Corruption-Based Data Augmentation for Arabic Essay Scoring: A Preliminary Study on the Organization Trait

May Bashendy and Tamer Elsayed
Computer Science and Engineering Department, Qatar University
{ma1403845, telsayed}@qu.edu.qa

Abstract

Despite significant advances in Automated Essay Scoring (AES), progress in Arabic AES remains limited by the scarcity and imbalance of publicly available datasets. Manual curation of such data is labor-intensive and lacks scalability. To address this, we introduce *COrE*, a *corruption-based* data augmentation method that targets the organization trait of Arabic essays. *COrE* generates synthetic essays by intentionally disrupting the organization of well-written essays through controlled, distance-aware sentence swapping. Our experiments are conducted on *TAQAE*, a dataset of 620 essays across 4 distinct writing prompts. We evaluate the effectiveness of *COrE* using two widely-adopted pre-trained models: AraBERTv2 and CAMeLBERT-mix. Both models show improved performance with *COrE*, achieving gains of 9-17% over the no-augmentation baseline. These results highlight the potential of trait-specific augmentation to address data scarcity and enhance AES performance for low-resource languages.

Keywords: Data Augmentation, Essay Scoring, Arabic Language

1. Introduction

Automated Essay Scoring (AES) is gaining importance in modern education assessment for its potential to provide consistent, scalable, and cost-effective evaluation of student writing, supporting both large-scale testing and formative feedback. However, developing effective AES systems for low-resource languages like Arabic remains challenging due to the language's rich morphology, lexical ambiguity, and complex morpho-syntactic structures. More importantly, the scarcity of large, balanced, and high-quality annotated Arabic essay datasets has limited the development of robust Arabic AES models and hindered fair evaluation and comparison across systems.

To mitigate data scarcity, synthetic data generation has emerged as a promising research direction. Unlike manual data curation, which is labor-intensive, time-consuming, and often limited by human resources, synthetic methods can produce varied data at scale in a cost-effective manner. Several studies have investigated data augmentation techniques as a form of synthetic data generation, such as paraphrasing, noise injection, and Large Language Models (LLMs) to expand training resources and enhance model robustness (Li et al., 2022; Ding et al., 2024; Liu et al., 2024; ElSabagh et al., 2025; Gupta, 2023).

In this paper, we introduce *COrE*, a **C**orruption-based data augmentation approach targeting the **Or**ganization trait[1] of Arabic **E**ssays. *COrE* intentionally disrupts the structural coherence and logical flow of high-quality essays by **randomly swapping sentences**, generating synthetic versions that simulate varying levels of organization quality. The method controls corruption intensity by considering both the number and distance of performed swaps. Our preliminary study is guided by the following research questions in the context of *prompt-specific* AES setup (i.e., training and testing on different essays written for the *same* prompt[2]): **(RQ1)** Can a corruption-based augmentation strategy (*COrE*) enhance Arabic AES model performance for scoring the organization trait in a very low-constrained setup in terms of training? **(RQ2)** How does *COrE* perform compared to baseline approaches?

We conduct our experiments on *TAQAE* (Sayed et al., 2025), a newly-formed dataset of 620 essays across 4 writing prompts. The results on two widely-adopted Arabic models, AraBERTv2 and CAMeLBERT-mix, show that *COrE* enhances performance, achieving 9-17% improvements over the no-augmentation baseline.

Our contribution in this work is three-fold:

- We introduce *COrE*, a novel, trait-specific corruption-based data augmentation method for Arabic AES.
- We evaluate *COrE*'s effectiveness across two widely-adopted pre-trained Arabic models.
- We publicly release the generated synthetic data to facilitate future research in Arabic AES for the organization trait in particular.[3]

The remainder of this paper is organized as follows. Section 2 reviews related work. Sec-

[1] A trait is an aspect of student writing.

[2] A prompt is the text describing an essay writing task.

[3] https://drive.google.com/drive/folders/1dzsXBTLgcLOtMW5usqCbOfRRGqiQHbVR

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 10525–10531
11–16 May 2026.

tion 3 presents our corruption-based augmentation method, *COrE*. Section 4 describes the experimental design, while Section 5 details the experimental setup. Section 6 presents and discusses the results. Finally, Section 7 concludes the paper with future research directions.

2. Related Work

Research on AES has advanced in recent years, yet progress in Arabic remains limited compared to high-resource languages such as English. Benchmark datasets like ASAP/ASAP++ (Mathias and Bhattacharyya, 2018) and PERSUADE (Crossley et al., 2024) have driven major advances in English AES. In contrast, Arabic AES continues to suffer from a lack of large and publicly available annotated datasets, which constrains both methodological development and fair benchmarking.

