LAILA: A Large Trait-Based Dataset for Arabic Automated Essay Scoring

May Bashendy[1]**, Walid Massoud**[1]**, Sohaila Eltanbouly**[1]**, Salam Albatarni**[1]
Marwan Sayed[1]**, Abrar Abir**[2]**, Houda Bouamor**[2]**, Tamer Elsayed**[1]

[1]Computer Science and Engineering Department, Qatar University
[2]Carnegie Mellon University in Qatar

{ma1403845, wmassoud, se1403101, sa1800633, me2104862, telsayed}@qu.edu.qa,
{aabir, hbouamor}@cmu.edu

Abstract

Automated Essay Scoring (AES) has gained increasing attention in recent years, yet research on Arabic AES remains limited due to the lack of publicly available datasets. To address this, we introduce *LAILA*, the *largest publicly available* Arabic AES dataset to date, comprising 7,859 essays annotated with *holistic* and *trait-specific* scores on *seven* dimensions: relevance, organization, vocabulary, style, development, mechanics, and grammar. We detail the dataset design, collection, and annotations, and provide benchmark results using state-of-the-art Arabic and English models in prompt-specific and cross-prompt settings. *LAILA* fills a critical need in Arabic AES research, supporting the development of robust scoring systems.

1 Introduction

Automated Essay Scoring (AES) has emerged as a key area of research in writing assessment, as it not only minimizes the time and effort required for manual grading but also enables educators to provide timely and detailed feedback. By emphasizing linguistic quality rather than factual recall, AES systems assess writing in a manner that is independent of specific curricular content.

Since the 1960s, most AES research has predominantly focused on English (Page, 1966), supported by the availability of several large-scale public datasets. In contrast, Arabic AES research faces a significant data scarcity problem, as publicly available annotated datasets remain limited. Existing Arabic resources are often small in scale such as QAES (Bashendy et al., 2024), lack trait-specific[1] annotations as in ZAEBAC (Habash and Palfreyman, 2022), or consist solely of unannotated essays such as ALC.[2] These limitations have constrained the development of robust and scalable Arabic AES systems, despite some promising efforts (Alghamdi et al., 2014; Gaheen et al., 2020).

Trait	Description
Relevance الصلة بالموضوع	Alignment with the prompt
Organization الهيكل العام	The structure of the essay
Vocabulary المفردات	Word choice and variety
Style الأسلوب والتماسك البنائي	Linking and transitions
Development تطور الأفكار والمضمون	Clarity and support of ideas
Mechanics الإملاء والترقيم والتنسيق	Spelling and punctuation
Grammar البناء والتراكيب	Grammatical accuracy
Holistic التقييم الكلي	Overall writing quality

Table 1: Brief description of the traits in *LAILA*.

In this paper, we introduce *LAILA*,[3] the *first large-scale* Arabic AES dataset featuring both holistic and trait-specific annotations of seven writing proficiency traits, described in Table 1. The dataset comprises 7,859 essays across 8 distinct prompts making it the most extensive publicly available Arabic resource of its kind.[4] To construct *LAILA*, our team of researchers conducted a large-scale data collection effort over the course of one academic year. Under Institutional Review Board (IRB) approval,[5] we visited 24 high schools in Qatar and collected essays directly from 4,372 students in authentic classroom settings. We describe the complete dataset development pipeline, from data collection design to annotation procedures and public release, to facilitate future research on Arabic AES and writing assessment. Our contributions are as follows:

- **Novel Dataset**: We introduce the first large-scale dataset of Arabic essays, filling a key gap in the existing Arabic AES literature.

- **Annotation Guidelines**: We develop and share comprehensive guidelines for data annotation to ensure transparency and reproducibility.

- **Trait-Specific Annotations**: We provide de-

---

[1]A trait is an aspect of student writing, e.g., organization.
[2]https://www.arabiclearnercorpus.com

[3]*LAILA* stands for "**L**inguistic **A**ssessment with tra**I**ts for **L**earning **A**rabic," and is pronounced in Arabic as "لَيْلَى".

[4]A prompt is the text that describes an essay writing task.
[5]IRB Number: QU-IRB 159/2024-EA

Dataset	Lang	Essays	Prompts	Len	Level	L1/L2	HOL	Traits	Public
ASAP++ (Mathias and Bhattacharyya, 2018)	EN	12,978	8	281	G7-10	L2	✓	✓	✓
TOEFL11 (Blanchard et al., 2013)	EN	12,100	8	348	-	L2	✓	✗	✗
ELLIPSE (Crossley et al., 2023a)	EN	6,500	44	427	G8-12	Both	✓	✓	✓
PERSUADE (Crossley et al., 2023b)	EN	25,000	15	418	G6-12	L1	✓	✗	✓
ACEA (He et al., 2022)	ZH	1,220	-	-	G12	L1	✓	✓	✗
TCFLE-8 (Wilkens et al., 2023)	FR	6,569	3	119	-	Both	✓	✗	✓
MERLIN (Boyd et al., 2014)	Eur	2,287	-	-	-	Both	✓	✗	✓
ZAEBUC (Habash and Palfreyman, 2022)	AR	214	3	156	College	L2	✓	✗	✓
QCAW (Ahmed et al., 2024)	AR	195	2	499	College	L1	✓	✗	✗
Abbir (Alghamdi et al., 2014)	AR	640	2	150	College	L1	✓	✗	✗
AAEE (Azmi et al., 2019)	AR	350	8	-	G7-12	L1	✓	✗	✗
QAES (Bashendy et al., 2024)	AR	195	2	489	College	L1	✓	✓	✓
TAQEEM 2025 (Bashendy et al., 2025)	AR	1,265	4	151	G10-12	L1	✓	✓	✓
LAILA	AR	7,859	8	171	G10-12	L1	✓	✓	✓

Table 2: Comparison of *LAILA* with existing essay datasets. "Len" is average essay length in words; "HOL" indicates holistic scoring; "Eur" covers German, Italian, and Czech; "L1/L2" denotes native or second-language learners; "Public" here means freely available.

tailed trait-specific annotations that capture multiple dimensions of writing proficiency.

- **Benchmarking**: We establish baseline AES results for Arabic under two evaluation setups: prompt-specific[6] and cross-prompt.[7]

- **Public Release**: We publicly release *LAILA*,[8] including essays with holistic and trait-specific annotations. We also release the benchmarking code [9] to support replication and future research.

The remainder of this paper is organized as follows. Section 2 reviews related work and existing AES datasets. Section 3 outlines the data collection design principles, while Section 4 describes the dataset construction process, including school and prompt selection, essay collection, and annotations. Section 5 presents dataset analysis, and Section 6 details the public dataset release. Section 7 reports preliminary experiments and results, and Section 8 concludes with final remarks.

2 Related Work

To contextualize our contribution, we review prior AES research across languages and modeling paradigms, outlining existing datasets and the evolution of Arabic AES from feature-based to neural and LLM-based approaches. Table 2 summarizes

major AES datasets across languages, underscoring the novelty and significance of *LAILA* as a benchmark resource for future Arabic AES research.

2.1 AES Datasets

Non-Arabic AES Datasets   Research on AES has advanced considerably in English, supported by large-scale, publicly available datasets such as ASAP/ASAP++ (Mathias and Bhattacharyya, 2018), TOEFL11 (Blanchard et al., 2013), ELLIPSE (Crossley et al., 2023a), and PERSUADE (Crossley et al., 2023b). These resources have enabled benchmarking and model development for both prompt-specific and cross-prompt AES tasks. In contrast, other languages have far fewer resources. Some smaller datasets exist, including ACEA (Chinese; annotated but not public) (He et al., 2022), TCFLE-8 (French; public but with only holistic scores) (Wilkens et al., 2023), and MERLIN (German, Italian, and Czech; public but limited to holistic annotations) (Boyd et al., 2014). While useful for AES studies, their limited scale and annotation depth restrict their utility.

