TAQEEM 2025: Overview of The First Shared Task for Arabic Quality Evaluation of Essays in Multi-dimensions

May Bashendy
Qatar University
ma1403845@qu.edu.qa

Salam Albatarni
Qatar University
sa1800633@qu.edu.qa

Sohaila Eltanbouly
Qatar University
se1403101@qu.edu.qa

Walid Massoud
Qatar University
wmassoud@qu.edu.qa

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

Tamer Elsayed
Qatar University
telsayed@qu.edu.qa

Abstract

Automated Essay Scoring (AES) has emerged as a significant research problem in natural language processing, offering valuable tools to support educators in assessing student writing. Motivated by the growing need for reliable Arabic AES systems, we organized the first shared Task for Arabic Quality Evaluation of Essays in Multi-dimensions (*TAQEEM*) held at the ArabicNLP 2025 conference. *TAQEEM* 2025 includes two subtasks: Task A on holistic scoring and Task B on trait-specific scoring. It introduces a new (and first of its kind) dataset of 1,265 Arabic essays, annotated with *holistic* and *trait-specific* scores, including relevance, organization, vocabulary, style, development, mechanics, and grammar. The main goal of *TAQEEM* is to address the scarcity of standardized benchmarks and high-quality resources in Arabic AES. *TAQEEM* 2025 attracted 11 registered teams for Task A and 10 for Task B, with a total of 5 teams, across both tasks, submitting system runs for evaluation. This paper presents an overview of the task, outlines the approaches employed, and discusses the results of the participating teams.

1 Introduction

Automated Essay Scoring (AES) systems automatically assess the writing quality of essays, providing holistic scores, trait-specific (i.e., multi-dimensional) scores, or both. Effective AES systems have brought benefits, such as saving teachers time and effort, and producing less-biased and consistent results. This is crucial in large-scale assessments, such as international exams with thousands of participants, making AES a high-stakes application (Burstein, 2013).

There are two AES paradigms: *prompt-specific* and *cross-prompt*. The dominant *prompt-specific* AES trains and tests models on essays from the same prompt[1] (Taghipour and Ng, 2016). This setup achieves high performance, but requires a large amount of labeled data for the target prompt. In contrast, *cross-prompt* AES trains a model on a set of source prompts and tests it on unseen target prompts (Ridley et al., 2021). This approach is more practical, reducing the reliance on large labeled data for every new prompt. However, it faces challenges in achieving high performance due to source and target prompt variations.

Despite significant advances in AES for languages such as English (Klebanov and Madnani, 2022), Arabic AES remains understudied due to the lack of publicly annotated datasets for Arabic essay scoring, and the language's complex nature. Nevertheless, there has been some work on prompt-specific Arabic AES (Gaheen et al., 2021, 2020); however, to the best of our knowledge, no work has been done on cross-prompt Arabic AES. This motivated us to organize the first shared Task for Arabic Quality Evaluation of Essays in Multi-dimensions (*TAQEEM*).[2] The task focuses on developing models for the automatic assessment of Arabic essays, both at a holistic level and across several traits. Through *TAQEEM*, we aim to advance research in Arabic AES by releasing the *first publicly available dataset* of 1,265 Arabic essays annotated with holistic and seven traits: relevance (الصلة بالموضوع), organization (الهيكل العام), vocabulary (المفردات), style (الأسلوب والتماسك البنائي), development (الأفكار والمضمون), mechanics (الإملاء والترقيم), and grammar (البناء والتراكيب).

TAQEEM 2025[3] focuses on cross-prompt AES setup, where models are evaluated on their ability to generalize to unseen prompts by leveraging knowledge learned from different labeled

[1] A prompt is the text of a specific essay writing task.

[2] Pronounced in Arabic as "تَقْيِيْم".

[3] https://sites.google.com/view/taqeem-2025

Proceedings of The Third Arabic Natural Language Processing Conference, pages 966–976

Team	Tasks	Team Size	Affiliations
912 (Vu and Đáng Văn, 2025)	A	2	University of Information Technology (UIT), VNUHCM, Vietnam
MarsadLab (Bessghaier et al., 2025)	A	3	University of Kairouan, Northwestern University, Hamad bin Khalifa University
ARxHYOKA (Alnajjar et al., 2025)	B	2	Nara Institution of Science and Technology, Tokyo University of Science
Taibah (Almarwani et al., 2025)	A,B	3	Taibah University
ANLPers3	A	5	Prince Sultan University

Table 1: Participating teams in *TAQEEM* 2025.

source prompts, thereby ensuring robustness and adaptability in real-world applications. *TAQEEM* 2025 includes two subtasks: (A) **Holistic Scoring**, which involves predicting a single overall score for a given essay reflecting its general quality, and (B) **Trait-specific Scoring**, which involves predicting separate scores for individual traits of the essay.

