912 at TAQEEM 2025: A Distribution-aware Approach to Arabic Essay Scoring

Trong-Tai Dam Vu, Dang Van Thin
University of Information Technology-VNUHCM,
Vietnam National University, Ho Chi Minh City, Vietnam
{taidvt,thindv}@uit.edu.vn

Abstract

We present our system for TAQEEM 2025 Task A on Arabic automatic essay scoring. Building on a pretrained Arabic encoder, our work focuses on two key design axes: (i) replacing the standard linear head with a lightweight multi-layer perceptron (MLP) and (ii) optimizing with distribution-aware objectives. We introduce a Weighted Mean-Squared Error loss, which assigns higher weights to less frequent scores to counteract the imbalanced, bell-shaped score distribution of the training data. On the official development folds, our system outperforms the baseline on Quadratic Weighted Kappa. Our findings underscore the importance of tailoring objective functions to specific data characteristics for achieving state-of-the-art results in AES.

1 Introduction

Automatic essay scoring (AES) aims to predict human-assigned holistic scores for free-form writing. The TAQEEM 2025 shared task focuses on Arabic AES (Task A), providing standardized data and an agreement-focused evaluation via QWK (Bashendy et al., 2025).

In line with the shared task guidelines, our goal is to conduct a transparent and reproducible study of what modifications yield reliable gains. Our work investigates two primary questions: 1) What is the optimal architecture for the prediction head? 2) Can a distribution-aware objective function, designed to address the specific characteristics of the score data, offer an advantage over standard regression losses?

Our system builds on the pretrained ArabicBERT v02 encoder (Antoun et al., 2020), which is based on the Transformer architecture (Vaswani et al., 2017; Devlin et al., 2019). We systematically explore the impact of MLP head depth and compare several objective functions. Our key contribution is the successful application of a Weighted MSE loss, which addresses the inherent imbalance in the dataset's score distribution. This simple, well-analyzed approach with a carefully chosen objective function can achieve state-of-the-art results.

2 Background

Task Description. TAQEEM 2025 is a shared task on evaluating Arabic student writing. We participated only in Task A (holistic AES). The official evaluation metric is QWK(Cohen, 1960).

The official dataset is composed of a training set of 426 Arabic essays and a test set of 840 essays, each covering two distinct writing prompts: explanatory and persuasive. The score distribution is bell-shaped and imbalanced toward mid-range scores, as shown in Figure 1. This observed imbalance is the primary motivation for our experiments with a weighted loss function, as standard MSE can be biased towards predicting the more frequent, mid-range scores.

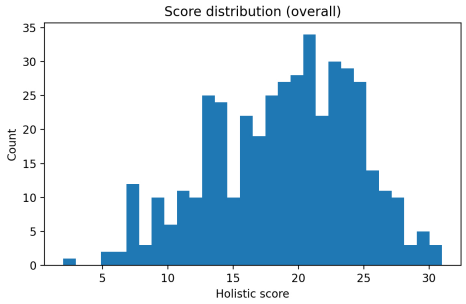

Figure 1: Overall distribution of holistic scores in the training set. The bell-shaped curve centered on mid-range scores (approx. 18-25) motivated our use of a weighted loss objective.

Related work. Automated Essay Scoring (AES) has evolved from early feature-based systems (Attali and Burstein, 2006) to deep learning. Currently, fine-tuning large pretrained Transformers like BERT (Devlin et al., 2019) is the state-of-

Proceedings of The Third Arabic Natural Language Processing Conference, pages 983–988

the-art approach, consistently achieving top results (Ludwig et al., 2021). Our work builds directly on this paradigm, leveraging a powerful pretrained Arabic model.

While much research focuses on English, Arabic AES is an active area (Ghazawi and Simpson, 2024), with models like AraBERT (Antoun et al., 2020) providing a strong foundation for the task. The paradigm of pre-training on large text corpora was popularized by both decoder-focused generative models like the GPT series (Radford et al., 2018, 2019; Brown et al., 2020) and encoder-focused models like BERT. Methodologically, our primary contribution—the use of a Weighted MSE loss—is inspired by established techniques for learning from imbalanced datasets (Cao et al., 2019; Ren et al., 2018), which we adapt to address the specific bell-shaped score distribution inherent in AES data.

3 System Overview

Our approach is centered on fine-tuning the AraBERTv02 model (Antoun et al., 2020). The overall architecture is depicted in Figure 2.

3.1 Backbone and Inputs

The core of our system is the AraBERT-v0.2 (Antoun et al., 2020) encoder, a pre-trained language model optimized for Arabic. To effectively present the task to the model, we explored two distinct input representations.