Arabic AES Datasets   Compared with English and other languages, Arabic AES lacks robust datasets. Few datasets have been developed, both public and proprietary, yet all face limitations in scale, annotation depth, or accessibility. Publicly available resources include the Zayed Arabic English Bilingual Undergraduate Corpus (ZAEBUC) (Habash and Palfreyman, 2022), which of-

---

fers linguistic annotations (POS tagging, lemmatization, and spelling corrections) and CEFR proficiency levels but lacks trait-level AES labels. The Qatari Corpus of Argumentative Writing (QCAW) (Ahmed et al., 2024) contains 195 essays with POS annotations but restricted holistic scores, while its extension, QAES (Bashendy et al., 2024), introduces both holistic and trait-level annotations. Building on these efforts, ZaQQ (Elsayed et al., 2025) merges QAES, ZAEBUC, and QALB (Mohit et al., 2014), annotated using both human raters and LLMs. More recently, the TAQEEM 2025 shared task dataset (Bashendy et al., 2025) released 1,265 essays across 4 prompts, and TAQAE (Sayed et al., 2025) combined QAES with 2 prompts from TAQEEM. It is important to note that *LAILA* reuses the 2 prompt texts from QAES dataset (with no essays taken from it), and fully include TAQEEM 2025 dataset (4 prompts with 1,265 essays).

Beyond these datasets, several proprietary resources have been used, including Abbir (Alghamdi et al., 2014), which contains essays by Saudi university students with holistic scores, and AAEE (Azmi et al., 2019), comprising essays from grades 7–12 evaluated for content, style, and spelling. However, their unavailability limits reproducibility and progress in Arabic AES.

Overall, existing Arabic AES resources remain fragmented and small in scale (typically under 2,000 essays), limiting robust model training and standardized benchmarking. In contrast to prior work, *LAILA* provides a large-scale, publicly-available Arabic AES dataset with holistic and trait-specific annotations. In terms of scale and annotation richness, *LAILA* is comparable to leading datasets in other languages and fills a critical gap in the current Arabic AES landscape.

2.2 Arabic AES Systems

Despite the scarcity of datasets, several studies have explored Arabic AES. Early approaches primarily relied on handcrafted feature engineering (Alqahtani and Alsaif, 2020; Alsanie et al., 2022), and text similarity techniques (Abdeljaber, 2021; Alobed et al., 2021a; Al Awaida et al., 2019; Alobed et al., 2021b; Azmi et al., 2019). More recent work has shifted towards leveraging pre-trained language models, notably through fine-tuning AraBERT (Ghazawi and Simpson, 2024), integrating handcrafted features (Machhout and Zribi, 2024), or applying parameter-efficient adaptation

methods (Mahmoud et al., 2024). Most recently, Sayed et al. (2025) initiated cross-prompt research in Arabic AES by developing a set of engineered features and evaluating their effectiveness across feature-based and encoder-based models. Additionally, Ghazawi and Simpson (2025) pioneered the use of LLMs for Arabic AES, assessing systems such as GPT and LLaMA under various prompting configurations. In this work, we benchmark *LAILA* dataset with state-of-the-art (SOTA) Arabic and English models to establish strong and reproducible baselines for future research.

3 Dataset Design Principles

The development of *LAILA* was guided by nine core design principles [D1–D9], each reflecting deliberate choices to ensure a high-quality, representative, and ethically sound resource for Arabic AES. Below, we outline these principles.

- **[D1] Ensuring Data Integrity:** We designed *LAILA* to capture authentic, classroom-produced writing created without external assistance. This design choice preserves genuine linguistic variation, learner errors, and authentic complexities essential for real-world AES applications.

- **[D2] Maximizing Diversity:** To reflect the breadth of Arabic writing proficiency, *LAILA* was planned to collect essays from multiple educational institutions and student populations with diverse academic and demographic backgrounds. Prompts were intentionally varied across genres, ensuring cultural relevance and age appropriateness, thereby promoting model generalization and reducing bias.

- **[D3] Achieving Gender Balance:** We intentionally aimed to balance the number of essays from male and female students to promote fairness, reduce gender-based bias, support inclusive evaluation, and enhance model generalization.

- **[D4] Supporting Full Trait Coverage:** *LAILA* was built to capture multiple dimensions of writing quality through annotations of relevance, organization, vocabulary, style, development, mechanics, and grammar. This design supports fine-grained assessment and meaningful feedback.

- **[D5] Ensuring Authentic Writing Conditions:** To emulate real academic or high-stakes testing scenarios, *LAILA* was designed to collect essays written under timed conditions using a digital submission platform. This setup ensures consis-

Figure 1: An overview of the construction process of *LAILA* dataset.

tency in data collection and eliminates manual transcription errors.

- **[D6] Protecting Participant Privacy:** To ensure ethical research and data protection, *LAILA* was planned with strict privacy measures by anonymizing all essays and collecting them with informed consent from students and guardians.

- **[D7] Standardizing the Annotation Process:** To ensure reliability, *LAILA* aimed to use a standardized rubric for evaluating essays across all traits, with planned training for annotators to maintain consistent interpretations of the guidelines and achieve high inter-annotator agreement.

- **[D8] Promoting Scalability and Reusability:** *LAILA* was planned at a large scale to enable robust AES model training. All data and annotations are provided in standardized, machine-readable formats with comprehensive documentation and an open license to support future reuse.

- **[D9] Supporting Multiple Evaluation Setups:** We structured *LAILA* to support both *prompt-specific* and *cross-prompt* AES setups, facilitating research on model generalization.

4 Dataset Construction

The construction of *LAILA* follows three stages: *Sampling*, involving the selection of students, schools, and writing prompts; *Data Collection*, covering essay collection and data cleaning; and *Annotations*, focusing on scoring the collected essays. Figure 1 shows an overview of the overall process.

4.1 Sampling

The first stage of constructing *LAILA* involved sampling schools, students, and writing prompts. Data were collected from 24 high schools in Qatar, evenly divided between female and male institutions [D3], selected for having Arabic as the primary language of their studies. The selected schools provided a representative sample encompassing diversity in geographic location, size, and academic performance [D2]. All participating insti-

tutions were public and followed a unified national curriculum, ensuring a consistent academic context for the writing tasks.

Students were drawn from grades 10 to 12 to capture a wide range of writing proficiency. This age group was chosen because students at this level typically produce essays of sufficient length and complexity to address a variety of prompts. To reflect the diversity of academic performance, approximately 30% of participants were low-performing, 40% medium-performing, and 30% high-performing [D2]. Participation was voluntary and required parental consent to ensure compliance with ethical research standards [D6].

Writing prompts were selected to elicit critical thinking and showcase different aspects of writing ability. The dataset includes eight distinct prompts, five persuasive and three explanatory, designed to capture both persuasive reasoning and expository clarity [D2]. This diversity supports the development of AES models capable of generalizing across writing genres and skill levels.

4.2 Data Collection

The second stage of constructing *LAILA* focused on the systematic collection of high-quality student essays under controlled conditions. This phase involved coordinating with participating schools, preparing the technical setup for essay submission, conducting supervised on-site writing sessions, and performing post-collection data cleaning to ensure completeness, authenticity, and reliability.

Briefing To ensure a structured data collection, an online briefing was held with the heads of Arabic departments from all participating schools. It outlined the essay collection setup and emphasized two key requirements: the writing task had to be conducted in an exam-like setting without access to external materials [D1], and student participation requires *both* parental and student consents.

Pre-visit Setup To facilitate essay collection, we prepared printed sheets, one for each student, con-

نص الموضوع: يلجأ كثير من الشباب إلى السهر لساعات طويلة ليلًا، ما يؤثر في نشاطهم اليومي وأدائهم الدراسي. اكتب مقالًا مكوّنا من ثلاثمائة (٣٠٠) كلمة توضح فيه العوامل التي تدفع الشباب إلى السهر، وما يترتب على ذلك من آثار جسدية ونفسية، مُراعيًا سمات المقال التفسيري، وسلامة اللغة، ومُوظّفًا علامات الترقيم وأدوات الربط بشكلٍ صحيح.

Prompt (English translation): Many young people stay awake late at night, which affects their daily activity and academic performance. Write an essay of approximately 300 words explaining the factors that drive young people to stay up late, and the resulting physical and psychological effects, taking into account the characteristics of explanatory essay features, using correct language, punctuation, and linking words.

