Journal of Theoretical and Applied Information Technology
30th November 2025. Vol.103. No.22

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

TRANSFORMERS IN ARABIC SHORT ANSWER GRADING: BRIDGING LINGUISTIC COMPLEXITY WITH DEEP LEARNING

WAEL HASSAN GOMAA[1], MENA HANY[2], EMAD NABIL[3,*] , ABDELRAHMAN E. NAGIB[4], HALA ABDEL HAMEED [5,6]

[1] Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef, Egypt
[2] King Fahd University of Petroleum & Minerals, Dhahran, KSA
[3] Faculty of Computer and Information Systems, Islamic University of Madinah, Madinah 42351, KSA
[4] Faculty of Computer Science, October University for Modern Sciences and Arts (MSA), Giza, Egypt
[5] Faculty of Computers and Artificial Intelligence, Fayoum University, Fayoum, Egypt
[6] Khaybar Applied College, Taibah University, Madinah 42353, KSA

E-mail: [1] wael.goma@gmail.com, [2] g202411920@kfupm.edu.sa, [3] emadnabil@iu.edu.sa, [4]abezzeldin@msa.edu.eg, [5] Ham07@fayoum.edu.eg, [6] hammohamed@taibahu.edu.sa
Corresponding Author Mail: emadnabil@iu.edu.sa

ABSTRACT

Automating the evaluation of Arabic short answers is a crucial step in advancing educational technology, as it enables rapid feedback, consistent scoring, and a significant reduction in educators' workload. However, the structural richness and semantic complexity of Arabic—characterized by its extensive morphology, flexible word order, and diverse vocabulary—make reliable grading especially challenging. To address these difficulties, this study introduces a three-stage framework built upon fine-tuned transformer architectures. In the first stage, both the question and the learner's response are encoded into dense semantic embeddings. The second stage applies comprehensive fine-tuning to a pre-trained transformer model, allowing it to capture task-specific nuances and better represent the intricate patterns of Arabic. In the final stage, a regression layer generates a numerical score, which is then compared against the human-assigned reference grade for evaluation. The proposed framework was rigorously tested on two benchmark datasets for Arabic short answer grading, AR-ASAG and Philosophy. Experimental results demonstrated strong performance, achieving Pearson correlation scores of 0.85 and 0.97, respectively, and outperforming previously reported state-of-the-art methods. These outcomes confirm the effectiveness of transformer-based models in handling the linguistic subtleties of Arabic while also demonstrating their scalability and adaptability across domains. Overall, the findings position fine-tuned transformers as a promising foundation for building accurate, efficient, and equitable automated grading systems in Arabic educational contexts.

Keywords: *Transformers, Arabic Short Answer Grading, Natural Language Processing, Deep Learning, Model Fine-Tuning*

1. INTRODUCTION

Grading student responses is often a demanding and time-intensive responsibility for educators. Advances in Natural Language Processing (NLP) have the potential to reduce this workload by enabling automated systems to interpret questions, compare them with model answers, and generate appropriate scores. Various NLP-based grading methods have been explored over the years [1], covering a spectrum of assessment formats — from straightforward multiple-choice questions (MCQs) to more complex short-answer and essay evaluations [2]. This work focuses exclusively on short-answer grading, which generally involves two main objectives: assigning an appropriate score and offering feedback [2]. The present study concentrates on the grading function.

A key motivation for building automated short-answer grading systems is to minimize human errors that may occur due to fatigue, stress, or grader bias toward certain individuals [3]. In addition, automation significantly accelerates the

evaluation process. Unlike human assessors, who must complete grading for all submissions before releasing results, automated systems can provide immediate feedback [4]. Although many automated short-answer grading systems exist, the vast majority target the English language. Systems designed for Arabic remain scarce, and available Arabic datasets are limited compared to English resources [2], [5], [6].

This research is guided by the following questions:

1- How can transformer-based models be adapted for effective Arabic Short Answer Grading (ASAG)?

2- What is the impact of fine-tuning pre-trained language models on their ability to grade Arabic short answers, given the language's unique structural challenges?

3- Which transformer architecture offers the best accuracy and robustness for ASAG tasks?

4- Can multilingual transformers outperform Arabic-specific models in capturing semantic and syntactic details?

5- What trade-offs exist between model size, architecture, and grading accuracy, and how do these affect overall efficiency?

The contributions of this work are fourfold. First, it presents an extensive evaluation of multiple transformer-based models (both multilingual and Arabic-specific) using two widely recognized ASAG datasets: AR-ASAG and Philosophy. This comparison provides insight into how Arabic-optimized models differ from those trained on multilingual corpora. Second, the study validates the benefit of fine-tuning pre-trained models for ASAG tasks. Third, a novel framework is proposed that achieves new state-of-the-art performance, with XLM-RoBERTa-large emerging as the top-performing internal model.

Finally, the work analyzes trade-offs between model complexity, architecture, and computational demands, offering guidance for selecting suitable models according to task needs and resource constraints. Overall, this research contributes to advancing transformer-based approaches in Arabic NLP and establishes a basis for extending automated grading to other languages and educational contexts.

The paper is structured as follows: Section 2 reviews prior research on Arabic short-answer grading. Section 3 describes the proposed framework and methodology. Section 4 reports experimental results and analysis. Section 5 discusses the implications of the findings. Section 6 concludes with future research directions.

2. LITERATURE REVIEW

Research on automated short-answer grading has been dominated by work in the English language, with a considerable body of studies exploring diverse techniques [7]–[14]. However, the Arabic language has received comparatively less attention, and only a limited number of ASAG systems have been developed for it. This section reviews key contributions focusing on Arabic short-answer grading.

