In-Context Meta-Learning vs. Semantic Score-Based Similarity: A Comparative Study in Arabic Short Answer Grading

Menna Fateen
Kyushu University
menna.fateen@m.ait.kyushu-u.ac.jp

Tsunenori Mine
Kyushu University
mine@ait.kyushu-u.ac.jp

Abstract

Delegating short answer grading to automated systems enhances efficiency, giving teachers more time for vital human-centered aspects of education. Studies in automatic short answer grading (ASAG) approach the problem from instance-based or reference-based perspectives. Recent studies have favored instance-based methods, but they demand substantial data for training, which is often scarce in classroom settings. This study compares both approaches using an Arabic ASAG dataset. We employ in-context meta-learning for instance-based and semantic score-based similarity for reference-based grading. Results show both methods outperform a baseline and occasionally even surpass human raters when grading unseen answers. Notably, the semantic score-based similarity approach excels in zero-shot settings, outperforming in-context meta-learning. Our work contributes insights to Arabic ASAG and introduces a prompt category classification model, leveraging GPT3.5 to augment Arabic data for improved performance.

1 Introduction

Automatic short answer grading (ASAG) has been a prominent subject of discussion in the field of AI in education, studied for more than half a century (Page, 1966). This is not surprising given the potential ASAG systems hold for enhancing various aspects of educational systems. By automating routine grading tasks, teachers can focus more on their unique human role of being motivators of learning and nurturing students' curiosity, ultimately enriching the educational experience (Keller, 1983). The shift toward automation in grading not only enhances efficiency and eliminates human bias but also empowers educators to dedicate their time and expertise to the critical aspects of teaching that require human insight and empathy.

Over the last decade, progress in the field of ASAG has significantly accelerated, driven by advancements in deep learning techniques and the availability of large datasets. ASAG systems can be broadly categorized into two main approaches: instance-based and reference-based (Horbach and Zesch, 2019). The majority of research in ASAG has primarily focused on the instance-based approach, which involves scoring individual student answers independently. On the other hand, reference-based approaches rely on measuring the similarity between the student's response and the reference answer and assigning a score based on this similarity. Reference-based approaches not only have the potential to be more robust to variability but also have the advantage of being more interpretable and less data-hungry (Bexte et al., 2023). However, only a few studies have been conducted comparing the 2 approaches and showing that the performance of reference-based approaches compared to instance-based approaches often yields worse or comparable results (Bexte et al., 2022).

While instance-based approaches have dominated the ASAG landscape, it's important to note that most of this research has been conducted in the context of the English language. English is one of the most widely studied languages, and therefore, a substantial amount of educational content and resources are available for it. However, the need for ASAG systems in other languages, such as Arabic, is equally significant. Even though Modern Standard Arabic (MSA) could also be considered a thriving language (Simons et al., 2022), datasets for ASAG in Arabic are still scarce.

In this study, we hope to contribute to the field of ASAG by presenting a comparison of two distinct approaches to ASAG in Arabic. In our first instance-based approach, we leverage a pre-trained language model (i.e. BERT) and train it on different questions with a shared type. For each instance, we create an input structure that provides contextual information for the model. In our reference-based approach, we train a score-based semantic similar-

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 350–358
December 7, 2023 ©2023 Association for Computational Linguistics

ity model using SentenceTransformers (Reimers and Gurevych, 2019). Our results demonstrate that while both techniques perform similarly in conventional training circumstances, score-based semantic similarity has considerable potential for delivering superior results in zero-shot settings. We additionally propose a "prompt category" classification model to facilitate the selection of the most suitable scoring model for a given question. We show the effectiveness of this model in low-resource settings by augmenting the training data with synthetic examples generated by GPT-3.5. To the best of our knowledge, this is the first study to apply and compare the two distinct approaches to the problem of ASAG in Arabic. Finally, we make the code and models publicly[1] available to facilitate future research in this area.