الموضوع: لا شك ان ظاهرة السهر عند شباب هذا الجيل مشكلة تؤثر على نشاطهم الصحي و أدائهم الدراسي بشكل سلبي و غير صحي، فقد اثبتت الدراسات ان السهر لاوقاتٍ طويلةٍ للشباب و الاطفال مشكلة تؤثر على صحتهم النفسية و العقلية مما قد يؤدي الى تراجع مستواهم الدراسي و الصحي و تودي الى تقلبات بالمزاج عندهم، و يوجد عدة اسبابٍ و عوامل تدفع الشباب الى السهر لاوقاتٍ متأخرة و منها:

اولًا: التركيز في العاب الفيديو: فشباب هذا الجيل يقضون معظم اوقاتهم في اللهو و اللعب خاصة بالعاب الفيديو، فان العاب الفيديو مستويلةً على عقل شباب هذا الجيل مما جعلهم ينسون العالم الخارجي و يلتهون عن اعمالهم و ينسون الوقت الذي يقضونه على العاب الفيديو مما يؤدي الى تضييع تركيزهم و تلاف عقلهم نتيجة السهر لاوقاتٍ طويلة.

ثانيًا: شرب المنبهات: فكثيرةً ما نرى شباباً باول اعمارهم يشربون المنبهات بشكلٍ مبالغ فيه لكي يستطيعُ السهر و اضافة الوقت، فقد اثبتت الدراسات ان شرب المنبهات من عمر صغيرٍ قد يؤدي الى مشاكل بالقلب او بالصحة و بعض الاحيان تفعل مشاكل اكبر من هذه بكثير.

ثالثًا: استخدام الهواتف المحمولةؤ بشكلٍ مبالغ فيه: فقد اصبح شباب هذا الجيل عتلكون الهواتف النقالة باحدث انواعها و ينسون الوقت بسببها فلا يفرقوا بين الليل و النهار لانهم ... مشغولون بالهواتف و قد نسوا وقتهم و اعمالهم، و ايضاً و قد ذكر في القرآن دليل على مخاطر السهر في قوله تعالى:((و جعلنا الليل لبسا و جعلنا النهار معاشا)).

و ختاماً ارى ان السهر احد اسوء و اشد المشاكل التي قد تمر بها الشباب فهي تدمر العقل و المستقبل و عوامها قد تؤدي احيانا الى الموت، ان السهر لاوقاتٍ طويلة مرض لعينٌ من الصعب التخلص منه. و انا ارى ان السهر احد احقر و اسوء المشاكل التي قد تدمر جيلنا الحالي او الاجيال القادمة و يجب على الجميع ان يتحد بحل الموضوع لانه قد يؤدي الى هلاك المجتمعات.

الدرجات: الصلة بالموضوع: 2 ، الهيكل العام: 5 ، المفردات: 4 ، الأسلوب والتماسك البنائي: 5 ، تطور الأفكار والمضمون: 5 ، الإملاء والترقيم والتنسيق: 4 ، البناء والتراكيب: 4 ، التقييم الكلي: 29

Scores (English translation): Relevance: 2, Organization: 5, Vocabulary: 4, Style: 5, Development: 5, Mechanics: 4, Grammar: 4, Holistic: 29

Figure 2: An example from *LAILA*, containing the prompt (P7) with its English translation, essay (070773) without translation, to preserve Arabic-specific linguistic errors, and annotations (scores) with their English translation.

---

taining one persuasive and one explanatory prompt, along with a unique, pre-generated ID to maintain students' anonymity while enabling metadata tracking [D6]. We also configured a Microsoft Form for digital essay submission, set with a 65-minute time limit to enforce a time-restricted setup [D5].

On-site Visits Data collection spanned one academic year through 27 school visits (24 initial, three follow-ups), by nine team members, to meet the target student count. Each visit, which lasted around six hours, was supported by at least two team members, an IT teacher, and proctoring teachers. During each visit, students received the printed sheets and accessed the online submission form to submit their essays within the allotted time.

Data Cleaning The collected essays then underwent a cleaning process to ensure data quality before the annotation phase. First, we resolved duplicate and misidentified submissions, by correcting ID mismatches through manual verification and name cross-referencing. Next, a manual review eliminated submissions lacking meaningful content, e.g., irrelevant personal content. Lastly, essays with 10 or fewer words were excluded for being insufficient for AES analysis. This approach preserved the data integrity for subsequent processing.

4.3 Data Annotation

The third stage of constructing *LAILA* involved annotating the collected essays through hiring reliable annotators, adopting standardized rubrics, and

developing an annotation guidebook. All annotations were managed and executed using an online platform. Figure 2 shows a sample annotated essay.

Annotators The hired annotation team consisted of 6 annotators and 3 supervisors, all of whom were Arabic language teachers or lecturers with educational backgrounds. Five members of the team hold advanced degrees (MSc or PhD) in the Arabic language. The supervisors oversaw training, quality assurance, and dispute resolution, while the annotators performed the primary scoring tasks.

Annotation Guidelines All essays in *LAILA* dataset were scored using the Core Academic Skills Test (CAST) rubric [D7], developed by the Qatar University Testing Center (QUTC).[10] The rubric covers 7 writing traits [D4]: Relevance (REL), Organization (ORG), Vocabulary (VOC), Style (STY), Development (DEV), Mechanics (MEC), and Grammar (GRA). Additionally, a Holistic score (HOL) was computed as the sum of the 7 trait scores. Six traits (all but REL) were rated on a 6-point scale (0 = lowest, 5 = highest), while REL was rated on a 3-point scale (0 = not relevant, 1 = partially relevant, and 2 = fully relevant). Furthermore, if an essay received a REL score of 0, all other trait scores were set to 0, as irrelevant responses are not subject to further evaluation. This mixed-scale allowed the rubric to capture both

---

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

P#	Type	Topic	Essays	Len (Max)	A3
P1	EXP	Sports	1,122	162 (696)	10.4%
P2	PER	Social Media	1,168	175 (690)	15.5%
P3	PER	Technology	521	159 (643)	9.4%
P4	PER	Communication	500	152 (682)	15.0%
P5	EXP	Heritage	1,181	157 (690)	23.3%
P6	PER	Homework Load	1,162	160 (706)	20.3%
P7	EXP	Staying Up Late	1,143	202 (689)	10.6%
P8	PER	Video Games	1,062	186 (690)	15.7%
Total			**7,859**		

Table 3: Descriptive statistics of *LAILA*. "EXP" and "PER" denote explanatory and persuasive prompts; "Len" is the average word count; "A3" indicates the percentage of essays annotated by a third annotator (supervisor).

Figure 3: Overall score distributions *per trait* in *LAILA*.

broad dimensions of writing proficiency and the degree to which essays addressed the assigned topic.

To ensure consistency, 2 supervisors developed an annotation guidebook[11] with detailed terminology, exemplars, and annotated practices for each prompt type. Annotators were required to review these materials and complete training sessions before formal annotation. Moderation sessions followed, where discrepancies were discussed and interpretations of the rubric were harmonized.

Annotation Process The scoring process was conducted using the Assessment Gourmet Platform, which supports large-scale, anonymized essay scoring.[12] The platform ensured that annotators were blinded to student identity [D6], while only supervisors had access to annotator metadata for monitoring purposes.

The essays were randomly distributed among the annotators to minimize systematic bias, and the scoring sessions were capped to prevent annotators' fatigue. Each essay was independently scored by two annotators across all traits. If the difference in HOL scores between the two annotators was less than 6 points,[13] the mean of the two scores was computed and then rounded down to the nearest integer; this rounded value was adopted as the final score for each trait. However, essays with large discrepancies between annotators (≥ 6 points difference in the HOL score) were flagged and escalated to a supervisor (a third annotator in this case), who provided the final adjudicated score and offered feedback to the annotators to improve alignment and consistency in subsequent batches

---

[13]The 6-point threshold was defined by QUTC experts, corresponding to an average of 1-point discrepancy across the 6 core traits, excluding REL, composing the holistic score.
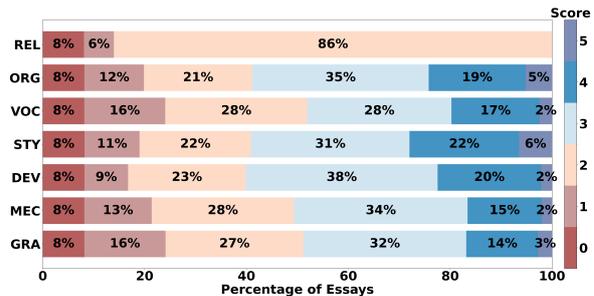
of essays. This multi-layered workflow, calibration, double-scoring, adjudication, and feedback ensured reliable and replicable annotations.