In [15], the authors proposed a method that grades short answers by comparing student responses with reference answers using various similarity measures. These include direct text matching, leveraging background knowledge, and analyzing word meanings, without relying on rigid templates or detailed syntactic parsing. Their system incorporated preprocessing, similarity computation, and score scaling through support vector regression, producing scores between 0 and 5. Evaluations on the Cairo University, Texas, and Extended Texas datasets yielded Pearson correlation scores of 0.84, 0.59, and 0.55, respectively, with corresponding RMSE values of 0.89 and 0.91, where applicable.

In [16], the AR-ASAG dataset was introduced as a benchmark for Arabic short-answer grading. It contains 2,133 pairs of student and ideal responses, offered in multiple formats for flexibility. The authors applied the COALS algorithm to assess semantic similarity, with variations that included preprocessing, word weighting, and stemming. Their approach achieved Pearson correlation scores of 0.7037 on AR-ASAG and 0.7220 on the SemEval dataset, with RMSE values around 1.03.

The work in [17] compared 14 different similarity metrics, both string-based and corpus-based, for evaluating student answers against a reference. They experimented with preprocessing methods such as stop word removal and stemming using the ISRI Arabic Stemmer. The system supported both holistic and partitioned answer evaluation, providing immediate feedback to learners. Tested

on the Philosophy dataset, the approach achieved a Pearson correlation of 0.862 and an RMSE of 0.76.

In [18], student answers were preprocessed by removing extraneous characters and applying lemmatization to enhance efficiency. Two baseline approaches were considered: a reference-based cosine similarity method and a TF-IDF + SVM classifier. Deep learning architectures, including RNN, LSTM, and Bi-LSTM, were later tested, followed by fine-tuning ELECTRA and BERT. On AR-ASAG, ELECTRA achieved the best performance with a Quadratic Weighted Kappa (QWK) score of 0.78, outperforming the reference-based baseline (0.46).

The system in [19] utilized tokenization, stop-word removal, and lemmatization before applying Latent Semantic Analysis (LSA) to compare student and model answers. Term weighting and singular value decomposition were employed to capture semantic relationships, with cosine similarity used for scoring. The AR-ASAG dataset evaluation yielded an F1-score of 82.82%, a recall of 81.41%, and a precision of 83.23%.

In [20], a dataset of 15,000 short answers (graded on a 0–5 scale) was used to train a deep learning model. The process included cleaning, tokenization, stemming, and Bag-of-Words vectorization. The authors used an LSTM layer optimized with Grey Wolf Optimization (GWO), achieving an RMSE of 0.807, an R^2 of 0.826, and a Pearson correlation of 0.909.

The Ans2vec approach in [21] employed a pre-trained Skip-Thought model to vectorize both the student's and the reference answers. Mathematical operations on the vectors, such as element-wise multiplication and absolute difference, were performed before feeding them to a linear regression model for scoring. This method reached a Pearson correlation of 0.63 and an RMSE of 0.91 on the Cairo University dataset.

In [22], the Longest Common Contiguous Subsequence (LCCS) technique was combined with Arabic WordNet to enhance answer evaluation. After preprocessing, synonyms from Arabic WordNet were added to expand semantic coverage. The method was tested on a dataset of 330 student responses, obtaining a Pearson correlation of 0.94 and an RMSE of 0.81.

Finally, [23] explored short-answer grading in the context of Environmental Science. Their dataset included varied question types and scoring schemes. They evaluated multiple similarity measures, concluding that N-gram performed best among string-based methods, while the DISCO algorithm excelled in corpus-based approaches. Translating Arabic responses into English improved performance for corpus-based methods due to richer linguistic resources in English. Their combined approach achieved a Pearson correlation of 0.83 and an RMSE of 0.75.

Table 1 summarizes the reviewed studies, detailing the datasets used and their reported results.

3. Proposed Methodology

This section outlines the datasets, experimental setup, and the proposed framework for Arabic Short Answer Grading (ASAG). The approach integrates recent advancements in transformer architectures with a task-specific fine-tuning strategy, designed to capture both the syntactic and semantic intricacies of Arabic responses.

3.1 Datasets

Two benchmark datasets were used to evaluate the proposed method: AR-ASAG [16] and the Philosophy dataset [17]. These datasets were chosen because they are widely referenced in ASAG research, enabling direct performance comparisons with previous works.

3.1.1 AR-ASAG

The AR-ASAG dataset [16] is specifically tailored for automated Arabic short-answer grading and includes 48 distinct questions related to cybercrimes. These questions fall into five categories: definition, explanation, consequence identification, justification, and difference identification. Each question is paired with multiple student responses, along with a model (ideal) answer. For each student answer, two human-assigned grades (ranging from 0 to 5) are provided, and the average grade serves as the gold standard for evaluation.

The dataset comprises 2,133 (question, student answer, model answer) triplets. Response lengths

vary, with a minimum of 1 word, an average of approximately 20 words, and a maximum of 76 words. This variability allows the dataset to test model robustness across very short and relatively long responses. The AR-ASAG dataset is particularly challenging due to variations in vocabulary choice, use of synonyms, and differences in sentence structure, all of which require models to capture deep semantic relationships rather than relying solely on surface-level word matching.