TAQEEM 2025 attracted registrations by 11 teams for Task A and 10 teams for Task B. However, in the final evaluation phase, only 4 teams submitted a total of 9 runs for Task A, while 2 teams contributed 4 runs for Task B. With one team actively involved in both tasks, this resulted in a total of 5 unique teams overall participating in *TAQEEM* 2025. Table 1 lists the participating teams, along with their affiliations and team sizes.

The remainder of the paper is organized as follows. Section 2 reviews related work on AES datasets and systems. Section 3 formally defines the two tasks, presents the dataset, and describes the evaluation setup. Section 4 discusses the approaches adopted by the participating teams along with their performance results. Finally, Section 5 concludes with final thoughts on future directions.

2 Related Work

This section reviews prior AES research, with particular emphasis on Arabic datasets and systems.

Datasets Progress in English AES has been driven by large public datasets such as ASAP[4] and ELLIPSE[5] with around 13,000 and 6,500 annotated essays, respectively. In contrast, Arabic AES lags behind due to the scarcity of annotated datasets, which are often small, limited in annotations, or not publicly accessible. For instance, the Zayed Arabic English Bilingual Undergraduate Corpus (ZAEBUC) (Habash and Palfreyman, 2022) contains 214 essays but lacks holistic and trait annotations. The Arabic Learner Corpus (ALC)[6] includes 1,585 essays, though its annotations are not publicly available. More recently, QAES dataset (Bashendy et al., 2024) was released with only 195 essays annotated with holistic and trait scores, building on the larger Qatari Corpus of Argumentative Writing (QCAW) (Ahmed et al., 2024).

Other datasets with holistic or trait annotations exist but are not public, such as Abbir (Alghamdi et al., 2014), which contains essays from Saudi university students with holistic scores from 1 to 6, and AAEE (Azmi et al., 2019), which evaluates essays based on semantic analysis, writing style, and spelling accuracy. Other datasets was collected for Arabic short-answer scoring (Abdeljaber, 2021; Ouahrani and Bennouar, 2020). Despite prior efforts, Arabic AES research still lacks a publicly available dataset that provides both essays and corresponding scores. Our shared task addresses this gap by releasing *TAQEEM* dataset, annotated with holistic and seven-trait scores, thereby making a substantial contribution to Arabic AES resources.

Systems Despite limited datasets, several studies have explored Arabic AES. Early work relied on traditional approaches that required extensive feature engineering (Alqahtani and Alsaif, 2020; Alsanie et al., 2022; Sayed et al., 2025). Other methods incorporated reference essays for scoring (Abdeljaber, 2021; Alobed et al., 2021a; Al Awaida et al., 2019; Alobed et al., 2021b). More recent efforts have advanced Arabic AES through the use of AraBERT and large language models (LLMs). Ghazawi and Simpson (2024) fine-tuned AraBERT with notable success, while Machhout and Zribi (2024)

[4] https://www.kaggle.com/c/asap-aes
[5] https://github.com/scrosseye/ELLIPSE-Corpus
[6] https://www.arabiclearnercorpus.com

Trait	Description
Relevance	Relevance of the essay to the prompt
Organization	The structure of the essay
Vocabulary	Precision and variety of word choice
Style	Linking words and transition phrases
Development	The support and clarity of ideas
Mechanics	Spelling and punctuation
Grammar	Accuracy of grammatical structures
Holistic	The overall quality score

Table 2: A brief description of the scoring traits.

Data	Prompt	Type	Size	Len.
Training	1	Explanatory	215	137
Training	2	Persuasive	210	150
Test	9	Explanatory	420	153
Test	10	Persuasive	420	166

Table 3: *TAQEEM* 2025 dataset statistics. Size indicates number of essays, and length is indicated in words.

improved its performance by integrating handcrafted features for relevance evaluation. Mahmoud et al. (2024) further optimized AraBERT using parameter-efficient tuning strategies. In parallel, Ghazawi and Simpson (2025) tested LLM-based approaches, experimenting with different LLMs under different prompting setups.

3 TAQEEEM 2025

In this section, we formally define *TAQEEM* 2025 subtasks, introduce the dataset, and elaborate on the evaluation setup.

3.1 Task Description

TAQEEM 2025 comprises two subtasks: Task A focuses on holistic scoring, while Task B targets trait-specific scoring.

Task A: Holistic Scoring The task is defined as follows: Given a set of source prompts P_{src}, the aim is to train a *holistic* scoring model using those prompts to score essays written for an unseen target prompt $p_{trg} \notin P_{src}$. The model should produce a single holistic score that reflects the overall writing quality of each essay.