The first configuration, which we term Essay-only, provides the model with only the student's essay text. This approach tests the model's ability to infer scoring criteria directly from the text itself.

The second configuration, Essay and Prompt, uses a concatenation of the writing prompt and the essay text as input. This method provides the model with explicit context about the task's requirements. The choice between these two representations was determined empirically, as detailed in our ablation study.

3.2 Prediction Head

The standard approach for regression tasks with BERT-like models is to use a single linear layer (a regression head) on top of the `[CLS]` token representation. To explore if a more complex function could better map the learned features to a score, we experimented with replacing this linear head with a lightweight Multi-Layer Perceptron (MLP).

We systematically varied the depth of this MLP by changing the number of hidden layers, denoted by k. We tested configurations within the set $k \in \{0, 1, 2, 3\}$. The case where $k = 0$ is equivalent to the standard linear head, which serves as a direct baseline for this experiment. The optimal depth of the MLP was determined empirically, as we detail in our ablation studies.

3.3 Objectives

Our primary contribution in this work lies in the design and application of a distribution-aware objective function tailored to the specific characteristics of the AES dataset. We describe our proposed Weighted MSE (wMSE) loss below. To validate its effectiveness, we benchmarked it against the standard MSE loss and an agreement-aware MSE+QWK objective in our ablation studies.

Our proposed Weighted MSE (wMSE) loss is designed to counteract the imbalanced (Cao et al., 2019; Ren et al., 2018), bell-shaped score distribution of the training data. The core idea is to assign a weight, w_s, to each possible integer score s, where the weight is inversely proportional to the score's frequency in the training corpus $\mathcal{D}_{\text{train}}$. This forces the model to place greater importance on correctly predicting essays with rare scores.

First, for each unique integer score s in the range $[s_{\min}, s_{\max}]$, we calculate its frequency $N_s = |\{y_i \in \mathcal{D}_{\text{train}} \mid y_i = s\}|$. The weight w_s is then defined as the inverse of this frequency:

$$w_s = \frac{1}{N_s} \tag{1}$$

These weights are pre-calculated once over the entire training set. For a given batch of B samples, the Weighted MSE loss, $\mathcal{L}_{\text{wMSE}}$, is computed as the mean of the squared errors, where each error term is multiplied by the weight corresponding to its ground-truth label. For a prediction $\hat{y}_i$ and a true label y_i, the loss is:

$$\mathcal{L}_{\text{wMSE}} = \frac{1}{B} \sum_{i=1}^{B} w_{y_i} \cdot (\hat{y}_i - y_i)^2 \tag{2}$$

Since the ground-truth labels y_i are integers, the corresponding weight w_{y_i} can be retrieved directly.

To benchmark our proposed wMSE loss, we also evaluated two other objective functions. The standard **Mean Squared Error (MSE)** served as our main regression baseline. Additionally, we experimented with a combined **MSE+QWK** objective.

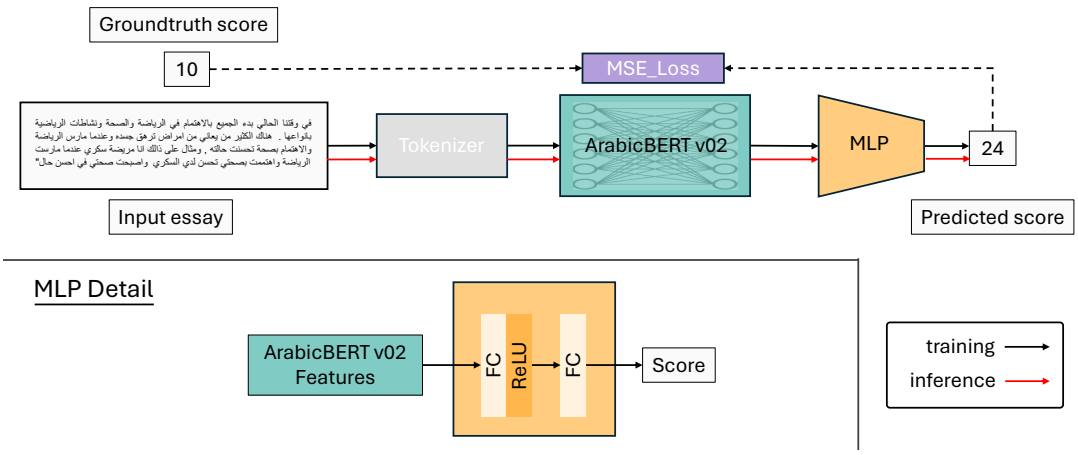

Figure 2: Our system architecture, showing the ArabicBERT encoder followed by a one-layer MLP head for score prediction.