2 Related Work

Research in ASAG can be categorized into two main paradigms, instance-based or similarity-based methods (Horbach and Zesch, 2019). Most recent ASAG research follows the instance-based paradigm, where algorithms are trained primarily using a large set of student answers to learn about the features of correct and incorrect responses. With the rise of large language models and transfer learning, most studies typically involve fine-tuning BERT such as in the work by Lun et al.. Another example is the work of Nael et al. where they fine-tune BERT and ELECTRA models on a machine-translated ASAP dataset. Condor et al. used SBERT embeddings to train a model with an instance-based approach rather than using it in a similarity-based approach. Fernandez et al. introduced a single shared scoring model for multiple questions using a specified input structure that provides contextual information for each item. Similarly, to score mathematical questions, Zhang et al. use an in-context learning approach that provides scoring examples as part of the input to a MathBERT model to promote generalization.

On the other hand, in the reference-based approach, student answers are evaluated by comparing them to one or more target answers. Judgments of correctness are thus determined based on their similarity to a reference solution. In early work, reference-based approaches mainly employed feature-engineering methods such as utilizing string-based or corpus-based similarity methods (Gomaa and Fahmy, 2014) and n-grams (Shehab et al., 2018). More recently, Meccawy et al. conducted a comparative study evaluating the efficiency of different word embedding approaches for conducting feature vectors. In their study, Wang et al. introduced innovative metrics for score-based similarity to construct a text representation space that is optimized for both inter and intra-level distinctions, leading to improved scoring efficiency. In our reference-based approach, we define score-based similarity in a manner similar to what they have presented in their research.

3 Dataset Description

In this study, we utilize the AR-ASAG dataset, which is the first publicly available dataset for automatic short-answer scoring in Arabic (Ouahrani and Bennouar, 2020). The dataset consists of 2133 short answers written by graduate students in response to 48 questions. The questions are taken from 3 different exams on cybercrime where each exam consists of 16 questions. The question prompts in the exams could be classified into 5 categories based on the type of answer they expect, namely: *define*, *explain*, *consequences*, *justify*, and *compare*.

The answers in this dataset were independently annotated by two human raters on a scale of 0 to 5 where 0 is completely incorrect and 5 is considered a perfect answer. The raters were instructed to assign a score based on the similarity of the student's answers to a reference answer given for each question. Determining the similarity between two answers not only is a subjective task but also requires a deep understanding of the topic. In cases like this, where no detailed scoring rubric is provided, the raters can find it especially difficult to determine the precise degree of similarity. This is reflected in the low inter-rate agreement of 35%. However, this is expected since the raters were also given the freedom to assign intermediate scores such as 4.5 or 3.25, etc.

In our study, we treat the scoring problem as a classification problem instead of a regression one. We discretize the scores into 6 categories, 0, 1, 2, 3, 4, and 5 by taking the rounded-down median after ceiling the scores to the nearest 0.5. This is done to increase the inter-rate agreement to 56% instead of 35%. The distribution of the scores in the dataset can be seen in Figure 1 where we can observe the

[1] Our code can be found at: `https://github.com/mennafateen/SSS-vs-InCML`

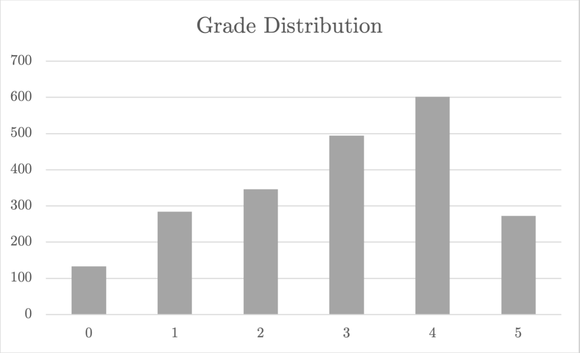

Figure 1: Distribution of the scores in the dataset.

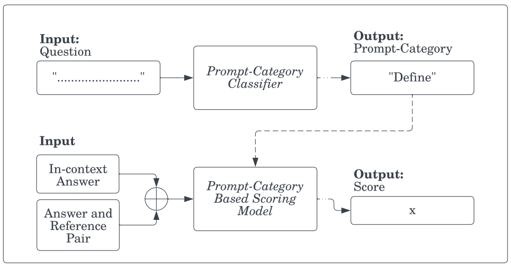

Figure 2: Overview of the in-context prompt-based scoring framework.

majority of the scores being concentrated in the range of 3 to 4.