5 Dataset Analysis

LAILA comprises 7,859 essays collected across 8 distinct prompts: 3 explanatory and 5 persuasive (3,446 and 4,413 essays, respectively) [D8]. Table 3 shows that the number of essays per prompt ranges from 500 to 1,181. Essay lengths demonstrate consistency across prompts, with an average of 171 words and a maximum of 706 words.

Score Distribution Figure 3 shows the distribution of the trait scores on all prompts. Most traits follow a near-normal pattern on the 6-point scale, indicating overall positive essay quality. The REL scores show that 86% of the essays agree strongly with the prompt. In contrast, GRA and VOC show the highest percentage of score 1 at 16%, followed by MEC at 13%, highlighting that these traits pose the greatest challenges. Other traits show a more balanced spread across mid-to-high scores, reflecting better performance. Extreme scores (0 or 5) are rare, with STY having the highest share of score 5 at 6%. More details are provided in Appendix B.

Inter-Annotator Agreement (IAA) We compute IAA using Quadratic Weighted Kappa (QWK) (Williamson et al., 2012) to assess agreement between the two main annotators (A1 and A2), prior to adjudication. Table 4 reports QWK per trait and prompt, with agreement strength following Landis and Koch (1977). Overall, IAA was *substantial* across prompts, with an average QWK ranging from 0.66 (P5) to 0.75 (P1), showing strong consistency. However, variability emerges by prompt and trait. In particular, P5 showed the lowest average agreement, noting that it has the highest percentage of essays that require a third annotator (A3: 23.3%; Table 3). This suggests greater initial disagreement

Trait	P1	P2	P3	P4	P5	P6	P7	P8
REL	0.79	0.60	0.67	0.59	0.58	0.68	0.75	0.77
ORG	0.78	0.72	0.83	0.77	0.74	0.78	0.77	0.78
VOC	0.74	0.71	0.76	0.69	0.71	0.75	0.79	0.80
STY	0.74	0.71	0.76	0.72	0.60	0.71	0.68	0.72
DEV	0.79	0.64	0.58	0.66	0.70	0.72	0.71	0.55
MEC	0.69	0.65	0.77	0.72	0.58	0.65	0.69	0.70
GRA	0.71	0.65	0.77	0.70	0.70	0.73	0.67	0.61
AVG	0.75	0.67	0.73	0.69	0.66	0.72	0.72	0.70

Table 4: Inter-Annotator Agreement assessed via QWK across all prompts. Colors indicate strength of agreement as Moderate , Substantial , and Almost Perfect .

between A1 and A2 for P5, potentially due to less precise prompt wording that made consistent interpretation more challenging. Importantly, this does not reflect lower final label quality as such disagreements were flagged and resolved through adjudication by A3, who provided the final scores. Across traits, ORG showed the strongest agreement due to its structural clarity, whereas REL, DEV, and MEC had lower agreement, reflecting greater subjectivity in semantic traits.

6 Dataset Release

We release *LAILA* as an open-access resource for Arabic AES, including writing prompts, essays, scoring rubrics, annotation guidebook, and annotations [D8]. The annotations include scores assigned by the main annotators, the third annotator used for adjudication, and the final adjudicated scores covering both holistic and trait-specific dimensions. Moreover, we provide two predefined data splits aligned with the commonly-applied AES setups: *prompt-specific* and *cross-prompt* [D9]. For the prompt-specific setup, essays of each prompt were randomly split over 5 folds, to support a 5-fold cross-validation setting. For the cross-prompt setup, we support 8-fold leave-one-prompt-out cross-validation setting as follows. For each of the 8 prompts, the target prompt is held out entirely for testing, 2 prompts (1 persuasive and 1 explanatory, selected at random) are designated for development, and the remaining 5 prompts are designated for training. For reproducibility and fair of comparisons, splits are made publicly available.[14]

7 Benchmarking Experiments

To establish performance references, we benchmarked *LAILA* under both prompt-specific and cross-prompt AES settings using strong baselines.

[14] https://gitlab.com/bigirqu/laila

For each setup, we investigate two research questions in the context of Arabic AES: **(RQ1)** how do different categories of models compare in overall performance?, and **(RQ2)** which models achieve the best results across individual traits?.

7.1 Experimental Setup

Model Selection We selected a diverse set of models with varying architectures to establish strong baselines for *LAILA*. The selection criteria included code availability, ease of implementation, and coverage of SOTA Arabic and English models. The models fall into three categories: feature-based (FB), encoder-based, and large language models (LLMs). LLMs were evaluated in zero-shot and few-shot settings, and all models followed a multi-task setup predicting holistic and trait scores jointly. Further details are provided in Appendix C.

Prompt-specific Models We selected 4 FB models: Linear Regression (LR), Random Forest (RF), Extreme Gradient Boosting (XGB), and a feed-forward Neural Network (NN). For encoder-based models, we fine-tuned two pre-trained SOTA systems: AraBERT (Antoun et al., 2020), combined with handcrafted features (Sayed et al., 2025), and AraT5 (Nagoudi et al., 2022), inspired by the strong performance of T5 in English AES (Do et al., 2024).[15] For LLMs, we evaluated three Arabic-centric models: ALLaM (Bari et al., 2025), Command-R7B-Arabic (R7B) (Alnumay et al., 2025), and Fanar (Team et al., 2025), under zero-shot and few-shot (5-shot) settings. For the prompt-specific experiments, we used 5-fold cross-validation using the predefined 5 splits. In each cross-validation iteration, one fold (20%) was used as the test set, and the remaining four folds were further split into training (70%) and development (10%) sets, with stratification applied to maintain consistent prompt distributions and ensure consistent evaluation across models.

Cross-prompt Models We selected the same FB models and LLMs. For encoder-based models, we employ the same AraBERT-based architecture, which performed strongly in Arabic AES (Sayed et al., 2025), along with two SOTA English AES models, ProTACT (Do et al., 2023) and MOOSE (Chen et al., 2025).

[15] https://huggingface.co/aubmindlab/bert-base-arabert, https://huggingface.co/UBC-NLP/AraT5v2-base-1024

Setup	Model	REL	ORG	VOC	STY	DEV	MEC	GRA	HOL	Avg^{-H}	Avg
Zero-shot	ALLaM	**0.242**	0.197	0.319	0.309	0.335	<u>0.341</u>	<u>0.312</u>	0.365	0.294	0.303
	R7B	**0.242**	0.348	0.350	**0.417**	0.340	0.150	0.292	0.367	0.306	0.313
	Fanar	<u>0.145</u>	**0.424**	**0.429**	<u>0.388</u>	**0.411**	**0.378**	**0.415**	**0.481**	**0.370**	**0.384**
Prompt-specific	LR	0.436	0.694	0.722	0.710	0.650	0.661	0.687	0.765	0.651	0.665
	RF	0.519	0.704	0.734	0.718	0.670	0.658	0.694	0.775	0.671	0.684
	XGB	0.522	0.716	0.743	0.731	0.685	0.686	0.711	0.791	0.685	0.698
	NN	0.500	0.723	0.758	0.743	0.699	0.699	0.727	0.797	0.693	0.706
	AraT5	**0.612**	0.731	0.759	0.759	0.717	0.709	<u>0.719</u>	0.787	0.715	0.724
	AraBERT	<u>0.587</u>	**0.761**	**0.776**	**0.768**	**0.731**	**0.724**	**0.744**	**0.833**	**0.727**	**0.740**
	ALLaM (5)	0.260	0.391	0.348	0.415	0.386	0.355	0.369	0.439	0.361	0.370
	R7B (5)	0.337	0.472	0.383	0.482	0.399	0.407	0.431	0.496	0.416	0.426
	Fanar (5)	0.490	0.559	0.562	0.614	0.575	0.572	0.580	0.644	0.565	0.575
Cross-prompt	LR	0.360	0.621	0.633	<u>0.643</u>	0.573	0.585	0.616	0.661	0.576	0.586
	RF	0.331	0.609	**0.644**	0.637	0.573	0.559	0.609	**0.682**	0.566	0.581
	XGB	0.360	**0.645**	0.641	0.641	0.583	0.577	<u>0.619</u>	0.679	<u>0.581</u>	0.593
	NN	0.353	0.609	0.621	0.631	0.566	0.565	0.597	0.651	0.563	0.574
	ProTACT	0.355	0.493	0.505	0.485	0.501	0.476	0.498	0.578	0.473	0.486
	AraBERT	0.322	0.596	0.604	0.593	0.529	0.546	0.585	0.620	0.539	0.549
	MOOSE	**0.411**	<u>0.627</u>	<u>0.642</u>	**0.649**	<u>0.585</u>	<u>0.586</u>	**0.623**	0.649	**0.589**	**0.597**
	ALLaM (5)	0.131	0.237	0.213	0.287	0.226	0.249	0.251	0.281	0.228	0.234
	R7B (5)	0.263	0.487	0.391	0.481	0.424	0.425	0.442	0.517	0.416	0.429
	Fanar (5)	<u>0.409</u>	0.580	0.541	0.623	**0.588**	**0.592**	0.587	0.669	0.560	0.574

Table 5: Average QWK performance over all prompts for each trait under the zero-shot, prompt-specific, and cross-prompt setups. Avg^{-H} denotes average performance without HOL scoring. **Bold** and <u>underlined</u> values indicate the best and second-best performance per trait, respectively, marked separately within each setup.