Table 1: Comprehensive overview of the related works

Ref No.	Method	Dataset Used	Results
[15]	Similarity scores with Support Vector Regressor	Texas Extended Texas Cairo university	Texas:0.59 Pearson's correlation Extended Texas:0.55 Pearson's correlation and 0.91 RMSE Cairo university:0.84 Pearson's correlation and 0.89 RMSE
[16]	COALS algorithm	AR-ASAG SemEval	AR-ASAG:0.7037 Pearson's correlation and 1.0454 RMSE SemEval:0.7220 Pearson's correlation and 1.03 RMSE
[17]	14 different string and corpus similarity algorithms	Philosophy	0.862 Pearson's correlation and 0.76 RMSE
[18]	Fine-tuning Electra Model	AR-ASAG	0.78 QWK (Quadratic Weighted Kappa)
[19]	LSA model	AR-ASAG	F1-Score 82.82%, Recall 81.41%, Precision 83.23%
[20]	LSTM with grey wolf optimization	private dataset	RMSE 0.807, R-Square 0.826, Pearson's correlation 0.909
[21]	Skip-thought model with Linear Regression model	Cairo university	0.63 Pearson's correlation and 0.91 RMSE
[22]	LCCS with Arabic wordnet	private dataset	0.81 Pearson's correlation and 0.94 RMSE
[23]	N-gram and DISCO algorithm	Environmental Science	0.83 Pearson's correlation and 0.75 RMSE

Table 2: A sample from the AR-ASAG dataset along with its English translation. The maximum score is five.

Question	Model Answer	student answer	Grader 1	Grader 2	avg. mark
عرف مصطلح الجريمة الإلكترونية Define the term cybercrime	هي كل سلوك غير قانوني يتم باستخدام الأجهزة الإلكترونية) الهاتف، الكمبيوتر، الانترنت (ينتج عنه حصول المجرم على فوائد مادية أو معنوية مع تحميل الضحية خسارة وغالبا ما يكون هدف هذه الجرائم هو القرصنة من أجل سرقة أو إتلاف المعلومات وتكون عادة الانترنت أداة لها أو مسرحا لها It is any illegal behavior carried out using electronic devices	**Student#1 answer** هي سلوك غير أخلاقي يتم عن طريق وسائل الكترونية يهدف الى عائدات مادية و يسبب اضرارا للضحية It is unethical behavior carried out through electronic means that aims to gain financial returns and causes harm to the victim	2.5	3.5	3

	(phone, computer, internet) that results in the criminal obtaining material or moral benefits while the victim incurs a loss. The goal of these crimes is often hacking in order to steal or destroy information, and the internet is usually a tool or a stage for them.	**Student#2 answer** هي كل سلوك غير أخلاقي يتم بواسطة الاجهزة الالكترونية ينتج عنها حصول المجرم على فوائد مادية او معنوية مع تحصيل الضحية خسارة مقابلة، هدفها القرصنة من اجل سرقة او اتلاف المعلومات It is all unethical behavior carried out by electronic devices that results in the criminal obtaining material or moral benefits while the victim receives a corresponding loss. Its goal is hacking in order to steal or destroy information.	5	5	5

Table 3: A sample from the Philosophy dataset and its translation into English. The maximum grade is ten.

Question	Model Answer	Student Answer	Average grade of two graders
وضح أهمية الفلسفة بالنسبة للفرد؟ Explain the importance of philosophy for the individual?	تعميق الوعى لدى الفرد تجعل الإنسان يفهم حياته و يدرك مكانته فى الوجود و فى المجتمع و تحدد أهدافه و توقظه من نومه العميق الإرتفاع بالمستوى العقلى و حل المشكلات لأنها تعتمد على التفكير العقلى و دراسة وجهات نظر الفلاسفة في مختلف المشاكل تساعد الفرد في حل مشاكله الخاصة. تأكيد و ترسيخ الإيمان و الثقة بقدرة الله تسعى الفلسفة إلى بث الثقة فى النفس وتأكيد إيمان الفرد بالله والدين على أساس من الاقتناع العقلى لا الإيمان الفطرى الموروث. Deepening the individual's awareness, It makes a person understand his life, realizes his place in existence and in society, defines his goals, and awakens him from his deep sleep. Raising the mental level and solving problems Because it depends on rational thinking and studying the views of philosophers on various problems, it helps the individual solve his own problems. Confirming and consolidating faith and confidence in God's power Philosophy seeks to instill self-confidence and affirm the individual's belief in God and religion on the basis of rational conviction, not innate, inherited faith.	**Student#1 answer** الإرتقاء بالمجتمع ككل .كشف مشكلات المجتمع .تحديد الإطار الأيديولوجى الذى ينظم السلوك الوطنى للأفراد. Improving society as a whole. Revealing community problems. Determine the ideological framework that regulates the national behavior of individuals.	2
		Student#2 answer تعمل الفلسفة على إشباع رغبة الفرد فى المعرفة و التأمل و حب الإستطلاع و ذلك من خلال الإجابة على التساؤلات التى تدور فى ذهنه مثل وجوده و هويته و قيمته. Philosophy works to satisfy the individual's desire for knowledge, contemplation, and love of curiosity by answering the questions that revolve in his mind, such as his existence, identity, and value.	3.5

3.1.2 Philosophy Dataset

The Philosophy dataset [17] contains 50 philosophical questions, each accompanied by 12 student responses, yielding 600 total answers. Each answer is scored independently by two human evaluators on a 0–10 scale, and the average is used as the ground truth label. Answers in this dataset are generally longer than in AR-ASAG, with a minimum of 3 words, an average of 25 words, and a maximum of 104 words.

Philosophy questions tend to require abstract reasoning and interpretation, making the dataset a

useful benchmark for testing a model's ability to assess responses where semantic equivalence is more important than lexical overlap.

Both datasets are split into 80% training and 20% testing, consistent with prior studies [16], [17].