4 Methodology

Problem Formulation

We formulate the problem of automatic short answer scoring as follows: Given a question q, a short answer a, and a reference answer a^*, the goal is to predict a score $s \in [0, 5]$ that represents the quality of the answer. Usually, each question is treated as a separate task where a separate model is trained for each task or question. However, this approach is not feasible in low-resource settings where there is a lack of annotated data. Hence, we propose a general in-context prompt-based scoring framework for automated scoring of short-answer questions where we divide the scoring problem into two sub-problems, prompt-category-based scoring and prompt category classifying.

The prompt-category-based scoring problem can be formally defined as follows: we have a set of tasks $T = \{t_i\}_{i=1}^N$, where each task t_i is defined by a question q_i, its reference answer a_i^*, and a set of instances $D_i = \{(a_{i,j}, s_{i,j})\}_{j=1}^{M_i}$, where M_i is the number of instances for the i-th task. The goal is to learn a function $f : (q, a^*, a) \rightarrow s$ that can generalize to both unseen answers and unseen questions with a small number of annotated examples. To solve the defined problem, we propose and compare two main approaches, namely, *in-context meta-learning* (InCML) and *score-based semantic similarity* (SSS). We describe each approach in detail in the following subsections. Within each approach, we train one model per prompt category, resulting in 5 models per approach.

In order to facilitate the selection of the most suitable prompt-category-based scoring model for a given question, an auxiliary prompt category classifier is trained to identify the prompt category of a given question. The output of this model should then serve as a guide for selecting the most suitable model for a given question. This is illustrated in Figure 2.

Prompt Category Classification

To construct a balanced training dataset, we train the model on the first 4 questions from each prompt category and set aside the remaining questions for testing. We utilize the SetFit framework (Tunstall et al., 2022) and use the pretrained AraBERT model (Antoun et al.) as the pretrained body. We then generate 50 pairs for contrastive learning and train the model for 5 epochs. The classification head is a logistic regression layer that takes the output of the last layer of the pretrained body as input. With this 4-shot training setup, the model achieves a mere accuracy of 0.357. To address this, we propose to augment the training data with synthetic examples generated by GPT-3.5. For each prompt category, we instruct the model to provide x more examples, for instance:

Provide five more examples that are similar to the following using the same "Define" prompt:

- عرف مصطلح الجريمة الإلكترونية (Define the term cybercrime)
- عرف مصطلح أمن المعلومات (Define the term information security)
- عرف مصطلح الهندسة الاجتماعية النفسية (Define the term psychological social engineering)
- عرف مصطلح تبييض أو غسيل الأموال (Define the term money laundering).

Table 1: Prompt category classification results on the test set with the different number of augmented examples.

	Original	Aug$_1$	Aug$_2$	Aug$_3$
Acc	0.357	0.607	0.714	**0.893**
F1	0.443	0.645	0.699	**0.821**
Precison	0.511	0.711	0.733	**0.82**
Recall	0.724	0.806	0.841	**0.90**

We experiment with different values of x and report the results in Table 1 where in Aug_1, Aug_2, and Aug_3, we augment the training data so that the number of examples per prompt category is 45, 125, and 250 respectively. As shown in the table, the model's performance significantly improves as we increase the number of augmented examples, achieving an accuracy of 0.893 with Aug_3 and an F1-score of 0.821. With this prompt classification model experiment, we show that in low-resource settings, a potential solution that could be explored is to augment available samples using generative large language models such as GPT-3.5.

