ARxHYOKA at TAQEEM2025: Comparative Approaches to Arabic Essay Trait Scoring

Mohamad Alnajjar
Nara Institute of Science and Technology
alnajjar.mohamad.al8@naist.ac.jp

Ahmad Almoustafa
Tokyo University of Science
1425501@ed.tus.ac.jp

Tomohiro Nishiyama
Nara Institute of Science and Technology
nishiyama.tomohiro.ns5@is.naist.jp

Shoko Wakamiya
Nara Institute of Science and Technology
wakamiya@is.naist.jp

Eiji Aramaki
Nara Institute of Science and Technology
aramaki@is.naist.jp

Takuya Matsuzaki
Tokyo University of Science
matuzaki@rs.tus.ac.jp

Abstract

Arabic automated essay scoring (AES) presents unique challenges due to the linguistic complexity of Arabic and the need for rubric-specific evaluation. In this paper, we present ARxHYOKA, our submission to TAQEEM2025 Task B, which targets trait-specific AES using the Core Academic Skills Test (CAST) rubric. We evaluate four approaches: (1) GPT-based few-shot prompting, (2) fine-tuning BERT-based models, (3) classical machine learning approaches with embeddings and handcrafted features, and (4) fine-tuning text-generation large language models (LLMs). Our best-performing system, GPT-4.1 with 10-shot CoT prompting, achieved the highest official score, outperforming all other approaches in average Quadratic Weighted Kappa (QWK) in the test phase. Fine-tuned BERT-based models performed on par with both the shared-task baseline and our GPT prompting setup in the development phase, while classical machine learning methods trailed these systems, and the fine-tuned Arabic LLM ranked last. We provide comparative analyses across systems to inform future research on Arabic AES.

1 Introduction

The TAQEEM2025 Task B (Bashendy et al., 2025) targets automated scoring of Arabic essays, evaluating seven traits defined by the Core Academic Skills Test (CAST) rubric.[1] A central challenge is cross-prompt generalization: systems trained on one prompt must accurately score essays from a different, unseen prompt. This task advances robust, rubric-aligned Arabic NLP evaluation and enables fair, scalable, and transparent assessment of student writing in high-stakes settings across real-world educational contexts. In our submission, we compared three main approaches: prompting, fine-tuning, and training traditional machine learning (ML) models. Our key findings are as follows:

- **GPT-based few-shot prompting** achieved the highest average QWK, outperforming the baseline in the test phase and closely matching it in the development phase. Performance was sensitive to the number and quality of examples as well as the language used in the prompt.
- **Fine-tuning BERT-based models** produced strong results close to the baseline in the development phase. Both Arabic-specific and multilingual models performed well.
- **Fine-tuning text-generation model Saka 14B** yielded poor results, suggesting that relatively small LLMs may not be optimal for this scoring task without further adaptation.
- **Classical ML approaches** remained competitive, with performance improving when linguistic features were combined with embeddings.

Code and prompts are available at our repository.[2]

2 Background

The task involves predicting numeric scores for seven traits: **Relevance** (0–2), **Organization**, **Vocabulary**, **Style**, **Development**, **Mechanics**, and **Grammar** (0–5 each). Essays are written in response to prompts that are either *explanatory* or *persuasive*, mimicking real classroom writing tasks.

The official dataset for TAQEEM2025 Task B is summarized in Table 1. It contains two prompt types in the training phase and two in the test phase, with essays of approximately 300 words each. All essays have been scored by expert raters using the official CAST rubrics for each trait.

[1] https://www.qu.edu.qa/en-us/testing-center/TestDevelopment/Pages/cast.aspx

[2] https://github.com/Mohamad-Alnajjar/ARxHYOKA

Proceedings of The Third Arabic Natural Language Processing Conference, pages 977–982

Split	Prompt ID	Type	# Essays
Development Phase	1	Explanatory	215
Development Phase	2	Persuasive	210
Testing Phase	9	Explanatory	420
Testing Phase	10	Persuasive	420

Table 1: Dataset composition for TAQEEM2025 Task B.

This setup poses challenges for both linguistic coverage and cross-prompt adaptability, particularly for traits such as **Relevance**, where alignment with the prompt topic is critical.

3 System Overview

We present the systems explored for Arabic essay trait scoring, covering GPT-based prompting, fine-tuned BERT models, classical ML baselines, and a fine-tuned generative LLM.