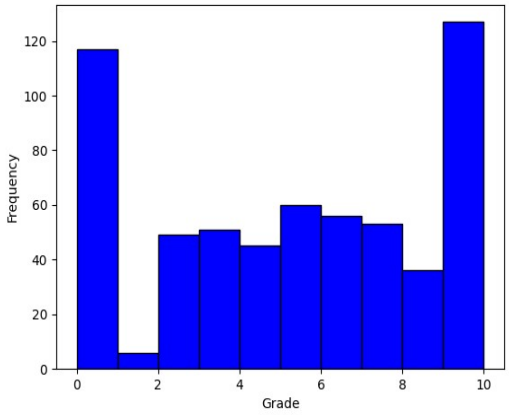

Fig. 1. The Grades' Distributions of the AR-ASAG dataset

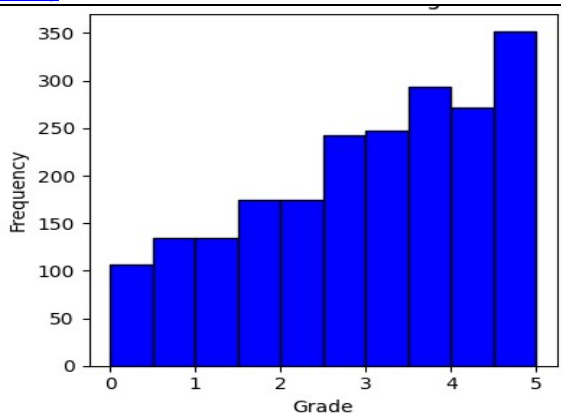

Fig. 2. The Grade Distribution of the philosophy dataset

3.2 Framework Overview

The proposed framework is designed as a three-stage pipeline, as shown in Figure 3:

1- **Embedding Layer**: Converts the model answer and the student's answer into dense vector representations.
2- **Fine-Tuning Stage**: Adapts the internal parameters of a pre-trained transformer to the ASAG task.
3- **Regression Layer:** Produces a continuous score, which is compared to the reference grade using correlation-based metrics.

This modular design allows flexibility in substituting different transformer architectures in the embedding and fine-tuning stages without altering the overall workflow.

3.2.1 Embedding Layer

Traditional transformer-based models such as BERT [33] or RoBERTa [24] are typically optimized for single-sentence classification or token-level prediction tasks. In the ASAG setting, however, the model must simultaneously process two related but distinct text inputs: the ideal (model) answer and the student answer.

To address this, the input pipeline was adapted to handle paired text sequences. Both answers are concatenated with special tokens (e.g., [CLS], [SEP]) so the model can jointly encode them while preserving their semantic relationship. This dual-input encoding enables the transformer to learn nuanced mappings between the model answer and the student's phrasing, even when synonyms, paraphrases, or varied syntactic structures are present.

3.2.2 Fine-Tuning Strategy

Unlike approaches that only adjust the final classification or regression layers, our method performs full-parameter fine-tuning — unfreezing all transformer layers. This allows the network to adapt deep, context-aware representations specifically for ASAG, which often requires capturing subtle semantic shifts and domain-specific terminology.

Fine-tuning is conducted using the AdamW optimizer [36] with a learning rate of 1e-5 and a batch size of 16. The number of training epochs is set to 20, with early stopping based on validation set performance. The best-performing checkpoint is saved for evaluation.

This comprehensive fine-tuning is particularly beneficial for Arabic because its rich morphology, flexible word order, and complex syntactic patterns

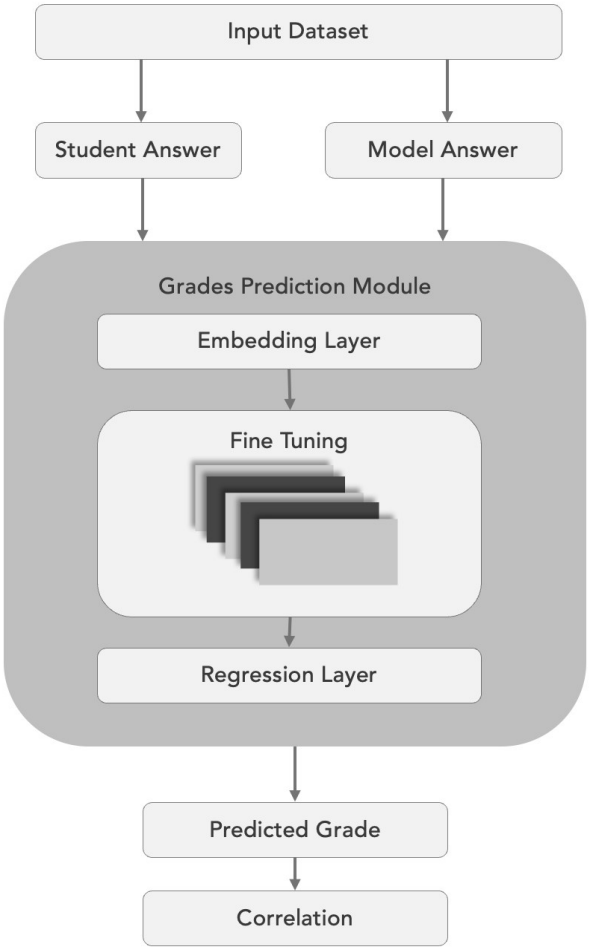

Figure 3: Proposed Model Pipeline

require models to adjust internal attention distributions more extensively than is often necessary for English.

3.2.3 Regression Layer

The final stage is a regression head that outputs a single scalar value, the predicted score for the student's answer. The regression output is trained to minimize the Mean Squared Error (MSE) between the predicted and gold-standard grades.

Performance is assessed using Pearson's correlation coefficient, which measures the linear relationship between predicted and actual scores. This metric is widely used in ASAG research because it captures the strength of alignment between automated and human scoring [15]–[17].

3.3 Model Selection Criteria

The set of transformer models evaluated in this study was chosen according to two main criteria:

1- **Arabic Competence:** Models must be either pre-trained on large Arabic corpora or be multilingual models with documented high performance on Arabic NLP tasks.
2- **Parameter Efficiency:** Models were restricted to those with fewer than one billion parameters to ensure reasonable training and inference times, while still allowing evaluation of large models like XLM-RoBERTa-large.