Instance-based: In-Context Meta Learning Model

The in-context meta-learning model (InCML) approach draws inspiration from the work of (Fernandez et al., 2022). Building upon their foundational concepts, we apply this approach to the unique domain of cybersecurity short answer scoring in Arabic. To introduce context, we input the answers using a template that is constructed by concatenating the target answer to be scored a_j, question q_i, and its reference answer, a_i^*. We additionally include a set of K in-context examples E_i that are randomly sampled from the training set D_{train} for the i-th task or question. We build a template for each component by adding semantically meaningful task instructions as shown in Table 2. Moreover, we convert the numeric scores s_j in the in-context examples E_i to meaningful words such that: *0: ضعيف جدا (very poor), 1: ضعيف (poor), 2: متوسط (fair), 3: جيد (good), 4: جيد جدا (very good), 5: ممتاز excellent*. We then concatenate the templates to form the final input to the model. During inference time, the same templates are created for the input components where the in-context examples are fetched from the training set for seen questions only.

We train our model on the union of training datasets for all items or questions per prompt category $\cup_{i=1}^{5} D_{train}^{i}$, instead of training a separate model for each item, thus reducing the number of model parameters and required storage space. Figure 3 illustrates the in-context meta-learning model.

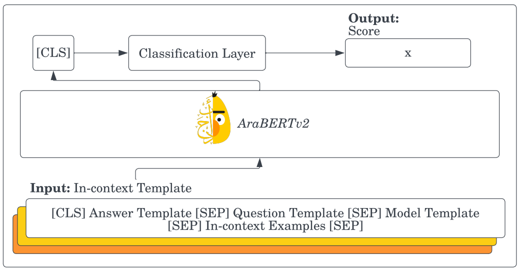
Output:
Score
[CLS]
Classification Layer
x
AraBERTv2
Input: In-context Template
[CLS] Answer Template [SEP] Question Template [SEP] Model Template [SEP] In-context Examples [SEP]

Figure 3: in-context meta-learning model

Reference-based: Score-based Semantic Similarity

When human experts are asked to score answers to open-ended questions, they usually compare the answers to a reference answer and assign a score based on the similarity between the two answers. In this approach, we propose to train a model that can mimic this process by learning to assign a score to a given answer based on its similarity to a reference answer.

To achieve this, we perform the following steps: First, we construct a simple in-context template for each answer to be graded a_j by prepending the question q_i to the answer. It has been shown that incorporating the question in the input can improve the performance of ASAG models (Lv et al., 2021). Then, we define score-based similarity as follows: Given a pair of answers a_x and a_y with their scores s_x and s_y, the similarity between the two answers is defined as:

$$sim(a_x, a_y) = \frac{s_x}{s_y}; (s_x \leq s_y) \tag{1}$$

For each task/question i, we then construct a dataset $\mathcal{D}_i$ of answer pairs annotated together via the score-based metric indicating their similarity as shown in Equation 2, where X is the number of examples per score k and a_k is a student answer that was graded k.

$$\mathcal{D}_i = \{\{(a_k, a_{\neg k}, sim(a_k, a_{\neg k}))\}_{x=0}^{X}\}_{k=0}^{5} \tag{2}$$

Table 2: Input components and templates for the in-context meta-learning model.

Input Component	Template	Sample
Student Answer	قيم هذه الإجابة: x	قيم هذه الإجابة: العلم الذي يستخدم التحليل الاحصائي لصفات الانسان الحيوية وذلك للتأكد من هويته الشخصية (Grade this answer: The science that uses statistical analysis of a person's vital characteristics to confirm his identity)
Question	السؤال: q_i	السؤال: عرف مصطلح القياس الحيوي (Question: Define the term biometrics)
Reference	النموذج: a_i^*	النموذج: هو العلم الذي يستخدم التحليل الإحصائي لصفات الإنسان الحيوية وذلك للتأكد من هويته الشخصية بإستخدام صفاته الفريدة وهي صفات سلوكية وصفات فيزيائية (Reference: It is the science that uses statistical analysis of a person's vital characteristics to confirm his identity using his unique characteristics, which are behavioral characteristics and physical characteristics.)
Grades	تقيم : ...	تقيم : ضعيف جدا ضعيف متوسط جيد جيد جدا ممتاز (Grades: very poor, poor, fair, good, very good, excellent)
Examples	مثال: $x_{\neg j}$	مثال: هو علم يدرس حالة الإنسان الفريدة التي تميز شخصًا عن آخر تقيم : ضعيف (Example: It is a science that studies the unique human condition that distinguishes one person from another Grade: Weak)