A writing prompt p in this task is defined as a tuple (a_p, E_p), where a_p is the textual description of the writing task of the prompt and E_p is a set $\{(e, h_e)\}$ of essays written for the prompt p; each essay e is associated with a holistic score h_e.

Task B: Trait-specific Scoring The task is defined as follows: Given a set of source prompts P_{src}, the aim is to train a *trait-specific* scoring model using those prompts to score essays written for an unseen target prompt $p_{trg} \notin P_{src}$. For each essay written for p_{trg}, the model should produce a score *for each trait* that reflects the quality of the essay for that trait.

A writing prompt p in this task is defined as a tuple (a_p, T_p, E_p), where a_p is the textual description of the writing task of the prompt, T_p is a set $\{(t, r_t)\}$ of traits; each trait t is associated with a rubric r_t, and E_p is a set $\{(e, \{s_{e,t}\})\}$ of essays written for the prompt p; each essay e is associated with a score $s_{e,t}$ for each trait $t \in T_p$. While each prompt has its own trait rubrics, those rubrics are usually common across different prompts for specific traits.

In *TAQEEM* 2025, all essays of all prompts are annotated for the same seven traits: Relevance (REL), Organization (ORG), Vocabulary (VOC), Style (STY), Development (DEV), Mechanics (MEC), and Grammar (GRA). We note that the holistic (HOL) score, used in Task A, represents the sum of all trait scores. Table 2 provides a brief description of each trait.

3.2 Dataset

The absence of standardized Arabic essay corpora, even modestly sized ones, has slowed the progress in Arabic AES. To address this gap, we introduce a *novel* dataset[7] of 1,265 Arabic essays written by native high school and first-year university students under test-like conditions. The essays span 4 distinct writing prompts, ensuring diversity in content and structure. Table 3 provides an overview of the prompts used in both the training and test sets, including the number of essays per prompt and their average length in words. Notably, the test set is, unusually, larger than the training set. This is because, at the time of releasing the training data, only 425 fully annotated essays from prompts 1 and 2 were available for use. The remaining essays (from prompts 9 and 10) were still undergoing annotation, which was finalized by the time of the test set release, thereby allowing these additional essays to be included for evaluation.