Selected models include Arabic-specific variants such as AraBERTv2 [25] and asafaya/bert-large-arabic [26], as well as multilingual models such as mDeBERTa-v3-base [27], [28] and XLM-RoBERTa [24].

3.4 Hardware and Experimental Setup

All experiments were conducted on a server with 128 GB RAM, an NVIDIA GPU with 24 GB VRAM, and a 20-core Intel Xeon processor. This configuration ensured that even the largest models could be fine-tuned without memory overflow issues.

4. EXPERIMENTS AND RESULTS

This section details the evaluation process of the proposed framework, including experimental settings, comparative analyses, and performance results on the AR-ASAG and Philosophy datasets. The aim was to rigorously test whether the integration of fine-tuned transformer models improves Arabic Short Answer Grading (ASAG) performance compared to existing state-of-the-art approaches.

4.1 Experimental Settings

All experiments were conducted using the hardware configuration described in Section 3.4. This setup was sufficient to train large-scale models like XLM-RoBERTa-large without resorting to gradient accumulation or severe batch-size reductions.

Table 4: The selected transformers' models for the experiments.

Model Name	Multi-lingual support Arabic	Arabic	No. of params. in Millions
xlm-roberta-large [24]	✓		561
aubmindlab/bert-large arabertv2 [25]		✓	371
asafaya/bert-large-arabic [26]		✓	338
microsoft/mdeberta-v3-base [27], [28]	✓		280
papluca/xlm-roberta-base language-detection	✓		278
asafaya/bert-base-arabic [26]		✓	111
asafaya/albert-large-arabic [26]		✓	18
facebook/xmod-base [29]	✓		852
Davlan/afro-xlmr-base [30]	✓		278
CAMeL-Lab/bert-base-arabic camelbert-msa [31]		✓	110
aubmindlab/bert-large arabertv02 [25]		✓	371
xlm-roberta-base [24]	✓		278
aubmindlab/bert-base arabertv02-twitter [25]	✓		135
MoritzLaurer/mDeBERTa-v3 base-mnli-xnli [32]	✓		279
ChaimaaBouafoud/arabicSent ChamaBert	✓		135
Osaleh/sagemaker-bert-base arabic-ar-SAS		✓	111
bert-base-multilingual-cased [33]	✓		179
bert-base-multilingual uncased [33]	✓		168
distilbert-base-multilingual cased [34]	✓		135
sentence transformers/paraphrase multilingual-MiniLM-L12-v2 [35]	✓		118
CAMeL-Lab/bert-base-arabic camelbert-da [31]		✓	109
sentence transformers/paraphrase xlm-r-multilingual-v1 [35]	✓		278
Ammar-alhaj-ali/arabic MARBERT-dialect identification-city		✓	163
CAMeL-Lab/bert-base-arabic camelbert-da-sentiment [31]		✓	109
sentence transformers/distiluse-base multilingual-cased [35]	✓		135

The experimental pipeline adhered to the following training configuration, unless otherwise stated:

- Epochs: 20
- Batch size: 16
- Optimizer: AdamW [36]
- Learning rate: 1e-5
- Train-test split: 80% training, 20% testing (as per [16], [17])
- Evaluation metric: Pearson's correlation coefficient (primary), supplemented by RMSE for some baseline comparisons.

The best model checkpoint (based on validation correlation) was saved for final testing.

4.2 Models Compared

To ensure comprehensive benchmarking, the study evaluated a set of both Arabic-specific and multilingual transformer models:

- **Multilingual models:**
  - XLM-RoBERTa-large [24]
  - mDeBERTa-v3-base [27], [28]
  - XLM-RoBERTa-base [24]
  - DistilBERT-multilingual-cased [34]
- **Arabic-specific models:**
  - AraBERTv2-large [25]
  - asafaya/bert-large-arabic [26]
  - asafaya/albert-large-arabic [26]
  - CAMeL-Lab/bert-base-arabic-camelbert-msa [31]

Models were chosen according to the criteria outlined in Section 3.3. The full fine-tuning strategy described earlier was applied uniformly to all, enabling a fair comparison of their capabilities for ASAG.

4.3 Results on AR-ASAG Dataset

Table 5 summarizes the top five performing models after fine-tuning on the AR-ASAG dataset.

Table 5: The results of the fine-tuning of the AR-ASAG dataset.

Model Name	Pearson's Correlation
xlm-roberta-large	0.85025596
aubmindlab/bert-large-arabertv2	0.833962274
asafaya/bert-large-arabic	0.828966321
asafaya/albert-large-arabic	0.824739273
microsoft/mdeberta-v3-base	0.82379811

These results indicate that XLM-RoBERTa-large outperformed all other models, confirming the advantage of large multilingual pretraining even for Arabic-specific tasks. Notably, AraBERTv2-large came second, demonstrating that Arabic-focused pretraining remains competitive.

4.4 State-of-the-Art Comparison on AR-ASAG

Table 6 compares our best-performing model with a well-cited baseline method from the literature, the COALS algorithm [16].

Table 6: Comparison of the proposed approach with the state of the art on the AR-ASAG dataset.

Models	Pearson's Correlation
Proposed approach	0.85025596
State-of-the-art COALS Algorithm [16]	0.7037

The proposed method shows a substantial improvement ($\Delta \approx +0.1466$), highlighting the benefits of transformer-based contextual embeddings over older similarity-based approaches.

4.5 Results on Philosophy Dataset

To assess the generality of our approach, the same pipeline was applied to the Philosophy dataset [17]. The results, summarized in Table 7, demonstrate that our approach maintained exceptional performance across domains.

The improvement was even more pronounced ($\Delta \approx +0.1037$), suggesting that the proposed method is well-suited for tasks requiring deep semantic understanding rather than mere lexical similarity.