The MSE+QWK loss function incorporates a differentiable surrogate of Quadratic Weighted Kappa (QWK) directly into the training objective. This aims to align the model's optimization more closely with the final evaluation metric. Standard QWK is non-differentiable because it is calculated from a confusion matrix, which requires rounding the model's continuous regression outputs (e.g., 13.7) into discrete integer predictions (e.g., 14). This rounding step prevents gradients from flowing during backpropagation.

To create a differentiable surrogate, we implemented a "soft" version of the QWK calculation. The process begins with soft assignment, where instead of rounding, each continuous prediction is represented as a soft probability distribution over all possible integer scores. This is achieved by calculating the distance from the prediction to each integer class center (e.g., the distances from 13.7 to) and converting these distances into a probability vector using a softmax function. A prediction of 13.7 will thus have high probabilities assigned to the nearby classes 13 and 14. This process is also applied to the ground-truth labels, naturally handling non-integer scores. These resulting probability vectors are then used to construct a "soft" confusion matrix for the batch by summing the outer product of each prediction-label vector pair. With this fully differentiable confusion matrix, the observed and expected agreement can be calculated using standard matrix operations, allowing gradients to flow back through the entire QWK formula to the model's outputs.

The combined loss is then formally defined as:

$$\mathcal{L}_{\text{MSE+QWK}} = \mathcal{L}_{\text{MSE}} + (1 - \text{QWK}) \quad (3)$$

where QWK is the fully differentiable surrogate of the QWK metric, calculated as described above.

4 Experimental Setup

To ensure reproducibility and isolate the impact of our design choices, we conducted a systematic ablation study. All models were fine-tuned using the AdamW optimizer with a learning rate of 5e-5, a batch size of 16, for up to 100 epochs. The best checkpoint for each run was selected based on the highest average QWK on the development folds. Our study evaluated three primary design axes: 1) the objective function (our proposed Weighted MSE vs. standard MSE and MSE+QWK), 2) the MLP head architecture (varying the number of hidden layers $k \in \{0, 1, 2, 3\}$), and 3) the input type (essay-only vs. prompt+essay). The results presented in Table 1 compare the best-performing configuration found for each objective to ensure a fair and comprehensive analysis.

5 Results and Analysis

Our main experimental results are summarized in Table 1, which presents a comprehensive ablation study. The final official scores on the private test set are shown in Table 2.

5.1 Overall Performance

Our best single model achieved an average QWK of **0.766** on the development set (0.784 on Fold 1 and 0.747 on Fold 2). As shown in Table 2, our

Configuration	Loss	Input	MLP Depth (k)	Dev QWK (Fold 1)	Dev QWK (Fold 2)	Avg QWK
Baseline (Linear Head)	MSE	Essay	0	0.705	0.727	0.716
Our Best Model	**Weighted MSE**	**Essay**	**1**	**0.784**	0.747	**0.766**
Ablation on Objective Function						
- use MSE Loss	MSE	Essay	3	0.768	**0.753**	0.761
- use MSE+QWK Loss	MSE+QWK	Essay	2	0.741	0.752	0.747
Ablation on Architecture						
- use 2 hidden layers	Weighted MSE	Essay	2	0.768	0.752	0.760
- use 3 hidden layers	Weighted MSE	Essay	3	0.781	0.740	0.761
Ablation on Input Type						
- use Prompt+Essay	Weighted MSE	Prompt+Essay	1	0.764	**0.753**	0.759

Table 1: Main results and a comprehensive ablation study on the development set. Performance is measured by Quadratic Weighted Kappa (QWK), averaged over two folds. The table compares our best model (in bold) against the official baseline. It also presents three sets of ablation studies, each starting from our best model's configuration and varying a single component: the objective function, architecture, or input type.

Configuration	**QWK (Fold 9)**	**QWK (Fold 10)**	**Official QWK**	**Official RMSE**
Baseline	0.608	0.670	0.639	5.372
Our Best Model	**0.662**	**0.683**	**0.673**	**5.333**

Table 2: Final performance on the private test set, comparing our best model to the official baseline. We report the official QWK and RMSE, along with the QWK scores from the last two cross-validation folds.

best model significantly outperforms the baseline on the private test set, confirming the effectiveness of our approach on unseen data.

5.2 Analysis of Findings

Our ablation studies, detailed in Table 1, provide several key insights into the factors driving performance.