Feature Set We implemented the 816 hand-crafted features introduced by Sayed et al. (2025) for Arabic AES, encompassing surface, readability, lexical, semantic, and syntactic aspects. These features were applied to the FB models, ProTACT, MOOSE, and AraBERT. To mitigate noise, we performed feature selection based on Pearson and Spearman correlations (Li and Ng, 2024), retaining features whose absolute correlation with any trait exceeded a predefined threshold for either metric. For MOOSE, however, we did not apply this explicit feature selection step.

Evaluation & Hyperparameters We evaluate model performance using QWK to measure agreement between human and model scores. Hyperparameters are tuned via Bayesian optimization with the Tree-structured Parzen Estimator algorithm (Bergstra et al., 2011), implemented using Optuna's TPESampler.[16] We ran 20 trials with 5 startup trials and a fixed random seed of 11. The best configuration, selected based on average QWK on the development set, was used for final evaluation on the test data. Additional hyperparameter details are provided in Appendix C.

---
[16]Optuna TPESampler documentation

7.2 Results and Discussions

We discuss the results of our benchmarking experiments for the two AES setups.

7.2.1 Prompt-specific Results

Performance Across Categories (RQ1) Table 5 summarizes performance differences across model categories. Among FB models, more complex architectures performed better, with NN leading, followed by XGB, RF, and LR. Both encoder-based models, AraT5 and AraBERT, outperform all other baselines, with AraBERT achieving SOTA performance (average QWK: 0.74) and AraT5 trailing by 1.6 points, highlighting their effectiveness in capturing prompt-specific patterns. For LLMs, adding prompt-specific few-shot examples improved the performance acorss all models. However, they still lagged behind, with the best variant, Fanar (5), scoring 16.5 points below AraBERT.

Performance Across Traits (RQ2) Encoder-based models consistently outperform other categories across traits, with AraBERT achieving the highest scores in 7 of 8 traits and AraT5 leading in REL. Their advantage over FB models is most pronounced in REL, where AraT5 and AraBERT exceed the best FB model by 9 and 6.5 points, respectively, demonstrating strength in modeling contextual and semantic relationships within essays.

7.2.2 Cross-prompt Results

Performance Across Categories (RQ1) FB models showed comparable performance, with only a 2-point difference between the top performer (XGB) and the lowest (NN), highlighting consistency of these feature-based approaches. LR's strong results further highlight the robustness and the quality of the feature set. Among encoder-based models, ProTACT exhibited the lowest performance overall, suggesting limited transferability from its English architecture, while AraBERT trailed XGB by 4.4 points. Notably, MOOSE achieved the highest performance among all evaluated cross-prompt models. For LLMs, few-shot prompting improved Fanar and R7B (+19 and +11.6 points), whereas ALLaM dropped by 7 points, indicating high sensitivity to prompt design.

Performance Across Traits (RQ2) No single model consistently outperformed the others across individual traits. MOOSE performed best in REL, STY, and GRA traits. For the remaining traits, RF led in VOC and HOL, XGB in ORG, and Fanar (5) in DEV and MEC.

7.2.3 Discussion

The performance of models differ significantly between prompt-specific and cross-prompt setups. *Encoder-based models* excel in prompt-specific tasks, approaching the IAA (Table 4), which indicates a close reach to the human performance. For instance, AraBERT exhibits the best performance overall in the prompt-specific setup with an average QWK performance of 0.74. However, its cross-prompt performance drops significantly (average QWK: 0.549), underscoring its ability to capture prompt-dependent patterns rather than generalizable features. We note that the applicability of the prompt-specific setup is relatively limited due to reliance on target prompt essays that are labeled, which is often unavailable in practice. In the cross-prompt setup, while MOOSE achieves the highest average performance (average QWK: 0.597), we observe that classical FB models (XGB) remain remarkably competitive. This confirms that handcrafted features capture prompt-independent linguistic properties that generalize better in the more challenging and realistic cross-prompt setup than some standard encoders. LLMs, specifically Fanar and R7B with few shots, exhibit consistent performance across both setups, indicating they capture general scoring criteria. Their relative underperformance in prompt-specific tasks reflects their limited ability to exploit prompt-dependent patterns without fine-tuning.

Overall, the findings of the benchmarking experiments above highlight the need to prioritize future research on cross-prompt AES, which is practically closer to the real world, while best models are still far from human performance.

8 Conclusion

While research on automated scoring of English essays began more than 55 years ago, Arabic essay scoring is hindered by the lack of data resources. To bridge this gap, this paper introduced *LAILA*, the first large-scale publicly available dataset for Arabic AES. The dataset comprises 7,859 essays written on 8 different prompts by 4,372 high school students, and provides annotations for 7 writing traits as well as a holistic score with *substantial* inter-annotator agreement. This makes it a *comprehensive* and *reliable* resource for training models and evaluating writing quality of Arabic essays. For reproducibility, we detailed the data collection and annotation process. We also benchmarked *LAILA* using SOTA Arabic and English AES models in both prompt-specific and cross-prompt settings, showing strong baselines for future research.

Limitations

While *LAILA* represents a significant step forward for Arabic AES, it has notable limitations. First, *LAILA* was collected from students in a single country, Qatar, and specific grade levels, grades 10 to 12,

which may limit its generalizability to other Arabic-speaking populations due to diverse educational systems. However, this concern is partially mitigated by Qatar's large population of Arab expatriates, resulting in student participants representing a wide variety of Arabic dialects and backgrounds. Second, *LAILA* includes only explanatory and persuasive writing prompts, limiting genre diversity and potentially affecting model robustness across other styles, such as narrative or descriptive writing. Third, two of the prompts (Prompts 3 and 4) have fewer essays than the rest, which could introduce imbalance and lead to skewed performance in prompt-specific evaluations. Fourth, the relevance trait is scored on a narrow scale from 0 to 2, which limits the granularity of evaluation and may cause QWK to underrepresent subtle differences in model predictions. This restricted range likely contributes to the lower performance scores observed for the REL trait compared to others. Fifth, with an average essay length of only 171 words, the dataset lacks extended writing samples, restricting the assessment of models on long-form academic tasks requiring more complex, extended compositions. Finally, we acknowledge that the experimental results reported in Table 5 do not aim to exhaustively benchmark state-of-the-art AES models; rather, they serve as baseline performance for *LAILA*. We emphasize that the primary contribution of this work lies in the construction and public release of a large-scale Arabic AES dataset, rather than in proposing novel AES model architectures.

Ethical Considerations

The collection of *LAILA* received approval from an Institutional Review Board (IRB Number: QU-IRB 159/2024-EA) and was conducted with informed consent from both students and their legal guardians. Essays were written under exam-style supervised conditions, with participation entirely voluntary and independent of students' academic evaluations or grades. To protect participant privacy, all submissions were assigned pre-generated anonymous identifiers, and personally identifiable information was systematically removed during data cleaning. Annotation was conducted on a platform that prevented annotators from accessing student identities or demographic information. Essay prompts were carefully designed to be developmentally appropriate for the target age groups, culturally relevant to the Arabic-speaking context,

and free of sensitive topics or potentially harmful content. The dataset includes balanced gender representation and covers grades 10-12.