Table 7: Comparison of the proposed approach with the state of the art on the Philosophy dataset.

Model	Pearson's Correlation
Proposed approach	0.965737707
State-of-the-art models: SMOreg with CombineBest [17]	0.862

4.6 Results Insights:

The performance patterns across datasets reveal several insights:

1- **Multilingual vs. Arabic-specific models:** While large multilingual transformers like XLM-RoBERTa-large achieved the highest overall scores, Arabic-specific models such as AraBERTv2-large were close contenders, particularly on the AR-ASAG dataset. This suggests that when computational resources are constrained, Arabic-specific models offer an excellent balance between performance and efficiency.

2- **Impact of fine-tuning:** The improvements over baselines highlight the necessity of full-parameter fine-tuning for ASAG tasks. Merely adding a classifier or regression head without deeper adaptation is unlikely to capture the nuanced semantic mappings needed for high-stakes grading.

3- **Dataset characteristics**: The AR-ASAG dataset contains shorter, more fact-oriented answers, while the Philosophy dataset includes longer, more interpretive responses. The fact that the proposed approach excels on both suggests strong adaptability to different question types and semantic demands.

4- **Computational trade-offs**: Although XLM-RoBERTa-large delivers the best accuracy, it requires more training time and GPU memory. Lighter architectures like ALBERT or mDeBERTa can be suitable alternatives in production environments where hardware constraints are significant.

5. DISCUSSION

The experimental results presented in Section 4 demonstrate the clear advantage of transformer-based architectures for Arabic Short Answer Grading (ASAG) when combined with a full fine-tuning strategy. Across both the AR-ASAG [16] and Philosophy [17] datasets, the proposed approach consistently outperformed existing state-of-the-art methods, with XLM-RoBERTa-large achieving the highest correlation scores in all scenarios.

One of the most striking findings is the strong performance of multilingual models particularly XLM-RoBERTa-large on an Arabic-only task. This suggests that large-scale multilingual pretraining on diverse corpora equips models with robust cross-lingual semantic representations that can be effectively adapted to specific languages, even those with rich morphology like Arabic. At the same time, the near-competitive results achieved by Arabic-specific models, such as AraBERTv2-large [25], indicate that domain-focused pretraining still holds significant value, especially when computational resources are constrained.

The improvements over baselines like COALS [16] and SMOreg with CombineBest [17] can be attributed to several factors:

1- **Contextual embeddings**:

Unlike traditional similarity-based or feature-engineered methods, transformers can model long-range dependencies and subtle meaning shifts between model and student answers.

2- **Full-parameter fine-tuning**:

Adjusting all transformer layers enables the model to specialize its attention mechanisms and hidden representations for ASAG-specific requirements, capturing the syntactic flexibility and semantic richness of Arabic.

3- **Dual-input encoding:**

Feeding both the model and student answer into the network allows for joint representation learning, which is crucial for accurately assessing content similarity when surface forms differ significantly.

From a practical perspective, these results have important implications for the deployment of ASAG systems in educational settings:

- **Speed and scalability:** Once fine-tuned, transformer models can grade answers in milliseconds, enabling real-time feedback in large-scale online learning platforms [4].
- **Fairness and consistency**: Automated grading reduces inter-grader variability, mitigating human biases and fatigue-related errors [3].
- **Adaptability**: The ability of multilingual models to perform well suggests that a single fine-tuned system could be adapted to other low-resource languages with minimal retraining.

However, the results also highlight important trade offs between performance and computational demands that cannot be overlooked. While XLM-RoBERTa-large consistently delivers superior accuracy and establishes new benchmarks in Arabic short answer grading, this gain comes with a substantial cost in terms of resource consumption. The model requires significantly more GPU memory, longer training cycles, and greater energy usage compared to smaller architectures such as ALBERT or mDeBERTa. These requirements may not be feasible for institutions or organizations operating with limited hardware, particularly in developing regions where advanced GPUs are not widely available. In such cases, lightweight yet competitive models become more attractive, as they can be fine-tuned and deployed on modest infrastructure without major compromises in grading accuracy. Thus, in practical terms, the sustainability of deployment should be considered alongside raw performance metrics, especially when scaling to thousands of users or integrating into large e-learning platforms.

Beyond resource considerations, the strong cross-domain performance; most notably on the Philosophy dataset; further emphasizes the generalization capability of the proposed approach. Unlike datasets with short, factual responses, Philosophy questions demand interpretive reasoning and abstract understanding, making them an effective stress test for semantic modeling. The model's ability to excel under these conditions indicates that it is not limited to rote lexical overlap but is capable of capturing deeper meaning. This adaptability opens the door to extending ASAG beyond traditional academic testing contexts. Potential applications include professional

certification exams that assess applied knowledge, language proficiency assessments that demand nuanced comprehension and expression, and even automated evaluation of open-ended survey responses where semantic accuracy matters more than exact wording. In our view, this broader potential underlines the practical value of transformer-based ASAG systems, suggesting that with careful calibration and resource-aware deployment, they can transform assessment practices across diverse educational and professional domains.

6. CONCLUSION AND FUTURE WORK

This study presented a comprehensive evaluation of transformer-based architectures for Arabic Short Answer Grading (ASAG), with a particular focus on the effectiveness of full fine-tuning strategies. Using two benchmark datasets: AR-ASAG [16] and Philosophy [17], we compared a range of Arabic-specific and multilingual pre-trained language models. The experimental results revealed that XLM-RoBERTa-large consistently achieved the highest performance, outperforming existing state-of-the-art methods such as COALS [16] and SMOreg with CombineBest [17] by a considerable margin.