[7] https://gitlab.com/bigirqu/taqeem2025

المعرّف	نص الموضوع و الموضوع	الدرجات
011069	**نص الموضوع:** باتَ إهْتمام وحماس المراهقين لِتعلُّمْ رِياضةٍ جديدة أو الإنتظام في حُضور التَدْريبات الرياضية الخاصة بِها يَتضَاءل يومًا بعد يوم؛ حتّى صَارَ يدُق نَاقُوسَ خَطرٍ مقلق ينْذر بوجُود جيلٍ ضَعيفِ البنيَة (الجسم). اكتب مقالًا مكوّنا من ثلاثمائة (٣٠٠) كلمة توضّح فيه أسباب انتشار هذه الظاهرة، مُراعيًا سمات المقال التفسيري، وسلامة اللغة، ومُوظفًا علامات الترقيم وأدوات الربط بشكلٍ صحيح. **الموضوع:** ان الرياضه في يومنا الحالي بات الاهتمام بها من قبل المراهقين لتعلم رياضه جديده او في الحضور للتدريبات الخاصه، ويوم بعد يوم يقل النشاط ةيزداد الكسل لديهم، وهذا خطر على صحتهم من قبل تاكيد الاطباء عليه. وكما ان زاد قل نشاط المراهقين في وقتنا الحالي يجب علينا ان نعالج هذا قبل ان يشكل خطرا اكثر عليهم ويكونون جيل ضعيف البنيه. ومن اسباب هذا الكسل عدم اهتمام وتفرغ بعض اولياء الامور لاطفالهم وتشجيعهم على الرياضه، وكذلك كثرة جلوس الاطفال على الاجهزه الالكترونيه مما يسبب الكسل في الحركه والخمول لدى الطفل المراهق، وايضا في بعض المدارس عدم ممارسة الرياضه في بداية اليوم الدراسي و عدم التوعيه بأهميه الرياضه في حياتنا من الناحيه الصحيه وكم هي تشكل خطرا على الجسم اذا ما مارسوا الرياضه، وكذلك عدم تعليم الطالب وتشجيعه من قبل المدرسه لتعلم رياضه جديد وعمل انشطه وفعاليات بين فترات للطلبه المراهقين. وجملة القول، يجب على الاسره الاهتمام باطفالهم والتفرغ لهم وكذلك المدرسه في عمل لهم نشاطات والتعرف على الرياضات الجديده لتعلمها.	الصلة بالموضوع: ٢ الهيكل العام: ٤ المفردات: ٣ الأسلوب والتماسك: ٣ الأفكار و المضمون: ٣ الإملاء و الترقيم: ٢ البناء و التراكيب: ٣ المجموع الكلي: ٢٠
100099	**نص الموضوع:** يرى البعض أنّ تقليل عبء الواجبات المنزليّة على الطّلبة لا يؤثّر على أدائهم الأكاديميّ، ويرفع رفاهيّتهم، ويزيد الأوقات التي يقضونها مع أسرهم. كيف ترى قضية تقليل الواجبات المنزلية؟ اكتب مقالًا مكونًا من ثلاثمائة (٣٠٠) كلمة لتقنع القارئ بوجهة نظرك في هذا الموضوع موظّفًا الأدلة والحجج الدّاعمة لهذا الرأي، ومراعيًا أساليب الإقناع، ومستخدمًا علامات الترقيم وأدوات الربط المناسبة. **الموضوع:** إنه مما لا شك فيه أن الواجبات المدرسية والتقييمات المنزلية أصبحت وسيلة أساسية يرتكز عليها المدرسين في تثبيت المعلومة لدى الطلبة. يُعطى الطالب الواجب المدرسي ويتم تخصيص مكافأة لمن أنجز أو عقوبة لمن لم ينجز العمل الموكّل به. وبذلك، يضمن المعلم والكادر التعليمي أن الطالب قد خصص وقتًا للدراسة، وأنه حاول ليرى نقاط الضعف لديه في هذا الدرس. لكن يرى البعض أن الواجبات المنزلية قد تزيد من العبء على الطالب وتحمله ضغوطًا أخرى قد يكون بغنى عنها . إن تكليف الطالب بأداء الواجبات المدرسية والحرص على تقديمها في الموعد المناسب يرفع من حس المسؤولية لدى الطالب : وذلك بتقليل الاعتماد على الغير ومحاولة الاعتماد على النفس وضبط أهدافه وتنظيم وقته ليناسب وقت تسليمه للواجب . إن الواجبات والتكليفات أسانًا من جميع النواحي العملية ترفع من حس المسؤولية عامةً ، حتى في ديننا الحنيف قد أمرنا الله تعالى بأداء خمسة فروض وواجبات مكلفة في كل يوم ، يعاقب من لم يؤدها ويجازى من قام بها . من هذا المنطلق ومن هذه الفكرة أقول بأن الواجبات والفروض والتكليفات التي تفرض على الطالب إنما لزيادة حرصه على دراسته وعدم التراخي وإشعال حس اليقظة والمسؤولية تجاه النفس . أضف إلى ذلك، أن أداء الطالب لواجباته المدرسية يرفع من المستوى الأكاديمي للطالب ويثبت المعلومات في ذاكرته: وذلك لأن الطالب قد يتذكر المعلومة التي التقطها من المعلم في الفصل لفترة قصيرة فقط ثم ينساها في حال أنه لم يراجعها ويثبتها؛ فعلميًا الذاكرة قصيرة المدى تخزن المعلومة الحسية في حدود ساعتين إلى ثلاث ساعتين ثم تبدأ بالتلاشي مع الوقت، أما إذا تم تكرار تلك المعلومة فإنها ستثبت مدة أطول، وكلما زاد التكرار زادت المعلومة ثباتًا واستقرارًا وزادت معها الفائدة العائدة منها. ختامًا أقول: إن تكليف الطالب بأداء مهامه الدراسية لهو من مصلحته الشخصية أولاً وقبل كل شيء، و قد يرى الطالب بأنه زيادة تكليف وأنه عبء عليه، لكن في المقابل يجب توعية الطلبة بضرورة المتابعة والدراسة حتى وإن لم يكلفوا بواجبات مدرسية؛ لتكون الفائدة خالصة ولينمو لديهم حس المسؤولية، وقيم ضبط ومجاهدة النفس، والصبر.	الصلة بالموضوع: ٢ الهيكل العام: ٥ المفردات: ٥ الأسلوب والتماسك: ٥ الأفكار و المضمون: ٥ الإملاء و الترقيم: ٥ البناء و التراكيب: ٤ المجموع الكلي: ٣١

Table 4: Annotated Essays from *TAQEEM* 2025 dataset.