All annotators were qualified Arabic language educators who received fair compensation at or above local professional rates. They completed structured training on annotation guidelines, and inter-annotator disagreements were resolved through systematic adjudication procedures. The dataset will be released exclusively for research and educational purposes in Arabic automated essay scoring. We explicitly prohibit its use for high-stakes educational decisions affecting individual students. Users must commit to: (1) preventing re-identification attempts, (2) avoiding demographic profiling or inference, and (3) refraining from deployment in operational assessment contexts without independent validation. We recommend that any system developed using *LAILA* undergo fairness auditing and stakeholder consultation before real-world application.

References

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.

Saeda A Al Awaida, Bassam Al-Shargabi, and Thamer Al-Rousan. 2019. Automated Arabic essay grading system based on f-score and Arabic worldnet. *Jordanian Journal of Computers and Information Technology*, 5(3).

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.

Yazeed Alnumay, Alexandre Barbet, Anna Bialas, William Darling, Shaan Desai, Joan Devassy, Kyle Duffy, Stephanie Howe, Olivia Lasche, Justin Lee, Anirudh Shrinivason, and Jennifer Tracey. 2025. Command R7B Arabic: A small, enterprise focused, multilingual, and culturally aware Arabic LLM.

Mohammad Alobed, Abdallah M M Altrad, and Zainab Binti Abu Bakar. 2021a. A comparative analysis of euclidean, jaccard and cosine similarity measure and Arabic wordnet for automated Arabic essay scoring.

In *2021 Fifth International Conference on Information Retrieval and Knowledge Management (CAMP)*, pages 70–74.

Mohammad Alobed, Abdallah MM Altrad, Zainab Binti Abu Bakar, and Norshuhani Zamin. 2021b. Automated Arabic essay scoring based on hybrid stemming with WordNet. *Malaysian Journal of Computer Science*, pages 55–67.

Abeer Alqahtani and Amal Alsaif. 2020. Automated Arabic essay evaluation. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 181–190, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Waleed Alsanie, Mohamed I Alkanhal, Mohammed Alhamadi, and Abdulaziz O Alqabbany. 2022. Automatic scoring of Arabic essays over three linguistic levels. *Progress in Artificial Intelligence*, pages 1–13.

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.

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham Abdullah Alyahya, Sultan AlRashed, Faisal Abdulrahman Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Saad Amin Hassan, Dr. Majed Alrubaian, Ali Alammari, Zaki Alawami, Abdulmohsen Al-Thubaity, Ahmed Abdelali, Jeril Kuriakose, Abdalghani Abujabal, Nora Al-Twairesh, Areeb Alowisheq, and Haidar Khan. 2025. ALLaM: Large language models for Arabic and English. In *The Thirteenth International Conference on Learning Representations*.

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, Suzhou, China. Association for Computational Linguistics.

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, Bangkok, Thailand. Association for Computational Linguistics.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A corpus of non-native English. *ETS Research Report Series*, 2013(2):i–15.

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Stindlová, and Chiara Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In *LREC*, pages 1281–1288. Reykjavik, Iceland.

Po-Kai Chen, Bo-Wei Tsai, Shao Kuan Wei, Chien-Yao Wang, Jia-Ching Wang, and Yi-Ting Huang. 2025. Mixture of ordered scoring experts for cross-prompt essay trait scoring. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18071–18084, Vienna, Austria. Association for Computational Linguistics.

Scott Crossley, Yu Tian, Perpetual Baffour, Alex Franklin, Youngmeen Kim, Wesley Morris, Meg Benner, Aigner Picou, and Ulrich Boser. 2023a. The English language learner insight, proficiency and skills evaluation (ELLIPSE) corpus. *International Journal of Learner Corpus Research*, 9(2):248–269.

Scott Andrew Crossley, Perpetual Baffour, Yu Tian, Alex Franklin, Meg Benner, and Ulrich Boser. 2023b. A large-scale corpus for assessing written argumentation: PERSUADE 2.0. *Available at SSRN 4795747*.

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, St. Julian's, Malta. Association for Computational Linguistics.

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, Toronto, Canada. Association for Computational Linguistics.

Yomna Elsayed, Emad Nabil, Marwan Torki, Safiullah Faizullah, and Ayman Khalafallah. 2025. ZaQQ: A new Arabic dataset for automatic essay scoring via a novel human–ai collaborative framework. *Data*, 10(9):148.

Marwa M. Gaheen, Rania M. ElEraky, and Ahmed A. Ewees. 2020. Optimized neural network-based improved multiverse optimizer algorithm for automated Arabic essay scoring. *International Journal of Scientific & Technology Research*, 9:238–243.

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*.

Rayed Ghazawi and Edwin Simpson. 2025. How well can llms grade essays in Arabic? *arXiv preprint arXiv:2501.16516*.

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.

Yaqiong He, Feng Jiang, Xiaomin Chu, and Peifeng Li. 2022. Automated Chinese essay scoring from multiple traits. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3007–3016, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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

Shengjie Li and Vincent Ng. 2024. Conundrums in cross-prompt automated essay scoring: Making sense of the state of the art. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7661–7681, Bangkok, Thailand. Association for Computational Linguistics.

Rim Aroua Machhout and Chiraz Ben Othmane Zribi. 2024. Enhanced BERT approach to score Arabic essay's relevance to the prompt. *Communications of the IBIMA*, 2024.

Somaia Mahmoud, Emad Nabil, and Marwan Torki. 2024. Automatic scoring of Arabic essays: A parameter-efficient approach for grammatical assessment. *IEEE Access*.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Abu Bakr Mohammad, Kareem Eissa, and Samhaa El-Beltagy. 2017. AraVec: A set of Arabic word embedding models for use in Arabic NLP. *Procedia Computer Science*, 117:256–265.

Behrang Mohit, Alla Rozovskaya, Nizar Habash, Wajdi Zaghouani, and Ossama Obeid. 2014. The first QALB shared task on automatic text correction for Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 39–47, Doha, Qatar. Association for Computational Linguistics.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. AraT5: Text-to-text transformers for Arabic language generation.

In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647, Dublin, Ireland. Association for Computational Linguistics.

Ellis B Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

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.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, Majd Hawasly, Mus'ab Husaini, Soon-Gyo Jung, Ji Kim Lucas, Walid Magdy, Safa Messaoud, Abubakr Mohamed, Tasnim Mohiuddin, Basel Mousi, Hamdy Mubarak, Ahmad Musleh, Zan Naeem, Mourad Ouzzani, Dorde Popovic, Amin Sadeghi, Husrev Taha Sencar, Mohammed Shinoy, Omar Sinan, Yifan Zhang, Ahmed Ali, Yassine El Kheir, Xiaosong Ma, and Chaoyi Ruan. 2025. Fanar: An Arabic-centric multimodal generative AI platform.

Rodrigo Wilkens, Alice Pintard, David Alfter, Vincent Folny, and Thomas François. 2023. TCFLE-8: a corpus of learner written productions for French as a foreign language and its application to automated essay scoring. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3447–3465, Singapore. Association for Computational Linguistics.

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.

A Grading Rubric

For annotating *LAILA*, we adopted the rubric from the Core Academic Skills Test (CAST), which is provided in Arabic, developed by the Qatar University Testing Center (QUTC).[17] This rubric guided scoring across 7 writing traits: relevance (REL), organization (ORG), vocabulary (VOC), style (STY), development (DEV), mechanics (MEC), and grammar (GRA). An English-translated version of the CAST rubric is presented in Table 6.