The findings highlight several key contributions to the field:

1- Demonstrating that multilingual models, when fine-tuned, can surpass Arabic-specific architectures for ASAG tasks, even in morphologically complex languages.

2- Confirming that full-parameter fine-tuning significantly boosts grading accuracy by enabling deeper task-specific adaptation.

3- Showing that transformer-based ASAG systems are robust across domains, handling both fact-oriented and abstract, interpretive responses with high reliability.

From a practical standpoint, the proposed framework offers scalability, speed, and grading consistency, making it highly suitable for large-scale educational platforms and online learning environments. The strong results across datasets also suggest promising cross-domain generalization, paving the way for applications beyond traditional classroom assessments, such as professional certification, language proficiency testing, and automated survey analysis.

While the results are promising, there are limitations to address in future work. Large models like XLM-RoBERTa-large, although accurate, require significant computational resources for both training and inference, which may limit deployment in low-resource settings.

From our perspective, the importance of this work extends beyond the numerical improvements reported in the experiments. We believe it demonstrates that Arabic short answer grading can now move from being an experimental research topic to a practical and impactful educational tool. In our view, the results confirm the maturity of transformer-based models in handling the complexity of Arabic, while also revealing that lightweight models still hold great promise for scalable deployment. We see this research as a step toward fairer, faster, and more consistent assessment practices, with the potential to reshape how educators and institutions in Arabic-speaking regions approach evaluation and feedback.

Future research could explore a wide range of strategies aimed at making Arabic Short Answer Grading (ASAG) systems both more accurate and more efficient. One promising direction is the use of knowledge distillation, where the capabilities of large and computationally expensive models such as XLM-RoBERTa-large are transferred into smaller student models that are easier to deploy in real-world educational environments. Alongside distillation, parameter-efficient fine-tuning techniques, including methods such as LoRA, adapters, and prefix-tuning—offer practical ways to adapt powerful pre-trained models to ASAG tasks without incurring the full cost of retraining or storing billions of parameters. These approaches not only reduce computational overhead but also make it feasible to update or customize models for specific institutions, domains, or curricula with minimal resources.

In addition to efficiency, future work should also broaden the scope of ASAG to more complex and diverse assessment scenarios. One direction is multi-turn question-answer contexts, where learners engage in a sequence of responses rather than a single isolated answer. Handling such interactive exchanges would allow automated systems to better reflect real classroom dynamics and provide more nuanced evaluation. Another frontier is the integration of multimodal inputs, combining textual answers with supplementary materials such as diagrams, figures, or even spoken responses. This would be particularly valuable for subjects like science, engineering, or language learning, where

understanding is often expressed through multiple forms of representation.

Finally, inclusivity remains a crucial area for future exploration. Current ASAG research has primarily focused on Modern Standard Arabic, yet the everyday reality of learners involves a wide range of dialects and code-switching between Arabic and other languages such as English or French. Developing systems that can grade across dialects and low-resource varieties of Arabic will not only enhance robustness but also ensure fairness for students from different cultural and linguistic backgrounds. By addressing efficiency, complexity, and inclusivity in tandem, we believe that future ASAG systems can evolve into comprehensive, adaptable, and equitable tools that serve a much broader spectrum of educational needs.

Funding: This work was funded by the Deanship of Scientific Research, Islamic University of Madinah, Saudi Arabia.

REFERENCE:

[1] Page, E. B. (1966). The imminence of... grading essays by computer. Phi Delta Kappan, 47(5), 238–243.

[2] Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. International Journal of Artificial Intelligence in Education, 25(1), 60–117. https://doi.org/10.1007/s40593-014-0026-8

[3] Haley, D., Thomas, P., De Roeck, A., & Petre, M. (2007). Measuring improvement in latent semantic analysis-based marking systems: Using a computer to mark questions about HTML. Proceedings of the 9th International Australasian Computing Education Conference. Ballarat, Australia. Retrieved August 17, 2024, from http://portal.acm.org/citation.cfm?id=1273677

[4] L, H. (2000). Automated grading of short-answer tests. IEEE Intelligent Systems, 15(5), 22–37.

[5] Bonthu, S., Rama Sree, S., & Krishna Prasad, M. H. M. (2021). Automated short answer grading using deep learning: A survey. In A. Holzinger, P. Kieseberg, A. M. Tjoa, & E. Weippl (Eds.), Machine Learning and Knowledge Extraction (pp. 61–78). Springer. https://doi.org/10.1007/978-3-030-84060-0_5

[6] Galhardi, L. B., & Brancher, J. D. (2018). Machine learning approach for automatic short answer grading: A systematic review. In G. R. Simari et al. (Eds.), Advances in Artificial Intelligence – IBERAMIA 2018 (pp. 380–391). Springer. https://doi.org/10.1007/978-3-030-03928-8_31

[7] Saeed, M., & Gomaa, W. (2022, May). An ensemble-based model to improve the accuracy of automatic short answer grading. Proceedings of MIUCC 2022 (pp. 337–342). IEEE. https://doi.org/10.1109/MIUCC55081.2022.9781737

[8] Sawatzki, J., Schlippe, T., & Benner-Wickner, M. (2022). Deep learning techniques for automatic short answer grading: Predicting scores for English and German answers. In E. C. K. Cheng et al. (Eds.), Artificial Intelligence in Education: Emerging Technologies, Models and Applications (pp. 65–75). Springer. https://doi.org/10.1007/978-981-16-7527-0_5

[9] Saha, S., Dhamecha, T. I., Marvaniya, S., Sindhgatta, R., & Sengupta, B. (2018). Sentence level or token level features for automatic short answer grading?: Use both. In C. Penstein Rosé et al. (Eds.), Artificial Intelligence in Education (pp. 503–517). Springer. https://doi.org/10.1007/978-3-319-93843-1_37