Annotation Process The annotations were conducted by two main native Arabic language specialists, with a third annotator resolving disagreements. Annotators were selected for their experience in teaching and assessing Arabic writing. To ensure score reliability, all annotators received training sessions to understand the assessment rubric and maintain consistent annotation procedures. The rubric itself was adapted from the Core Academic Skills Test (CAST) developed by the Qatar University Testing Center (QUTC).[8] A full detailed English-translated version of the rubric is in Appendix A. Each essay was annotated across 7 traits: REL, ORG, VOC, STY, DEV, MEC, and GRA, along with an overall quality score (HOL) computed as the sum of all trait scores. Traits are rated on a 0 to 5 scale, except for REL, which is from 0 to 2, and the HOL, which is from 0 to 32, all using 1-point increments. Table 4 shows two essays from the *TAQEEM* 2025 dataset, a training essay (ID 011069) from an explanatory prompt (Prompt 1) and a test essay (ID 100099) from a persuasive prompt (Prompt 10), along with their prompts and scores.

Inter-Annotator Agreement We assessed annotation quality using the Quadratic Weighted

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

Kappa (QWK) (Cohen, 1968), averaging trait-level scores per prompt to obtain prompt-level agreement. The resulting average agreements were 0.692 for Prompt 1, 0.640 for Prompt 2, 0.525 for Prompt 9, and 0.676 for Prompt 10. According to the scale outlined by Landis and Koch (1977), Prompts 1, 2, and 10 fall within the range of *substantial* agreement, while Prompt 9 shows *moderate* agreement, possibly due to less precise wording that made it more open to interpretation and increased variability in annotators' judgments. Nevertheless, the overall results indicate strong rater consistency across prompts.

This dataset is used in both subtasks of *TAQEEM* 2025, with distinct evaluation targets. Task A (Holistic Scoring) uses the holistic score assigned to each essay, whereas Task B (Trait-specific Scoring) uses the seven individual trait scores. Although this dataset is limited in scale, it represents a carefully curated first step resource to address data scarcity in Arabic AES.

3.3 Evaluation Setup

This section outlines the setup used to evaluate participating systems in *TAQEEM* 2025. We describe the leaderboard and repository infrastructure provided to participants, as well as the evaluation measures adopted to ensure consistent, and reproducible comparisons across submitted systems.

3.3.1 Leaderboard and Repository

The leaderboard for both Task A[9] and Task B[10] was hosted on Codabench, providing participants a platform to submit their runs, evaluate system outputs, and benchmark performance. Each team was required to submit their predictions in a single file, referred to as a run file. Submissions were restricted to a maximum of 30 runs on the development set and up to 3 runs on the test set. Typically, each run represented a distinct system or model.

To facilitate the submission process, we made the submission checker and evaluation scripts available through the shared task repository. These resources enabled participants to validate their runs before leaderboard submission. Additionally, we released a regression-based baseline by fine-tuning AraBERTv02 (Antoun et al.), along with the corresponding code, in the same repository.

[9] https://www.codabench.org/competitions/9282/
[10] https://www.codabench.org/competitions/9295/

Team	Run	QWK	MSE	RMSE
Taibah	1	**0.751**	**25.44**	**5.01**
912	1	0.673	28.51	5.33
912	2	0.673	28.51	5.33
ANLPers3	1	0.650	31.68	5.62
ANLPers3	2	0.642	28.28	5.28
baseline	0001	0.639	29.01	5.37
ANLPers3	3	0.602	29.73	5.45
Taibah	2	0.488	33.15	5.73
MarsadLab	1	0.438	50.56	7.07
MarsadLab	2	0.438	50.56	7.07

Table 5: Task A performance results on the test set. Bold values are the best for each measure.

3.3.2 Evaluation Measures

The primary evaluation metric for *TAQEEM* 2025 is the Quadratic Weighted Kappa, a standard AES performance metric that quantifies the agreement between human-assigned scores and system predictions. Additionally, we report the mean squared error (MSE) and the root mean squared error (RMSE) to provide a more comprehensive analysis of model performance, as these metrics capture the magnitude of prediction errors, penalize larger deviations more heavily, and allow for direct comparison of error scales across models.

The subtasks are evaluated independently. Task A is assessed based on the average QWK of the holistic score across the test prompts. For Task B, the average QWK for each trait across the test prompts is measured separately, and teams are ranked based on the average QWK over all traits.

4 Participating Systems and Results

This section presents the participating systems and their performance in *TAQEEM* 2025, highlighting the methods used and the corresponding evaluation results for both subtasks.

