياجات الخاصة برعاية كبيرة تمثلت في توفير التعليم الأمثل لهم عن طريق مساعدتهم في تلقيه بشتى مستوياته من خلال التكنولوجيا، مما منحهم الحصول على نفس نوعية وجودة التعليم المتلقاة لدى المتعلمين الأصحاء. بالتأكيد و بلا أدني شك أن التكنولوجيا قد نتج عنها فوائد عديدة لذوي الاحتياجات الخاصة كثيرا ولاسيما في السنوات القادمة. كما نعلم أن دولتة قطر من أكثر الدول التي تحرص على الاهتمام بالجانب التعليمي لكافة الفئات وخاصة ذوي الاحتياجات الخاصة، حيث أنها قد أسست مركز الشفلح الذي يسهم في تعليم ذوي الاحتياجات الخاصة ورعايتهم في الجانب التعليمي. مركز الشفلح يعتمد اعتمادا كبيرا على التكنولوجيا، فهم يحتاجون إلى اكتساب المهارات الأكاديمية اللازمة لتكيفهم مع مجتمعهم المحيط بهم، حيث يتطلب تعلم المهارة واكتسابها مشاهدة نموذج للأداء، وممارسة هذا الأداء، وكلا الأمرين يتطلب الاستعانة بوسائل تكنولوجيا التعليم. نتيجة لجميع تلك الجهود هو إنماء ذوي الاحتياجات الخاصة فتصبح الإعاقة ليست عائقا أمام التعلم في ظل ماوصلنا إليه من تكنولوجيا. إن التعليم هو الأساس الذي تقوم عليه حياة الأفراد، وتكمن أهميته في العديد من المحاور، فتعود فائدته على المتعلم ثم الدولة والمجتمع .كما نرى أيضا اهتمام دولة قطر في تكمين مفهوم الاستدامة في التعليم ،كما تعتمد عليه رؤية ٢٠٣٠ ، فإنه في غاية الأهمية أن يكون بأفضل مستوى كي تتمكن من إنتاج أجيال فعالة، فدمج التكنولوجيا بالعملية التعليمية يشكل عاملا مهما في تحسين الأجيال المقبلة
Assigned Scores: REL: 1, ORG: 3, VOC: 2, STY: 3, DEV: 2, MEC: 3, GRA: 3, HOL: 17	**Assigned Scores:** REL: 2, ORG: 4, VOC: 5, STY: 4, DEV: 4, MEC: 4, GRA: 5, HOL: 28

Table 8: Essays from *QAES* Color-coded by Traits: MEC/GRA in Blue, STY in Red, VOC in Green.

Feature Category	Feature Name	Description
Surface Features	Words_count	Total number of words in the text.
	Log_words_count	Logarithm of the total number of words.
	Unique_words_count	Number of distinct words in the text.
	Log_unique_words_count	Logarithm of the number of unique words.
	Average_word_length	Mean length of words in the text.
	Max_length_word	Length of the longest word.
	Min_length_word	Length of the shortest word.
	Standard_deviation_words	Standard deviation of word lengths.
	Chars_conut	Total number of characters in the text.
	Hmpz_count	Total number of همزة in the text
	Paragraphs_count	Total number of paragraphs.
	Is_first_paragraph <= 10	Indicates if the first paragraph has 10 or fewer words (binary).
	Average_length_paragraph	Mean length of paragraphs.
	Max_length_paragraph	Length of the longest paragraph.
	Min_length_paragraph	Length of the shortest paragraph.
	Has_parentheses	Indicates if the text contains parentheses (binary).
	Has_colon	Indicates if the text contains a colon (binary).
	Has_question_mark	Indicates if the text contains a question mark (binary).
	Sentences_count	Total number of sentences.
	Average_length_sentence	Mean length of sentences.
	Max_length_sentence	Length of the longest sentence.
	Min_length_sentence	Length of the shortest sentence.
	Standard_deviation_sentence	Standard deviation of sentence lengths.
Syntactic Features	noun_count	Total number of nouns in the text.
	verb_count	Total number of verbs in the text.
	adj_count	Total number of adjectives in the text.
	punc_count	Total number of punctuation marks in the text.
	pron_count	Total number of pronouns in the text.
	pre_count	Total number of prepositions in the text.
	adv_count	Total number of adverbs in the text.
	conj_count	Total number of conjunctions in the text.
	num_count	Total number of numerical values in the text.
	misspelled_count	Total number of misspelled words in the text.
	inna_count	Total number of إن وأخواتها in the text
	kana_count	Total number of كان وأخواتها in the text
Lexical Features	stop_words_count	Total number of stop words in the text.
	words_count_without_stopwords	Total number of words excluding stop words.
	first_paragraph_has_intro_words	Indicates if the first paragraph contains introductory words (binary).
	last_paragraph_has_conclusion_words	Indicates if the last paragraph contains concluding words (binary).
	lexical density	Ratio of content words (nouns, verbs, adjectives, adverbs) to the total number of words.
Semantic Features	cosine_similarity	Measure of similarity between two text embedding vectors.
	preplexity_score	Measure of how well a language model predicts the text.
N-gram Features	unigram and bigram	Features based on the frequency of single words (unigrams) and pairs of consecutive words (bigrams) in the text.

Table 9: List of all Extracted Features.