---

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

Trait	1	2	3	4	5
REL الصلة بالموضوع	Partially relevant to the topic	Completely relevant to the topic			
ORG الهيكل العام	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 المفردات	Use of a limited range of vocabulary and phrases that do not make sense together, with repetition and lexical errors, and generally inappropriate vocabulary that obscures meaning.	Use of a basic range of vocabulary, with repetition, lexical errors, and many inappropriate 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 meaning.	Use of a good and appropriate range of vocabulary with few lexical errors, inappropriate choices without affecting meaning, and occasional use of idiomatic expressions.	Use of a broad, correct, and appropriate range of vocabulary with few errors, showing good knowledge of idiomatic expressions and awareness of implicit levels of meaning.
STY الأسلوب والتماسك البنائي	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, appropriate use of different organizational patterns, and a range of structural cohesion devices.	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 تطور الأفكار والمضمون	Content is not related to the subject; ideas are random and lack coherence, sequence, and evidence.	Content is somewhat related; ideas are sequential, but the main idea disappears during writing, limited coverage, and poor use of supporting structures.	Content is completely related; ideas mostly follow sequence, the main idea gradually disappears, some evidence is present but disorganized.	Content is completely related; ideas are clear, organized, coherent, with the main idea connected to sub-ideas, a specific position adopted, some arguments and evidence presented coherently.	Content is completely related; ideas are clear, organized, coherent, the main idea connected to sub-ideas, a specific position adopted, arguments and evidence presented coherently, comprehensive coverage of opinions, and use of various persuasive methods.
MEC الإملاء والترقيم والتنسيق	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 pen slips.
GRA البناء والتراكيب	Use a limited set of simple grammatical structures and sentence patterns with little flexibility or precision.	Correct use of some simple structures with frequent systematic errors that may obscure meaning.	Use a variety of grammatical structures, with notable errors that can sometimes obscure meaning.	Good use of variety of structures with rare errors and minor imperfections that do not affect meaning.	Always correct and flexible use of a wide variety of grammatical constructions with minor slips.
Note: A score of 0 is assigned if the student does not attempt the task, provides a response that falls below the performance level described for score 1, or submits content that is not relevant to the topic of the given prompt.					

Table 6: CAST Persuasive/Argumentative Writing Rubric - English Translation (Bashendy et al., 2024).

Figure 4: Score distributions across prompts and traits in *LAILA*.

B Detailed Dataset Analysis

In this section, we present a detailed analysis of score distributions and essay lengths in *LAILA*.

B.1 Score Distributions

Figure 4 presents the score distributions in *LAILA* across 8 distinct prompts (P1–P8) and 7 writing traits, including Relevance (REL), Organization (ORG), Vocabulary (VOC), Style (STY), Development (DEV), Mechanics (MEC), and Grammar (GRA), alongside a Holistic (HOL) score representing the sum of all individual traits. All traits are rated on a scale of 0–5, except REL, which ranges from 0–2. Overall, most traits approximate a normal distribution with comparable patterns across prompts, suggesting that essays' quality remains relatively stable regardless of the prompt topic. However, variations emerge by prompt and trait.

For REL, distributions are consistently skewed toward score 2 (the highest) across all prompts, indicating that most students successfully addressed the assigned task and produced strongly relevant essays. However, P1 and P5 display relatively higher frequencies at score 0, compared to other prompts,

suggesting potential ambiguity in prompt wording that may have limited clear interpretation. Notably, both P1 and P5 are explanatory prompts, which often allow greater flexibility in structure and content. This openness may have led to varied interpretations among students and inconsistencies in alignment with the intended prompt focus.

The distribution of ORG scores varies across prompts but consistently peaks at score 3. P2 exhibits a pronounced peak at 3, whereas P3 and P4 show a wider spread toward lower scores, suggesting a weaker organizational structure in these essays. P5 and P6 display nearly identical, perfectly bell-shaped distributions, while P7 and P8 are skewed toward higher scores, reflecting stronger organization. Similarly, the VOC and STY traits generally follow comparable patterns, clustering around scores 2–3 with a common peak at 3, indicating that these dimensions present moderate challenges for most students.

The DEV, MEC, and GRA traits exhibit broadly comparable distributions across all prompts, though with subtle variations. The DEV trait shows distinct peaks at scores 2, 3, and 4, suggesting that

3087

the depth and elaboration of ideas vary depending on the prompt. In contrast, the MEC and GRA traits display closely aligned distributions, mirroring each other across prompts, with their peaks typically occurring at either score 2 or 3. This close alignment suggests a strong association between grammatical accuracy and mechanical correctness, as essays with better grammar tend to exhibit stronger adherence to writing conventions.

The HOL score, scaled up to 32, representing the sum of individual trait scores, provides a comprehensive measure of overall performance. Its distribution roughly approximates a bell curve, reflecting an overall normal pattern. However, prompt-specific effects are evident: P1, P3, P4, and P6 exhibit roughly centered distributions, whereas P2, P5, P7, and P8 are skewed toward higher scores, suggesting that students generally performed better on these prompts. This pattern indicates that certain prompts may have been easier or allowed for clearer demonstration of writing skills, while others elicited more variable performance. Also, it is worth noting that the scarcity of top scores (24+ holistically) suggests a ceiling effect, where achieving excellence demands strong performance across all dimensions.

Finally, the close uniform patterns across traits and prompts indicate a balanced scoring process, highlighting the robustness of the rubric and minimal influence from prompt variation or annotator bias. This consistency supports the development of generalized and robust Arabic AES models.

B.2 Essay Length

Figure 5 presents the distribution of essay lengths (in words) across all prompts in *LAILA* dataset, which contains a total of 7,859 essays. The data exhibits a unimodal pattern centered around an average essay length of 171 words, with the majority of essays concentrated between 90 and 210 words. A pronounced peak occurs in the 150 to 180 word range, where the count reaches its highest at 1,065 essays. Shorter essays are also represented, with 1,061 essays falling between 11 and 60 words. Notably, the minimum observed length is 11 words, which results from the data cleaning step that excluded essays containing 10 words or fewer. Beyond 420 words, the frequency drops sharply, with only a few essays exceeding 500 words, suggesting that longer essays are less common in the dataset. Overall, this distribution provides a clear view of
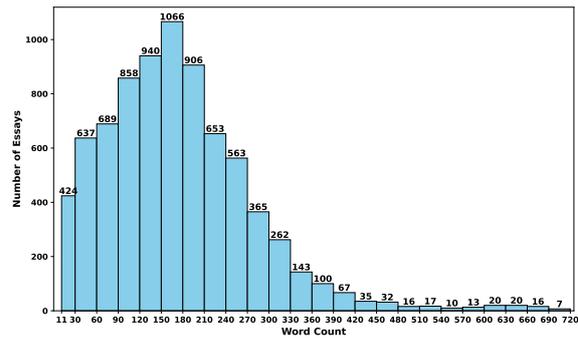

Figure 5: Distribution of essay lengths in *LAILA*.

the typical essay length within *LAILA* dataset, highlighting a strong central tendency and a sharp drop-off at the extremes.

C Implementation Details

In this section, we provide all the details of implementation and hyperparameter tuning for all the models used in the benchmarking experiments. Model-specific parameters are summarized in Table 7. All experiments were conducted on two machines: one equipped with an NVIDIA RTX A6000 GPU, and another with two NVIDIA A10 GPUs.

C.1 Feature-based Models

For the feature-based (FB) models, the architectures and hyperparameter configurations were kept consistent across both the *prompt-specific* and *cross-prompt* setups. We used the sklearn library[18] for Linear Regression (LR) and Random Forest (RF) models, and the XGBoost library[19] for the Extreme Gradient Boosting (XGB) model. The Neural Network (NN) model followed the architecture described by Li and Ng (2024), consisting of two hidden layers with ReLU activations and a sigmoid output layer, implemented in PyTorch. For the NN, the number of epochs was fixed at 50 with early stopping, setting patience to 7, and checkpointing the best epoch during hyperparameter tuning for final model training.

For all the FB models, feature selection was performed on the training set during hyperparameter tuning, with threshold values of [0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6], where 0 corresponds to using all features.

---

[18] https://scikit-learn.org/
[19] https://xgboost.readthedocs.io/en/stable/

Model	Hyperparameter Name	Value
RF	Max depth	[3-10] with a step of 1
	Max features	[0.1-0.9] with a step of 0.1
	Max samples	[0.1-0.9] with a step of 0.1
XGB	Max depth	[3, 6, 9]
	Learning rate	[0.01, 0.1, 0.2]
	Subsample	[0.8, 0.9, 1.0]
	feature subsampling rate per tree	[0.8, 0.9, 1.0]
	Number of estimators	[100, 200, 300]
NN	Hidden layer widths	[64, 128, 256]
	Dropout rate	0.3
	Learning rate	[1e-5, 1e-4, 1e-3]
	Batch size	16
AraT5	Learning rate	[5e-5, 1e-4, 2e-4]
	Batch size	4
AraBERT	Batch size	8
	Encoder learning rate	[1e-6, 2e-6, 5e-6]
	Dense-layers learning rate	[5e-5, 1e-4, 2e-4]
	Number of trainable layers	[2, 4, 8, 'all']
ProTACT	Learning rate	[1e-5, 1e-4, 1e-3]
	Trait similarity loss threshold	[0.5, 0.6, 0.7]
	Batch size	16
	Alpha	0.7
	LSTM units	32
	Self-attention heads	4
	CNN filters	100
	CNN kernel size	3
	Dropout rate	0.5
MOOSE	Learning rate	[1e-4, 2e-5]
	Hidden dimensions	256
	Batch size	8
	Weight decay	0.001
	Alpha	0.7
	Epochs	14
	Chunk sizes	[10,30,90,130]
	Dropout rate	0.5

Table 7: Model-specific hyperparameters

C.2 Encoder-based Models

C.2.1 AraT5

Motivated by the strong performance of the T5 model in *prompt-specific* English AES, we fine-tuned the AraT5 model using a similar setup to the ArTS model (Do et al., 2024). The AES task is formulated as a text generation problem, where the model predicts the trait scores sequentially. The input to the model consisted of an instruction to score the essay along with the essay text, and the model was fine-tuned to generate the trait names followed by their corresponding scores. The prediction order of the traits was aligned with the rubric to reflect the scoring order followed by human annotators.