4.1 Task A: Holistic Scoring

Task A attracted 4 teams in total, each adopting distinct methodological approaches, resulting in 9 runs submitted on the test set. The top-ranked team, Taibah) (Almarwani et al., 2025), employed a rubric-guided few-shot prompting strategy based on GPT-4o, utilizing exemplars to assess the holistic quality of essays. The 912 team (Vu and Đáng

Team	Run	QWK								MSE	RMSE
		REL	ORG	VOC	STY	DEV	MEC	GRA	Avg.		
Taibah	1	0.562	0.668	0.642	**0.678**	**0.703**	**0.644**	**0.664**	**0.652**	**0.762**	**0.857**
ARxHYOKA	1	0.553	0.709	0.633	0.654	0.640	0.515	0.580	0.612	0.760	0.848
ARxHYOKA	2	**0.585**	0.711	0.646	0.666	0.647	0.477	0.544	0.610	0.758	0.845
ARxHYOKA	3	0.545	**0.712**	**0.653**	0.620	0.629	0.482	0.506	0.592	0.797	0.867
Baseline	-	0.155	0.591	0.574	0.572	0.458	0.445	0.513	0.472	1.005	0.990

Table 6: Task B performance results on the test set. The best score for each metric is highlighted in bold.

Văn, 2025) adopted a pre-trained Arabic encoder (AraBERTv02) with a lightweight single-layer MLP head, coupled with a distribution-sensitive weighted MSE loss to address score imbalance. The MarsadLab system (Bessghaier et al., 2025) was also built on a fine-tuned AraBERT model, integrating lexical features into the embeddings to predict essay scores.

In terms of performance, which is summarized in Table 5, teams were ranked by their highest average QWK score across all test prompts. The highest performing system, submitted by the Taibah team, achieved a QWK of 0.751, significantly outperforming the baseline of the shared-task (QWK of 0.639). Team 912 followed by two identical runs reaching a QWK of 0.673. Team ANLPers3 also delivered competitive systems, with their best run achieving a QWK of 0.650. The two runs of MarsadLab resulted in the lowest performance across submissions (QWK of 0.438), and it was the only team that did not outperform the baseline. Overall, three of the four teams submitted at least one run that outperformed the baseline, reflecting both the effectiveness and diversity of the applied approaches.

4.2 Task B: Trait-specific Scoring

Task B featured two participating teams, who together submitted four runs on the test set, implementing different approaches for trait-specific scoring. Notably, the Taibah team, which had also ranked first in Task A, once again secured the top position in Task B. They adopted a GPT-4o-based few-shot prompting approach, leveraging trait-specific rubrics to achieve fine-grained scoring (Almarwani et al., 2025). The second-ranked team (ARxHYOKA) (Alnajjar et al., 2025) explored a broader methodological spectrum, including GPT-based few-shot prompting, fine-tuned BERT-based models, classical machine learning approaches with embeddings and handcrafted features, and fine-tuned text-generation LLMs. Their best-performing configuration used GPT-4.1 with 10-shot chain-of-thought prompting.

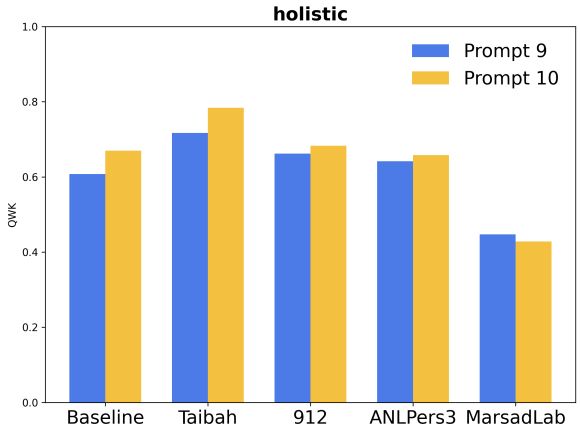

Figure 1: Performance of Task A teams across test prompts. The best submitted run was considered.

Performance was evaluated based on the average QWK across all 7 traits. The top-performing run, submitted by Taibah, achieved an average QWK of 0.652, with MSE of 0.762 and RMSE of 0.857, substantially outperforming the shared-task baseline (average QWK of 0.472, MSE of 1.005, RMSE of 0.990). ARxHYOKA also outperformed the baseline, reaching an average QWK of 0.612. These results underscore the potential of prompting strategies for trait-specific scoring in Arabic AES. Table 6 presents the test results for Task B, reporting QWK, MSE, and RMSE measures.