3.1 GPT-Based Few-Shot Prompting

This system leverages GPT-4.1 to score essays based on in-context learning. The model relies entirely on the design of the prompt and the quality of examples provided. We tested prompts in both Arabic and English with different random sets of examples from the dataset in the development phase. The prompt includes:

- Detailed instructions for scoring.
- The CAST rubric.
- The essay type (explanatory or persuasive).
- The original writing prompt given to students.
- Instructions for structured output formatting.

We systematically compared model performance across:

- Arabic vs. English rubrics and prompts (translated using GPT-4.1).
- Number of in-context examples (0, 1, 5, and 10 shots).

3.2 Fine-Tuning BERT Models

We fine-tuned three encoder-only transformers: mDeBERTa-v3-base (He et al., 2021), XLM-R-large (Conneau et al., 2019), and CAMeLBERT-mix (Inoue et al., 2021), as independent systems (no ensembling). Essays are tokenized with each model's native tokenizer, truncated to 512 tokens, and passed to a 7-dimensional regression head to jointly predict the seven trait scores; at inference, continuous outputs are rounded and clamped to valid per-trait ranges.

3.3 Classical ML Approaches

We generated embeddings for each essay using `CAMeL-Lab/bert-base-arabic-camelbert-mix` (Inoue et al., 2021) and fed them to regression models to predict scores across seven traits. Inspired by (Bashendy et al., 2024), we also extracted 14 handcrafted linguistic features (listed in Table 10 in the Appendix) and evaluated the best-performing models during our experiments with and without these features.

We tested several pooling strategies and trained five regressors: LASSO, ElasticNet, Ridge, XGBoost, and Random Forest. Pooling strategies evaluated include:

1. `[CLS]` token
2. Average pooling
3. Average pooling + `[CLS]` token

3.4 Fine-Tuning Text-Generation LLM

We adapted `Sakalti/Saka-14B` (Sakalti, 2024), an open-source Arabic LLM, for trait-specific scoring using parameter-efficient fine-tuning (PEFT) via LoRA (Hu et al., 2021). To encourage rubric-grounded reasoning, we manually created two datasets:

- **Simple CoT**: 5–6 concise reasoning steps per trait focusing on essential rubric criteria.
- **Advanced CoT**: 7–8 detailed reasoning steps with deeper justification aligned to rubric criteria.

Each training instance concatenated the writing prompt, student essay, and trait-specific reasoning sequence with the gold score, encouraging the model to emulate human evaluation.

4 Experimental Setup

All experiments, including hyperparameter tuning and prompt engineering, used a cross-prompt setting: models were trained on **Explanatory** essays and tested on **Persuasive** essays. After selecting the best configurations, we retrained on the *union* of both essay types for the final submission.

- **Classical ML Approaches:** We performed 3-fold cross-validation using `scikit-learn`, optimizing for QWK. Initial experiments (Table 2) showed Ridge (AVG pooling, 0.521) and ElasticNet (CLS+AVG pooling, 0.527) as the strongest models, so we selected them for further evaluation. Incorporating handcrafted linguistic features (LF) improved results across both models (Table 3), with Ridge (AVG + LF) achieving the best QWK of 0.539. These findings highlight the

complementary value of shallow linguistic cues when combined with transformer embeddings.

Model	CLS	AVG	CLS + AVG
Lasso	0.480	0.517	0.482
ElasticNet	0.474	0.518	0.527
XGBoost	0.472	0.495	0.479
Ridge	0.454	0.521	0.471
RandomForest	0.447	0.492	0.494

Table 2: Performance of different regression models using CLS, AVG, and CLS+AVG embeddings during experiments.

Pooling	Features	Ridge	ElasticNet
AVG	+ LF	0.539	0.529
AVG	– LF	0.524	0.514
AVG + CLS	+ LF	0.533	0.532
AVG + CLS	– LF	0.511	0.539

Table 3: Performance comparison of Ridge and ElasticNet across pooling strategies with and without linguistic features (LF) on the development dataset.

- **GPT-Based Few-Shot Prompting:** A structured CoT prompt was employed to score the essays. We used an English version of both the CAST rubric and the essay prompts, achieving better performance after translation (QWK improved from 0.539 to 0.579 compared to Arabic). We also tested different numbers of shots; Table 4 compares 0, 1, 5, and 10 shots, showing consistent improvement as the number of provided examples increased. The prompt strictly specified the output format, and the model outputs were parsed to extract trait name–score pairs, which were organized into a table with one row per essay. The total API cost was approximately USD 21, covering exploratory experiments, development dataset scoring, and test dataset scoring (around 21,346,941 tokens in total). The final version of the prompt template used for submission is included in the Appendix.