We experiment with 3 different settings of X when constructing the dataset with $X = 30$, $X = 50$, and a final configuration where we account for the distribution of the scores in the training set so that the number of samples per score is $50\frac{N_k}{\sum_{k=0}^{5} N_k}$

We then train one model on the union of training datasets for all items or questions per prompt category $\cup_{i=1}^{5} D_{train}^{i}$, instead of training a separate model for each item as described in the InCML approach. In this approach, using SBERT, we fine-tune a pretrained AraBERT model through a Siamese network structure where we train the model using a cosine-similarity loss function. For each answer pair in the union dataset, we pass both answers through the model which generates an embedding u and v for each answer. The gold similarity score is then compared with the cosine similarity between the generated embeddings. Figure 4 illustrates the score-based semantic similarity model.

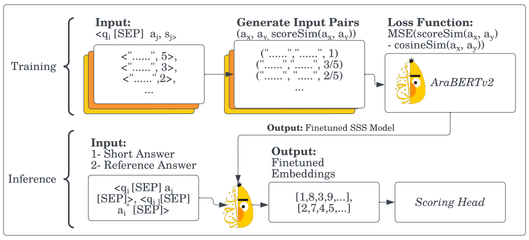

Figure 4: Score-based semantic similarity model

5 Experimental Results

Evaluation Metrics

To evaluate the performance of the proposed approaches, we use two metrics, namely, quadratic weighted kappa (QWK) and percentage of tick accuracy (PTA). QWK is a commonly used metric in ASAG that measures the agreement between two raters. In (Williamson et al., 2012), the authors suggest that the QWK between automated and human scoring should be at least 0.7 on datasets with normal distribution to be considered acceptable. Percentage of Tick Accuracy (PTA$_x$) measures the percentage of answers that are scored correctly or within x points of the gold score. PTA$_0$ would be equivalent to accuracy while PTA$_1$ also includes answers that are scored within 1 point of the gold score (e.g. 3 is considered correct if the gold score is 2 or 4) and so on.

Experimental Setup

We use AraBERT as the pretrained body for both approaches. In the InCML approach, we use the Adam optimizer with a learning rate of 1e-5 and a batch size of 8 and train the models for 6 epochs. Similarly, in SSS, we train the model for 6 epochs but use a batch size of 16 instead.

Results

We undertake two experiments to evaluate the performance of the proposed approaches. In the first experiment, we test the models' performance on unseen answers. We set aside 10% of the answers from each prompt category for testing. In the second experiment, we evaluate the performance of the models on unseen questions. We set aside the first question and its answers from each prompt category for testing. This setting is considered a zero-shot learning scenario since the models are not trained on any examples from the test set. The results of both experiments are shown in Table 3. As a baseline for comparison, we report the results of a majority class classifier and the QWK and PTA between the rounded-up grades of the two human raters.

6 Discussion

6.1 Unseen Answers

As shown in Table 3, compared to the majority class classifier, both approaches with different configurations outperform the baseline in all prompt categories in the unseen answers experiments. Comparing the model's performance to human performance, we observe that with prompts $P1$ and $P3$, InCML$_0$ and InCML$_3$ outperform the human raters in terms of QWK. In terms of PTA$_0$, InCML$_1$, InCML$_3$, and additionally SSS$_{50}$ outperform the human raters with prompt $P1$ while InCML$_1$ again outperforms the human raters with prompt $P3$ type questions. In the remaining prompt categories, the performance of both approaches in terms of QWK is marginally below the QWK achieved between the human raters with the in-context meta-learning approach showing a tendency to outperform the score-based similarity approach.

In the in-context meta-learning approach, we observe that the performance of the model does not necessarily improve as we increase the number of in-context examples. In fact, in some cases, the performance decreases. We speculate that this might be attributed to potential overfitting on the in-context examples. It is also important to note that the performance of InCML fluctuates depending on the in-context examples that are extracted from the training set which introduces inherent instability. On the other hand, in the case of the semantic score-based similarity approach, an increase in the number of examples per score generally corresponds to improved model performance.