4.3 Analysis and Discussion

This section provides a detailed analysis of the results from two perspectives: *trait-level* performance and *prompt-level* performance. Figure 1 shows Task A performance for the holistic scoring,

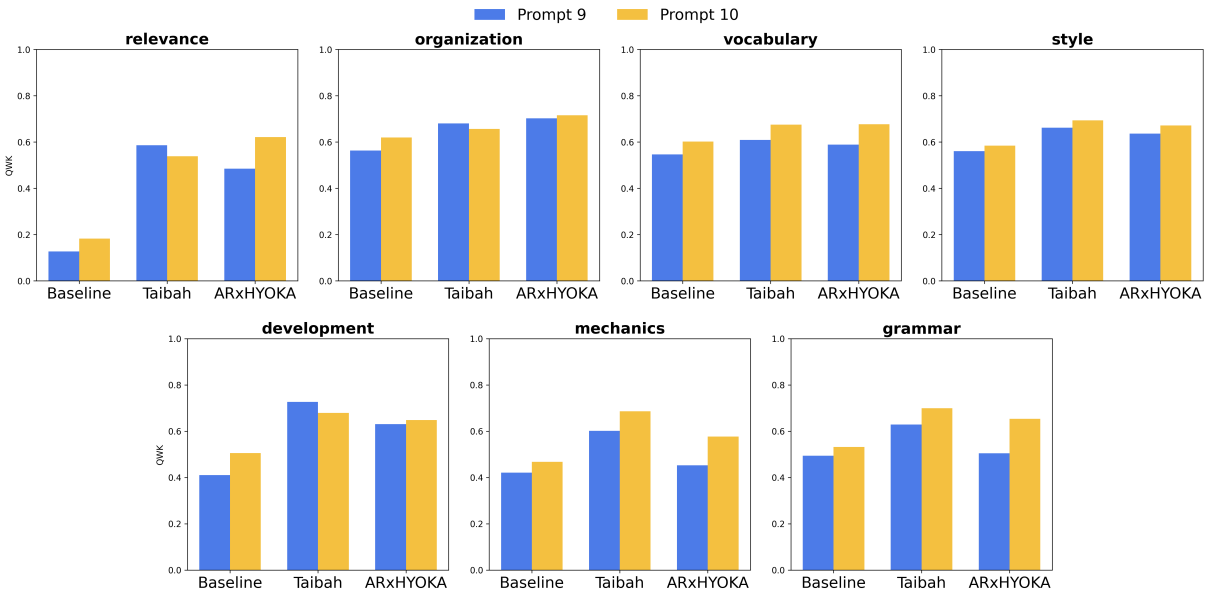

Figure 2: Trait-level performance of Task B teams on prompt 9 (Explanatory) and prompt 10 (Persuasive). The best submitted run was considered.

whereas Figure 2 illustrates Task B performance for the trait-specific scoring. In both cases, the figures report results from each team's best submitted run, providing a clear view of top-performing approaches for each task across the two test prompts.

Trait-level Analysis Figure 2 reveals clear differences in teams performance across the different traits. Notably, the REL trait appears to be the most challenging, as evidenced by the highest QWK score being only 0.585, achieved by the ARxHYOKA team across the two test prompts. The baseline, which relies on fine-tuning AraBERTv2, struggled the most with the REL trait. This suggests that smaller encoders like AraBERT have difficulty capturing the semantic alignment between essays and prompts. In contrast, both Taibah and ARxHYOKA show substantial improvements in REL, demonstrating that leveraging the advanced capabilities of GPT-4o and GPT-4.1 through few-shot prompting significantly enhances performance on this semantically complex trait. The MEC trait also proved difficulty, with an average QWK of 0.580 across the two participating teams and test prompts, reflecting the difficulty of capturing fine-grained linguistic correctness, such as punctuation, spelling, and syntax. Traits VOC and GRA showed moderate performance across teams, while ORG, STY, and DEV exhibited comparatively higher and more consistent performance.

Prompt-level Analysis Performance also varies depending on the prompt type. From Figures 1 and 2, it is evident that Prompt 10 (Persuasive) generally resulted in higher performance across most traits and teams compared to Prompt 9 (Explanatory). This pattern suggests that persuasive writing, which typically follows a predictable and structured format (e.g., a clear thesis statement, supporting arguments, counterarguments, and a conclusion), is easier for models to capture. Explanatory essays, on the other hand, exhibit greater structural and stylistic diversity, making it more difficult for models to identify consistent patterns.

Overall, these results clearly indicate that GPT-4-based models (Taibah & ARxHYOKA) generally outperform fine-tuned BERT models (Baseline, 912, MarsadLab) across most traits. This shows the potential of LLMs for automated essay scoring. Their ability to understand complex language and capture nuanced relationships leads to significantly higher agreement with human scores. While fine-tuned BERT models provide a reasonable baseline, they struggle to match the performance of LLMs.

5 Conclusion

Automated Essay Scoring has seen notable progress in writing evaluation, yet the development of AES systems tailored for the Arabic language remains very limited. This scarcity motivated the organization of TAQEEM, the first

shared task dedicated to Arabic AES, aiming to foster state-of-the-art research in this area, with a novel dataset of 1,265 essays across four different writing prompts.