Impact of Objective Function. The choice of objective function is the most critical factor for success. Our **Weighted MSE** model (Avg QWK 0.766) significantly outperforms the best configurations using standard MSE (0.761) and MSE+QWK (0.747). This confirms our hypothesis that explicitly addressing the dataset's imbalanced score distribution is crucial for achieving top performance. By forcing the model to pay more attention to less frequent scores, the wMSE objective mitigates the model's natural bias towards the populated mean of the distribution.

Interplay between Architecture and Objective. The architectural ablation study reveals a clear relationship between our proposed wMSE objective and the model's architectural complexity. As shown in Table 1, the performance of the wMSE-trained model peaks with a 1-layer MLP ($k = 1$). Performance degrades when the architecture is too simple ($k = 0$, a standard linear head) and also when it becomes overly complex ($k = 2, 3$).

This suggests that the wMSE loss, by increasing the importance of rare scores, creates a more challenging optimization landscape than standard MSE. A simple linear head ($k = 0$) appears to lack sufficient capacity to fully model the nuances of this distribution-aware objective. Conversely, deeper MLPs ($k = 2, 3$) seem prone to overfitting on this specialized task. Therefore, our results indicate that the benefits of a distribution-aware objective are best realized when paired with an architecture of appropriate, non-trivial complexity.

Impact of Input Type. The ablation on input type confirms that an essay-only approach is optimal for our best model. Including the prompt text resulted in a performance drop (from 0.766 to 0.759 Avg QWK). While the prompt provides essential context for evaluating aspects like relevance, our empirical results suggest that its explicit inclusion via concatenation is suboptimal in this setup. We hypothesize two potential reasons for this counter-intuitive finding. First, since the dataset contains only two distinct prompts, the model may be able to implicitly infer the necessary context from the essay's topic, vocabulary, and structure alone, making the explicit prompt text redundant. Second, concatenating the prompt might unfavorably shift the model's attentional focus. The model may allocate too much of its limited attention capacity to the initial prompt tokens, thereby diluting its focus on the nuanced linguistic features distributed

throughout the essay itself.

5.3 Error Analysis

To better understand our model's limitations, we analyzed its prediction errors on the development set. Figure 3 presents a binned confusion matrix of our best model's predictions, which visually confirms our two primary failure modes:

1. **Near-Boundary Confusion:** The strong concentration of predictions along the main diagonal and its adjacent cells is the most prominent pattern. This shows that the model's primary error is confusing similar, adjacent score ranges (e.g., predicting a score in the 17-21 bin for a true score in the 22-26 bin). This is a classic challenge in regression-based AES.

2. **Off-Prompt Responses:** The dataset contains some essays that do not fully address the prompt. Our model, trained on holistic writing quality, sometimes assigns a moderate score to a well-written but off-topic essay, whereas a human grader might penalize it more heavily for being non-responsive to the task.

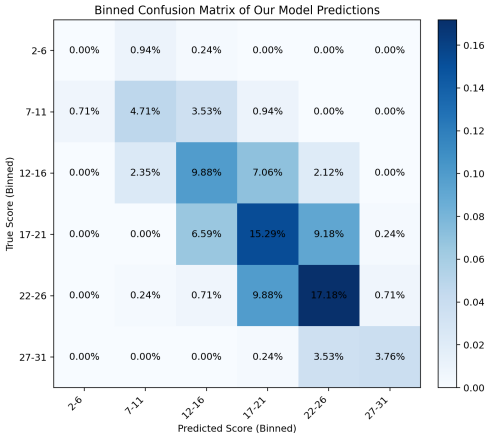

Figure 3: Binned confusion matrix of predictions on the development set. The concentration of values around the main diagonal highlights near-boundary confusion as the primary error type.

6 Conclusion

We presented our system for the TAQEEM 2025 Task A on Arabic AES. Our success was primarily driven by a custom Weighted MSE (wMSE) objective, designed to counteract the imbalanced, bell-shaped score distribution of the training data. Our analysis revealed a crucial finding: this distribution-aware objective not only significantly boosted performance but also achieved its best results with a simpler 1-layer MLP architecture compared to the deeper models required by standard MSE. Our work underscores the value of tailoring objective functions to data characteristics and demonstrates that a simple, well-analyzed approach can achieve state-of-the-art results in AES.

7 Acknowledgements

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

References

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 (OSACT)*.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Advances in neural information processing systems (NeurIPS)*, volume 33, pages 1877–1901.

Kaidi Cao, Chen Wei, Adrien Gaidon, Niki Arechiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*.

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*.

Sabrina Ludwig, Christian Mayer, Christopher Hansen, Kerstin Eilers, and Steffen Brandt. 2021. Automated essay scoring using transformer models. *Psych*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and 1 others. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. 2018. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning (ICML)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*.