Our setup differs from ArTS in data splitting: while their training and testing data included essays from all prompts (with a single model trained on all prompts), we adopted the conventional prompt-specific setup, training a separate model for each prompt using our predefined dataset splits.

For training, we used Seq2SeqTrainer[20] from the transformers library. For hyperparameter tuning, we explored the learning rates in the range [5e-5, 1e-4, 2e-4]. Other hyperparameters followed the configuration of (Do et al., 2024), with a batch size of 4 and a maximum of 15 epochs. Early stopping was applied based on the average QWK score on the development set.

C.2.2 AraBERT

AraBERT is used as a baseline for the *prompt-specific* and *cross-prompt* setups using the same architecture and hyperparameter search space. The model is fine-tuned with a custom regression head and combined with handcrafted features for trait scoring. The essay and prompt were provided as input to the encoder, with the max-pooled token representation concatenated with the handcrafted features. The resulting vector was then fed into eight parallel regression heads, each corresponding

---

[20]https://huggingface.co/docs/transformers/en/main_classes/trainer#transformers.Seq2SeqTrainer

to a single trait score. Sigmoid activation is used at the output layer to produce values in the range [0, 1], which were subsequently rescaled to the appropriate range for each trait.

For hyperparameter tuning, the learning rates for the encoder layers and the regression head are tuned separately to better accommodate the larger dataset, with a weight decay of 0.01. In addition, we tuned the number of trainable layers, with values of [2, 4, 8, all], where 'all' corresponds to training all encoder layers. For the feature selection threshold, we used the same search space as that used with the FB models. These configurations were considered better suited to the dataset, mitigating training instabilities.

C.2.3 ProTACT

For ProTACT, we used the official implementation released by the authors.[21] Essay representations were constructed using CNNs and LSTMs over Part-of-speech (POS) embeddings, while prompt representations combined POS and pre-trained GloVe embeddings. A multi-head attention mechanism was applied to obtain prompt-aware essay representations, which were then concatenated with handcrafted features and passed through a linear layer for scoring. The architecture was adapted for Arabic by replacing GloVe with AraVec embeddings (Mohammad et al., 2017).[22] POS embeddings were extracted using Camel Tools.[23]

For hyperparameter tuning, we adopted the same search spaces for learning rate and feature selection threshold as in the NN model, while additionally tuning the trait similarity loss threshold. All other parameters were kept consistent with those reported in the original study, including embedding dimension, maximum essay length, maximum prompt length, LSTM units, self-attention heads, CNN filters, and kernel size.

C.2.4 MOOSE

For MOOSE, we used the official implementation released by the authors.[24] The architecture was adapted for Arabic by replacing BERT with AraBERT (Antoun et al., 2020). Specifically, in the preprocessing stage, AraBERTv02[25] was used only for tokenization and input encoding, while

---

[21]https://github.com/doheejin/ProTACT
[22]https://github.com/bakrianoo/aravec
[23]https://camel-tools.readthedocs.io/
[24]https://github.com/antslabtw/MOOSE-AES
[25]https://huggingface.co/aubmindlab/bert-base-arabertv02

سيتم إعطاؤك مقال كتب ردًا على الموضوع المعطى. مهمتك هي تقييم جميع المعايير التالية للمقال.

—

الموضوع: موضوع المقال
المقال: المقال المراد تقييمه.
الدرجات: الرجاء إعطاء الدرجات لجميع المعايير بهذا الشكل: الصلة بالموضوع: 0-2 ، الهيكل العام: 0-5 ، المفردات: 0-5 ، الأسلوب والتماسك البنائي: 0-5 ، الأفكار والمضمون: 0-5 ، الإملاء والترقيم والشكل: 0-5 ، البناء والتراكيب: 0-5

—

الموضوع : هل تتفق أو تختلف جعلت الهواتف ورسائل البريد ...
المقال : إن مصطلح التكنولوجيا ...
الدرجات : الهيكل العام: 3.0 ، المفردات: 3.0 ، الأسلوب والتماسك البنائي: 3.0 ، الأفكار والمضمون: 3.0 ، الإملاء والترقيم والشكل: 3.0 ، البناء والتراكيب: 3.0 ، الصلة بالموضوع: 2

—

الموضوع : على الرغم من أهمية وسائل التواصل الاجتماعي ...
المقال : لا شك ان الافراط في استخدام وسائل التواصل ...
الدرجات : الهيكل العام: 1.0 ، المفردات: 2.0 ، الأسلوب والتماسك البنائي: 2.0 ، الأفكار والمضمون: 2.0 ، الإملاء والترقيم والشكل: 1.0 ، البناء والتراكيب: 1.0 ، الصلة بالموضوع: 1

—

الموضوع: بات إهتمام وحماس المراهقين لتعلُّم رياضةٍ جديدة ...
المقال: الصحة والجسم السليم من نعم الله على الإنسان ...
الدرجات:

Figure 6: An example of the LLM-prompt used in the few-shot setup, containing the base instructions, the input format, the 5-shot examples, and the input essay for scoring. For the zero-shot setup, the 5-shot examples are removed, and all trait rubrics are included instead.

the finetuned model for the scoring task used AraBERTv01.[26] This design choice was guided by empirical validation: we experimented with using AraBERTv01 and AraBERTv02 consistently for both preprocessing and modeling, as well as with mixed configurations, and found that using AraBERTv02 for tokenization and input encoding while finetuning AraBERTv01 for scoring yielded the best performance on the development set, thus it was adopted. For the incorporated features, We implemented the handcrafted features introduced by Sayed et al. (2025) for Arabic AES, without applying any feature selection step.

For hyperparameter tuning, the learning rate was tuned over the values 1e-4 and 2e-5, while all other parameters were kept consistent with those reported in the original study. Training was performed with a batch size of 8 for a maximum of 14 epochs, and no early stopping was applied.

C.3 LLMs

The selection of the LLMs was based on the Open Arabic LLM Leaderboard,[27] where we selected

---

[26]https://huggingface.co/aubmindlab/bert-base-arabertv01
[27]Open-Arabic-LLM-Leaderboard

the top 3 Arabic-centric open-weight models, at the time of the experiments, with a number of parameters < 10B. The selected models are: Fanar,[28] Command R7B Arabic,[29] and ALLaM.[30]

We evaluated the models under zero-shot and few-shot prompting. Figure 6 illustrates the prompt adopted for our LLM scoring experiments. In the zero-shot setting, the LLM is tasked with generating trait scores in JSON format, given the prompt text, essay, and trait rubrics. The holistic score was computed as the sum of the individual trait scores. In the few-shot setting, the rubric was omitted, as the meaning of the scores could be inferred from the examples. The number of examples was fixed to 5, balancing scoring context with the context length limit (4096 tokens). When a prompt exceeded this limit, we iteratively removed examples until it fit. Although Command-R7B supports a larger context length, we fixed the number of examples to 5 to ensure fair comparison across LLMs.

In the prompt-specific setup, few-shot examples are selected from the same prompt. In the cross-prompt setup, examples are selected from different source prompts, ensuring that each comes from a distinct training prompt to expose the model to varied contexts. To account for variability in example selection, each experiment was repeated three times with random seeds 1, 11, and 42, and we report the average performance across runs. For all the LLM experiments, we used vLLM[31] for inference with the outlines library[32] to enforce the JSON output format.