Shots	QWK Score
0-shot	0.579
1-shot	0.597
5-shot	0.603
10-shot	0.631

Table 4: Performance of few-shot prompting with varying numbers of examples during experiments.

- **Fine-Tuning BERT-Based Models:** Hyperparameter tuning explored adaptation scope {full, last-6, last-3 layers} and learning rates $\{1e{-}5, 2e{-}5, 3e{-}5\}$ under AdamW; training ran up to 100 epochs with early stopping on development macro-QWK.
- **Fine-Tuning Text-Generation LLM:** We fine-tuned `Sakalti/Saka-14B` with LoRA on all attention projections (r=32, α=64, dropout 0.08), using two rubric-aware supervision styles: *Simple CoT* (5–6 steps per trait) and *Advanced CoT* (7–8 steps per trait). Training ran on 5× NVIDIA TITAN RTX (24 GB) GPUs with learning rate $2e{-}5$, batch size 1, gradient accumulation 8, and fp16; we fixed the budget at **3 epochs** because training loss decreased monotonically, reaching **1.37** (Simple) and **0.31** (Advanced) by epoch 3, indicating continued fitting of the supervision. Inference used deterministic decoding (temperature 0.0, `max_new_tokens` 80), and outputs were parsed into seven trait-level integers and evaluated with QWK, MSE, and RMSE.

5 Results

This section presents system performance in both the development and testing phases. We report results using QWK, MSE, and RMSE across all traits. The analysis highlights the differences in agreement with human raters and calibration quality among all the models.

5.1 Development Phase

We first evaluated all four system families under the *cross-genre* setup (training on explanatory essays and testing on persuasive essays, and vice versa), averaging results across both directions. Table 5 reports the mean (QWK), mean squared error (MSE), and root mean squared error (RMSE) across the seven traits.

In Table 5, the "Fine-tuned BERT" row corresponds to **mDeBERTa-v3-base** trained with learning rate **2e−5**, **last-6 layers** unfrozen, with early stopping—the best single checkpoint among our BERT-based models—achieving the best development QWK *among our systems* (**0.575**), slightly below the shared-task baseline (**0.582**). GPT-based prompting (10 shots) was close (0.564), classical ML (AVG-pooling ridge with linguistic features) trailed (0.539), and the fine-tuned LLM lagged with lower QWK (0.480).

System	QWK	MSE	RMSE
Baseline	**0.582**	**0.504**	**0.699**
Fine-tuned BERT	0.575	0.596	0.758
GPT-based Few-Shot	0.564	0.549	0.727
Classical ML	0.539	0.624	0.777
Fine-tuned LLM	0.480	0.821	0.887

Table 5: Development set performance across system families.

5.2 Testing Phase

The official evaluation was conducted under a *cross-prompt* setting, where systems were tested on previously unseen prompts in challenging conditions. Table 6 reports macro-average results across all seven traits. GPT-based few-shot prompting achieved the strongest performance, improving from 0-shot (0.592 QWK) to 1-shot (0.610) and 10-shot (**0.612**), with GPT-1-shot also producing the lowest error rates (MSE **0.758**, RMSE **0.845**). Among non-GPT systems, **Classical ML** with AVG-pooling ridge and linguistic features reached 0.582 QWK, **Fine-tuned BERT** 0.554, and the **Fine-tuned LLM** 0.538; all exceeded the shared-task baseline (0.472).

Calibration differed by family. While GPT variants achieved both the highest QWK and the lowest error rates, **BERT** improved over the baseline on MSE (0.949 vs. 1.005) and RMSE (0.956 vs. 0.990) while maintaining moderate QWK. In contrast, **Classical ML** and the **Fine-tuned LLM** raised QWK but suffered from higher MSE (1.081 and 1.029, respectively). Taken together, these results suggest that GPT prompting is most effective for balancing *ordinal agreement* with human raters (QWK) and *absolute calibration* (MSE/RMSE), whereas other approaches achieve only partial and less consistent gains.

System	QWK	MSE	RMSE
Baseline	0.472	1.005	0.990
GPT-0-shot	0.592	0.797	0.867
GPT-1-shot	0.610	**0.758**	**0.845**
GPT-10-shot	**0.612**	0.760	0.848
Classical ML	0.582	1.081	1.038
Fine-tuned BERT	0.554	0.949	0.956
Fine-tuned LLM	0.538	1.029	0.995

Table 6: Official testing results.