In the unseen answers experiment, with a few training examples, we observe that while both approaches have comparable performance to the human raters, the instance-based in-context meta-learning approach generally gives better performance compared to the reference-based approach.

6.2 Unseen Questions

In Table 3, we see that the overall performance of the models in the unseen questions experiment, or in zero-shot settings, is lower than the performance of unseen answers. However, we observe that the PTA$_0$ of the models is still higher than the majority class classifier in most prompt models using our reference-based, SSS$_{30}$, SSS$_{50}$ and SSS$_{50W}$ methods.

Compared to the instance-based InCML approach, it is evident that the reference-based SSS approach proposed gives higher performance showcasing its reduced data hunger advantage and its ability to generalize to new questions.

7 Conclusion

In this paper, we propose a general in-context prompt-based scoring framework for automated scoring of short-answer questions. We divide the scoring problem into two sub-problems, namely, prompt category classification and prompt-category-based scoring. For prompt-category classification, we utilize a few-shot, prompt-free framework to train the model. We also show that with data augmentation using GPT3.5, the performance could be significantly increased. We then propose two main approaches for the prompt-category-based scoring problem, namely, instance-based in-context meta-learning and reference-based semantic similarity. Utilizing the only publicly available Arabic ASAG dataset, we evaluate both approaches in their ability to generalize to unseen answers and unseen questions. Experimental results show that both proposed approaches outperform the majority class classifier and are comparable to human raters when grading unseen answers. However, the performance is highly prompt-dependent and no particular approach is consistently better than the other. In zero-shot settings, when generalizing to unseen questions, we observe a tendency for the reference-based semantic similarity approach to outperform the instance-based in-context meta-learning approach. We thus believe that in class-

Table 3: Experimental results.