[10] Gaddipati, S. K., Nair, D., & Plöger, P. G. (2020). Comparative evaluation of pretrained transfer learning models on automatic short answer grading. arXiv preprint, arXiv:2009.01303. Retrieved August 22, 2024, from http://arxiv.org/abs/2009.01303

[11] Süzen, N., Gorban, A. N., Levesley, J., & Mirkes, E. M. (2020). Automatic short answer grading and feedback using text mining methods. Procedia Computer Science, 169, 726–743. https://doi.org/10.1016/j.procs.2020.02.171

[12] Pribadi, F. S., Permanasari, A. E., & Adji, T. B. (2018). Short answer scoring system using automatic reference answer generation and geometric average normalized-longest common subsequence (GAN-LCS). Education and Information Technologies, 23(6), 2855–2866. https://doi.org/10.1007/s10639-018-9745-z

[13] Hassan, S., & El-Ramly, M. (2018). Automatic short answer scoring based on paragraph embeddings. International Journal of Advanced Computer Science and Applications, 9(10). https://doi.org/10.14569/IJACSA.2018.091048

[14] Gomaa, W. H., Nagib, A. E., Saeed, M. M., Algarni, A., & Nabil, E. (2023). Empowering short answer grading: Integrating transformer-based embeddings and BI-LSTM network. Big Data and Cognitive Computing, 7(3), Article 122. https://doi.org/10.3390/bdcc7030122

[15] Magooda, A., Zahran, M. A., Rashwan, M., Raafat, H., & Fayek, M. B. (n.d.). Vector based techniques for short answer grading.

[16] Ouahrani, L., & Bennouar, D. (2020). AR-ASAG: An Arabic dataset for automatic short answer grading evaluation. In N. Calzolari et al. (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference (pp. 2634–2643). European Language Resources Association. Retrieved August 17, 2024, from https://aclanthology.org/2020.lrec-1.321

[17] Hassan Gomaa, W., & Fahmy, A. A. (2014). Arabic short answer scoring with effective feedback for students. International Journal of Computer Applications, 86(2), 35–41. https://doi.org/10.5120/14961-3177

[18] Nael, O., ELmanyalawy, Y., & Sharaf, N. (2022). AraScore: A deep learning-based system for Arabic short answer scoring. Array, 13, 100109. https://doi.org/10.1016/j.array.2021.100109

[19] Badry, R. M., Ali, M., Rslan, E., & Kaseb, M. R. (2023). Automatic Arabic grading system for short answer questions. IEEE Access, 11, 39457–39465. https://doi.org/10.1109/ACCESS.2023.3267407

[20] Salam, M. A., El-Fatah, M. A., & Hassan, N. F. (2022). Automatic grading for Arabic short answer questions using optimized deep learning model. PLOS ONE, 17(8), e0272269. https://doi.org/10.1371/journal.pone.0272269

[21] Gomaa, W. H., & Fahmy, A. A. (2020). Ans2vec: A scoring system for short answers. In A. E. Hassanien et al. (Eds.), The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019) (pp. 586–595). Springer. https://doi.org/10.1007/978-3-030-14118-9_59

[22] Abdeljaber, H. (2021). Automatic Arabic short answers scoring using longest common subsequence and Arabic WordNet. IEEE Access, 1–1. https://doi.org/10.1109/ACCESS.2021.3082408

[23] Gomaa, W. H., & Fahmy, A. A. (2014). Automatic scoring for answers to Arabic test questions. Computer Speech & Language, 28(4), 833–857. https://doi.org/10.1016/j.csl.2013.10.005

[24] Conneau, A., et al. (2020). Unsupervised cross-lingual representation learning at scale. arXiv preprint, arXiv:1911.02116. https://doi.org/10.48550/arXiv.1911.02116

[25] Antoun, W., Baly, F., & Hajj, H. (2021). AraBERT: Transformer-based model for Arabic language understanding. arXiv preprint, arXiv:2003.00104. https://doi.org/10.48550/arXiv.2003.00104

[26] Safaya, A., Abdullatif, M., & Yuret, D. (2020). KUISAIL at SemEval-2020 Task 12: BERT-CNN for offensive speech identification in social media. arXiv preprint, arXiv:2007.13184. https://doi.org/10.48550/arXiv.2007.13184

[27] He, P., Gao, J., & Chen, W. (2023). DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. arXiv preprint, arXiv:2111.09543. https://doi.org/10.48550/arXiv.2111.09543

[28] He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with disentangled attention. arXiv preprint, arXiv:2006.03654. https://doi.org/10.48550/arXiv.2006.03654

[29] Pfeiffer, J., et al. (2022). Lifting the curse of multilinguality by pre-training modular transformers. arXiv preprint, arXiv:2205.06266. https://doi.org/10.48550/arXiv.2205.06266

[30] Alabi, J. O., Adelani, D. I., Mosbach, M., & Klakow, D. (2022). Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. arXiv preprint, arXiv:2204.06487. https://doi.org/10.48550/arXiv.2204.06487

[31] Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., & Habash, N. (2021). The interplay of variant, size, and task type in Arabic pre-trained language models. arXiv preprint, arXiv:2103.06678. https://doi.org/10.48550/arXiv.2103.06678

[32] Laurer, M., van Atteveldt, W., Casas, A., & Welbers, K. (2024). Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI. Political Analysis, 32(1), 84–100. https://doi.org/10.1017/pan.2023.20

[33] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint,

arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805

[34] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv preprint, arXiv:1910.01108. https://doi.org/10.48550/arXiv.1910.01108

[35] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. arXiv preprint, arXiv:1908.10084. https://doi.org/10.48550/arXiv.1908.10084

[36] Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. arXiv preprint, arXiv:1711.05101. https://doi.org/10.48550/arXiv.1711.05101