5.3 Error Analysis

Across both development and testing phases, distinct error patterns emerged for each model family. **GPT few-shot** yields the highest exact-match rate, especially on *Relevance* and *Development*, with a mild tendency to under predict extremes. **BERT** systematically skews high, over predicting most on *Vocabulary*, *Style*, and *Grammar*. **Saka-14B (fine-tuned)** also overestimates, most visibly for *Vocabulary*/*Style*, and sporadically under predicts *Relevance*, indicating weaker calibration under unseen prompts. In contrast, the **Ridge** baseline consistently under predicts across traits, most notably for *Organization* and *Development*. Overall, GPT is best-calibrated, BERT/Saka tend to score high, and Ridge tends to score low, with these tendencies persisting from development (Figure 1) and testing (Figure 2).

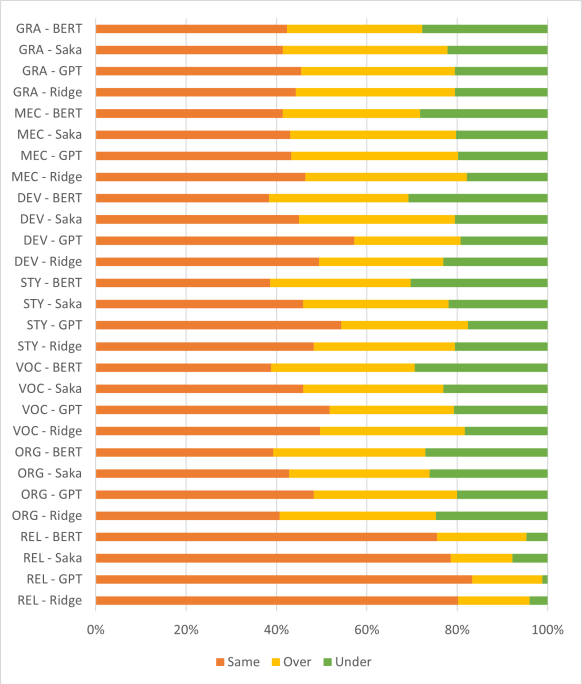

Figure 1: Development-phase calibration across traits and models. Bars show the proportion of predictions that were exact matches (*Same*), overestimates (*Over*), or underestimates (*Under*); stacks sum to 100%.

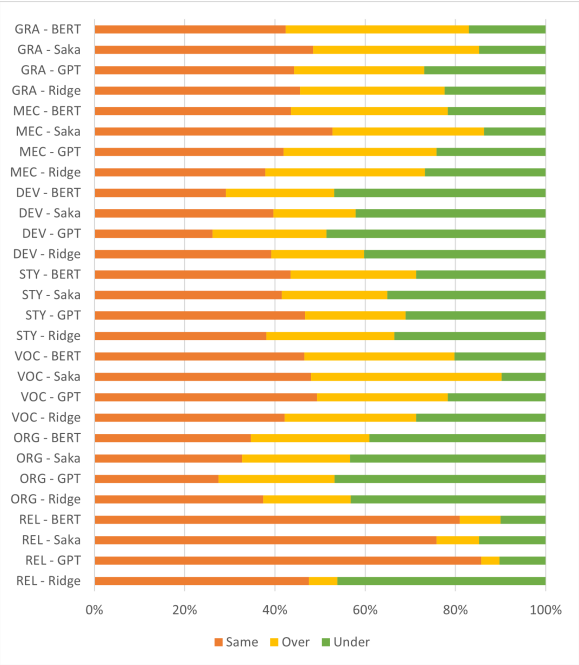

Figure 2: Testing-phase calibration across traits and models. Bars show the proportion of predictions that were exact matches (*Same*), overestimates (*Over*), or underestimates (*Under*); stacks sum to 100%.

6 Conclusion

In this study, we systematically compared multiple approaches to automated essay scoring, with particular emphasis on cross-genre generalization and alignment with trait-specific rubric criteria. By concatenating prompts, essays, and reasoning sequences with gold scores, our systems were explicitly encouraged to approximate human evaluation. Experimental results showed that fine-tuned BERT-based models achieved the highest QWK on the development set, slightly outperforming GPT-based few-shot prompting and classical ML approaches, while text-generation LLMs struggled under cross-genre conditions despite detailed CoT guidance.