			Human	**MV**	**InCML**$_0$	**InCML**$_1$	**InCML**$_3$	**SSS**$_{30}$	**SSS**$_{50}$	**SSS**$_{50W}$
Unseen Answers	***P1***	**QWK**	*0.676*		0.611	0.656	**0.697**	0.591	0.593	0.632
		PTA$_0$	*0.357*	0.357	0.286	**0.500**	0.404	0.357	0.393	0.357
		PTA$_1$	*0.893*		0.857	0.843	0.889	0.857	0.857	**0.893**
		PTA$_2$	*0.929*		0.964	0.954	0.964	**1.000**	**1.000**	**1.000**
		PTA$_3$	*1.000*		**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	P2	**QWK**	*0.788*		0.722	0.668	0.760	0.718	**0.763**	0.718
		PTA$_0$	*0.568*	0.239	0.432	0.443	**0.451**	0.375	0.443	0.364
		PTA$_1$	*0.920*		0.909	0.875	0.936	0.932	**0.955**	0.943
		PTA$_2$	*0.977*		0.989	0.963	**0.998**	0.989	0.989	0.977
		PTA$_3$	*1.000*		**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	P3	**QWK**	*0.749*		**0.798**	0.774	0.727	0.557	0.581	0.660
		PTA$_0$	*0.385*	0.308	0.385	**0.542**	0.385	0.154	0.385	0.385
		PTA$_1$	*0.846*		**0.923**	0.869	0.919	0.846	0.846	0.885
		PTA$_2$	*0.962*		**1.000**	**1.000**	0.950	**1.000**	0.962	0.962
		PTA$_3$	*1.000*		**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	P4	**QWK**	*0.666*		0.602	**0.649**	0.519	0.614	0.613	0.515
		PTA$_0$	*0.533*	0.311	0.400	0.464	0.351	**0.467**	**0.467**	0.444
		PTA$_1$	*0.822*		0.867	0.813	0.733	**0.911**	**0.911**	0.844
		PTA$_2$	*0.978*		**1.000**	0.989	0.936	**1.000**	**1.000**	**1.000**
		PTA$_3$	*1.000*		**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	P5	**QWK**	*0.716*		**0.696**	0.000	0.070	0.529	0.606	0.405
		PTA$_0$	*0.526*	0.421	0.421	0.421	0.368	**0.474**	0.421	0.368
		PTA$_1$	*0.842*		**0.842**	0.684	0.737	0.789	**0.842**	0.789
		PTA$_2$	*0.947*		**1.000**	0.789	0.842	**1.000**	**1.000**	**1.000**
		PTA$_3$	*1.000*		**1.000**	0.947	0.947	**1.000**	**1.000**	**1.000**
Unseen Questions	***P1***	**QWK**	*0.743*		0.145	0.000	0.063	0.620	0.599	**0.627**
		PTA$_0$	*0.596*	0.383	0.213	0.000	0.064	**0.447**	0.404	**0.447**
		PTA$_1$	*0.957*		0.553	0.043	0.149	**1.000**	**1.000**	**1.000**
		PTA$_2$	*1.000*		0.936	0.085	0.404	**1.000**	**1.000**	**1.000**
		PTA$_3$	*1.000*		0.979	0.319	0.766	**1.000**	**1.000**	**1.000**
	P2	**QWK**	*0.876*		0.000	-0.055	0.026	**0.645**	0.558	**0.645**
		PTA$_0$	*0.833*	0.416	0.167	0.083	0.167	**0.500**	**0.500**	**0.500**
		PTA$_1$	*0.916*		0.167	0.167	0.167	**0.917**	0.833	**0.917**
		PTA$_2$	*1.000*		0.417	0.417	0.333	**1.000**	**1.000**	**1.000**
		PTA$_3$	*1.000*		0.833	0.417	0.833	**1.000**	**1.000**	**1.000**
	P3	**QWK**	*0.752*		0.000	-0.014	0.000	0.488	0.446	**0.495**
		PTA$_0$	*0.714*	**0.469**	0.082	0.082	0.082	0.204	0.224	0.204
		PTA$_1$	*0.918*		0.184	0.184	0.184	**0.918**	0.878	**0.918**
		PTA$_2$	*0.959*		0.347	0.347	0.347	**1.000**	**1.000**	**1.000**
		PTA$_3$	*0.980*		0.531	0.531	0.531	**1.000**	**1.000**	**1.000**
	P4	**QWK**	*0.774*		0.101	-0.015	-0.161	0.254	0.284	0.365
		PTA$_0$	*0.395*	0.271	0.208	0.125	0.042	**0.333**	**0.333**	**0.333**
		PTA$_1$	*0.875*		0.500	0.292	0.375	0.688	**0.708**	0.688
		PTA$_2$	*0.979*		0.646	0.375	0.771	0.938	0.938	**1.000**
		PTA$_3$	*1.000*		0.896	0.646	0.958	**1.000**	**1.000**	**1.000**
	P5	**QWK**	*0.358*		0.000	**0.069**	0.003	0.000	0.029	0.024
		PTA$_0$	*0.667*	**0.500**	0.000	0.146	0.000	0.042	0.042	0.021
		PTA$_1$	*0.979*		0.000	0.396	0.021	0.500	**0.563**	0.500
		PTA$_2$	*1.000*		0.000	0.563	0.188	**1.000**	**1.000**	**1.000**
		PTA$_3$	*1.000*		0.042	0.604	0.667	**1.000**	**1.000**	**1.000**

room settings, the reference-based semantic similarity approach could be a more suitable solution due to its superiority in zero-shot settings.

Limitations

In this paper, we presented a comparison between a specific instance-based and reference-based approach, thus our findings are limited to these methods and cannot be generalized to different methods. This study was also limited to a prompt-category-based scoring framework and while preliminary experiments were conducted, we did not compare with specific prompt-based models or cross-prompt-category models for a more straightforward and comprehensible comparison. Due to the scarcity of resources, our comparison also relies on a specific dataset, which does not encompass the full diversity of responses or topics encountered in a real-world educational setting. Furthermore, since we utilize an Arabic dataset, we adapted a BERT model pre-trained on Arabic data but have not presented a comparison with a language-agnostic model. Finally, while we briefly touched upon the potential of reference-based approaches in offering explainability, we have not delved into the topic of interpretability of the provided models. Understanding why a model assigns a specific score to an answer is essential for educational applications, as it can provide valuable feedback to students, however, it is beyond the scope of this paper and is left for future work.