TAQEEM 2025 attracted 15 researchers and practitioners across five teams from different institutions. It comprised two cross-prompt subtasks: holistic scoring (Task A), with four participating teams, and trait scoring (Task B), with two teams. The participating teams explored diverse solutions, including fine-tuning transformer-based models and employing classical machine learning approaches. However, as expected, LLMs were heavily adopted by multiple teams, achieving state-of-the-art performance and outperforming the baseline. For task A, the teams employed different solutions that mainly focused on fine-tuning different transformer-based models and prompting LLMs using different prompting techniques. For task B, the best results were achieved with few-shot in-context learning and chain-of-thought prompting using GPT-4 variants.

Overall, *TAQEEM* 2025 established the first benchmark for Arabic AES, providing a foundation for future research and community efforts to develop AES systems for the Arabic language. In the next iteration, we plan to expand the shared task by incorporating a larger training set that encompasses a wider range of essay types, topics, and student populations, thereby fostering deeper research advancements and broader community contributions in this area.

6 Limitations

One key limitation of TAQEEM is the size and diversity of the dataset. Although it provided a useful benchmark for Arabic AES, the training and test sets were relatively small and may not fully capture the variety of essay topics, writing styles, or proficiency levels. Moreover, the test set was larger than the training set due to the challenges and time required to provide high-quality annotated data. This limitation could affect the generalizability of the models trained and evaluated in this shared task. Another limitation is the small number of participating teams, which reduces the variety of approaches evaluated.

Acknowledgments

We would like to thank our dedicated annotators who contributed to building the dataset. We also extend our sincere thanks to the Qatar University Testing Center (QUTC), the Ministry of Education, the participating schools, and the students for their essential role in making this work possible. 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

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.

Nada Almarwani, Alaa Alharbi, and Samah Aloufi. 2025. Taibah at TAQEEM 2025: Leveraging GPT-4o for Arabic Essay Scoring. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, China.

Mohamad Alnajjar, Ahmad Almoustafa, Tomohiro Nishiyama, Shoko Wakamiya, Eiji Aramaki, and Takuya Matsuzaki. 2025. ARxHYOKA at TAQEEM2025: Comparative Approaches to Arabic Essay Trait Scoring. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, China.

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. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

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.

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.

Mabrouka Bessghaier, Md. Rafiul Biswas, Amira Dhouib, and Wajdi Zaghouani. 2025. Marsadlab at TAQEEM 2025: Prompt-Aware Lexicon-Enhanced Transformer for Arabic Automated Essay Scoring. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, China.

Jill Burstein. 2013. *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

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

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.

Marwa M Gaheen, Rania M ElEraky, and Ahmed A Ewees. 2021. Automated students arabic essay scoring using trained neural network by e-jaya optimization to support personalized system of instruction. *Education and Information Technologies*, 26:1165–1181.

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.

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

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

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*.

Leila Ouahrani and Djamal Bennouar. 2020. ARASAG 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.

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

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 Arabic Natural Language Processing Conference (ArabicNLP 2025)*, China.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.

Trong-Tai Dam Vu and Thìn Đáng Văn. 2025. 912 at TAQEEM 2025: A Distribution-aware Approach to Arabic Essay Scoring. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, China.

A Grading Rubric

For annotating *TAQEEM*2025 dataset, we utilized the rubric from the Core Academic Skills Test

(CAST) designed by the Qatar University Testing Center (QUTC)[11],which is provided in Arabic. This rubric guided the scoring of seven traits: relevance (REL), organization (ORG), vocabulary (VOC), style (STY), development (DEV), mechanics (MEC), and grammar (GRA). An English-translated version of the CAST grading rubric for each trait is provided in Table 7.

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

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, 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 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, 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	Content is not related to the subject; ideas are random and lack coherence, sequence, and evidence.	Content is somewhat related; ideas are sequential but main idea disappears during writing, limited coverage, and poor use of supporting structures.	Content is completely related; ideas mostly follow sequence, main idea gradually disappears, some evidence present but disorganized.	Content is completely related; ideas are clear, organized, coherent, with main idea connected to sub-ideas, specific position adopted, some arguments and evidence presented coherently.	Content is completely related; ideas are clear, organized, coherent, main idea connected to sub-ideas, specific position adopted, arguments and evidence presented coherently, comprehensive coverage of opinions, and use of various persuasive methods.
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 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 occasional minor slips.
Note: A score of zero is given if the response is completely memorized or copied from the prompt, if the student did not attempt the task, or if the content is irrelevant to the given topic.					

Table 7: CAST Persuasive/Argumentative Writing Rubric - English Translation (Bashendy et al., 2024).