The testing phase further demonstrated the robustness of GPT-based few-shot methods: providing in-context examples consistently improved performance, and translating rubrics and prompts into English enhanced trait calibration. Overall, this work shows that combining rubric-grounded reasoning with modern NLP architectures can yield reliable, trait-specific scoring of Arabic essays. These findings provide insights for practical deployment in educational contexts and point to future research directions focused on improving generalization, calibration, and interpretability in automated writing evaluation systems.

Acknowledgments

This work was supported by Cross-ministerial Strategic Innovation Promotion Program (SIP) on "Integrated Health Care System" Grant Number JPJ012425. We also extend our gratitude to the TAQEEM2025 organizers and reviewers for their valuable efforts in dataset preparation, rubric design, and evaluation. Their contributions were instrumental in enabling this research.

References

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 (ArabicNLP 2025)*, China.

May Bashendy, Salam Albatarni, Sohaila Eltanbouly, Eslam Zahran, Hager 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 (ArabicNLP 2024)*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *arXiv preprint* arXiv:2111.09543.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised Cross-lingual Representation Learning at Scale. *arXiv preprint* arXiv:1911.02116.

Go Inoue, Badr Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models. *arXiv preprint* arXiv:2103.06678.

Sakalti. 2024. Saka-14B: An Arabic large language model. https://huggingface.co/Sakalti/Saka-14B.

Edward J. Hu, Yelong Shen, Phillip Wallis, Z. Allen-Zhu, Yuanzhi Li, Sébastien Bubeck, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint* arXiv:2106.09685.

Appendix

Prompt Specification for GPT Scoring

This section includes the final structured prompt template used in our GPT experiments, ensuring the reproducibility of our results.

```
"role": "system",
"content": "You are an expert Arabic language
    teacher responsible
for evaluating Arabic essays written by
    students based on specific
traits and rubrics."}

fixed_user_message = {
"role": "user",
"content": f"""Think step-by-step about the
    following criteria, then start scoring the
    provided essay:
- Essays are evaluated on the following traits:
    {traits}.
- Each trait is described in this rubric (the
    dictionary of each trait is
    score-explanation pairs): {rubric}.
- The essay was written in response to the
    following prompt: {essay_prompt}
- Essay type: {essay_type}
- A score of zero is given if the response is
    completely memorized, copied from the
    prompt,
  if the student did not attempt to complete
    the task, or wrote something unrelated to
    the required topic.

Scoring steps:
1. Check the trait and its rubric.
2. Read the essay.
3. Provide a score.
4. Repeat from step 1 for each trait.
5. After scoring all traits, format the output
    as follows:
<trait_name>: <score>
Do not provide any additional text or
    explanation.

Example essays with scores:
{examples}

"""}
```

GPT-Based Few-Shot Prompting

This section provides detailed results for GPT-based few-shot prompting. We report trait-level evaluation metrics for different shot settings, complementing the aggregate results in the main text.

Trait	QWK	MSE	RMSE
Relevance	0.545	0.170	0.411
Organization	0.712	0.800	0.894
Vocabulary	0.653	0.783	0.881
Style	0.620	0.981	0.986
Development	0.629	0.761	0.872
Mechanics	0.482	1.038	1.009
Grammar	0.506	1.048	1.014

Table 7: GPT-0-shot: Trait-level evaluation results.

Trait	QWK	MSE	RMSE
Relevance	0.585	0.158	0.395
Organization	0.711	0.802	0.894
Vocabulary	0.646	0.798	0.889
Style	0.666	0.841	0.914
Development	0.647	0.716	0.846
Mechanics	0.477	1.023	1.004
Grammar	0.544	0.969	0.972

Table 8: GPT-1-shot: Trait-level evaluation results.

Trait	QWK	MSE	RMSE
Relevance	0.553	0.168	0.406
Organization	0.709	0.821	0.905
Vocabulary	0.633	0.837	0.911
Style	0.654	0.863	0.926
Development	0.640	0.750	0.865
Mechanics	0.515	0.972	0.979
Grammar	0.580	0.908	0.944

Table 9: GPT-10-shot: Trait-level evaluation results.

Classical ML Features

This section lists the handcrafted linguistic features extracted from essays, which were combined with embeddings in the classical ML experiments.

Feature
Total words in the essay
Unique words in the essay
Punctuation marks count
Total sentences in the essay
Average word length (chars)
Average words per sentence
Total characters in the essay
Stopwords count
Total bigrams
Total trigrams
Unique bigrams
Unique trigrams
Unique/total bigrams ratio
Unique/total trigrams ratio

Table 10: Linguistic features extracted from essays.