Acknowledgements

This work was partially supported by JST, the establishment of university fellowships towards the creation of science technology innovation, Grant Number JPMJFS2132, and KAKENHI No. JP20H01728, JP21H00907, and JP23H03511.

References

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.

Marie Bexte, Andrea Horbach, and Torsten Zesch. 2022. Similarity-based content scoring - how to make sbert keep up with bert. In *Proceedings of The 17th Workshop on Innovative Use of NLP for Building Educational Applications*.

Marie Bexte, Andrea Horbach, and Torsten Zesch. 2023. Similarity-based content scoring-a more classroom-suitable alternative to instance-based scoring? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1892–1903.

Aubrey Condor, Max Litster, and Zachary Pardos. 2021. Automatic short answer grading with sbert on out-of-sample questions. *International Educational Data Mining Society*.

Nigel Fernandez, Aritra Ghosh, Naiming Liu, Zichao Wang, Benoît Choffin, Richard Baraniuk, and Andrew Lan. 2022. Automated scoring for reading comprehension via in-context bert tuning. In *International Conference on Artificial Intelligence in Education*, pages 691–697. Springer.

Wael Hassan Gomaa and Aly Aly Fahmy. 2014. Automatic scoring for answers to arabic test questions. *Computer Speech & Language*, 28(4):833–857.

Andrea Horbach and Torsten Zesch. 2019. The influence of variance in learner answers on automatic content scoring. In *Frontiers in Education*, volume 4, page 28. Frontiers Media SA.

John M Keller. 1983. Motivational design of instruction. *Instructional design theories and models: An overview of their current status*, 1(1983):383–434.

Jiaqi Lun, Jia Zhu, Yong Tang, and Min Yang. 2020. Multiple data augmentation strategies for improving performance on automatic short answer scoring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13389–13396.

Gaoyan Lv, Wei Song, Miaomiao Cheng, and Lizhen Liu. 2021. Exploring the effectiveness of question for neural short answer scoring system. In *2021 IEEE 11th International Conference on Electronics Information and Emergency Communication (ICEIEC) 2021 IEEE 11th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, pages 1–4. IEEE.

Maram Meccawy, Afnan Ali Bayazed, Bashayer Al-Abdullah, and Hind Algamdi. 2023. Automatic essay scoring for arabic short answer questions using text mining techniques. *International Journal of Advanced Computer Science and Applications*, 14(6).

Omar Nael, Youssef ELmanyalawy, and Nada Sharaf. 2022. Arascore: a deep learning-based system for arabic short answer scoring. *Array*, 13:100109.

Leila Ouahrani and Djamal Bennouar. 2020. Ar-asag an arabic dataset for automatic short answer grading evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2634–2643.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Abdulaziz Shehab, Mahmoud Faroun, and Magdi Rashad. 2018. An automatic arabic essay grading system based on text similarity algorithms. *International Journal of Advanced Computer Science and Applications*, 9(3).

Gary F Simons, Abbey L Thomas, and Chad K White. 2022. Assessing digital language support on a global scale. *arXiv preprint arXiv:2209.13515*.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*.

Bo Wang, Billy Dawton, Tsunenori Ishioka, and Tsunenori Mine. 2023. Optimizing answer representation using metric learning for efficient short answer scoring. *The 20th Pacific Rim International Conference on Artificial Intelligence (PRICAI)*.

David M Williamson, Xiaoming Xi, and F Jay Breyer. 2012. A framework for evaluation and use of automated scoring. *Educational measurement: issues and practice*, 31(1):2–13.

Mengxue Zhang, Sami Baral, Neil Heffernan, and Andrew Lan. 2022. Automatic short math answer grading via in-context meta-learning. *arXiv preprint arXiv:2205.15219*.