Arabic AES Datasets Despite challenges, several Arabic AES datasets have been introduced, yet all remain limited in size, annotation granularity, or accessibility. Among them, the Zayed Arabic English Bilingual Undergraduate Corpus (ZAEBUC) (Habash and Palfreyman, 2022) provides CEFR annotations for 214 essays. The Qatari Corpus of Argumentative Writing (QCAW) (Ahmed et al., 2024) includes essays with POS tagging and holistic scores. Building on it, QAES (Bashendy et al., 2024) provides their holistic and trait-level scores. TAQEEM 2025 shared task dataset (Bashendy et al., 2025) includes 1,265 annotated essays across 4 prompts, and TAQAE (Sayed et al., 2025) combines QAES with 2 prompts from TAQEEM. Other proprietary resources such as Abbir (Alghamdi et al., 2014) and AAEE (Azmi et al., 2019) include essays annotated for only holistic scores. Overall, most Arabic AES datasets, summarized in Table 1, are small, non-standardized, and lack trait-level annotations, making it difficult to train robust AES models. This scarcity highlights the need for data augmentation strategies that can enrich training data and support trait-specific modeling.

Data Augmentation for AES Data augmentation has emerged as a promising solution not only to mitigate data scarcity in AES but also to improve model generalization by automatically generating additional training samples, without the costly and time-consuming process of manual curation and annotation. In English AES, Gupta (2023) utilized transformer-based models for essay augmentation, while Yoo et al. (2025) generated 20K synthetic essays through a corruption-based strategy (CASE), both demonstrating the effectiveness of augmentation in improving scoring performance. In Arabic AES, the only known study employing data augmentation is that of Qwaider et al. (2025), who utilized LLM prompting combined with error injection to generate synthetic Arabic essays annotated with holistic CEFR-based scores.

Dataset	Essays	Len	Tasks	Public	HOL	Traits
ZAEBUC	214	156	3	✓	✓	×
QCAW	195	499	2	×	✓	×
Abbir	640	150	2	×	✓	×
AAEE	350	-	8	×	✓	×
QAES	195	489	2	✓	✓	✓
TAQEEM	1,265	151	4	✓	✓	✓

Table 1: Existing Arabic AES datasets. 'Len' denotes average essay length in words, and 'HOL' refers to holistic scoring.

In contrast to prior studies, our work focuses on trait-specific augmentation, particularly targeting the organization trait. Our proposed *COrE* methods builds on CASE framework (Yoo et al., 2025) by introducing a *distance-aware*, controlled swapping mechanism that more realistically models variations in essay organization quality.

3. Corruption-Based Augmentation

Inspired by the Corruption-based Augmentation Strategy for Essays (CASE) proposed by Yoo et al. (2025), we introduce an enhanced corruption-based augmentation method, *COrE*, specifically designed to target the *organization* trait in essays. *COrE* generates synthetic lower-scoring essays by deliberately disrupting the sentence order of high-scoring well-written original essays (organization score ≥ 4), thereby systematically degrading structural coherence while maintaining the original essay's linguistic content. This targeted approach is guided by a *deterministic scoring function* that controls the degree of corruption based on two key factors: the number of swaps performed and the distance between swapped sentences as follows:

$$S_c = S_o - \left((S_o - 1) \cdot \frac{Swaps}{MaxSwaps} \cdot \frac{\sum_{i=1}^{Swaps} d_i}{MaxDist} \right)$$

Here, S_c denotes the trait score of the synthetic essay after *corruption*, S_o is the *original* score, $Swaps$ refers to the number of *sentence swaps* performed, and $MaxSwaps$ is the maximum allowable swaps that can be performed in an essay, calculated as $\lfloor \frac{N}{2} \rfloor$, where N is the total number of sentences in an essay. Notably, our implementation constraint that each sentence participates in at most one swap, preventing redundant swaps that could negate intended disruptions.

For each swap i, we define d_i as the absolute distance between the original positions of the two

swapped sentences. The cumulative distance of swaps $\sum_{i=1}^{Swaps} d_i$ is then normalized by $MaxDist$. The value of $MaxDist$ depends on the total number of sentences N in the essay, and represents the maximum possible cumulative distance under the one-swap-per-sentence constraint. It reflects the theoretical upper bound of structural disruption, obtained by performing the most disruptive swaps (e.g., first $\leftrightarrow$ last sentences). Formally, $MaxDist$ is calculated as: $\sum_{i=0}^{\lfloor N/2 \rfloor - 1}(N - 1 - 2i)$, where for each swap i, the sentence at position i is paired with the maximally distant sentence at position $(N - 1 - i)$, yielding a distance of $(N - 1 - 2i)$.

Overall, *COrE* ensures that score degradation increases progressively with both the number of swaps and the cumulative distance between swapped sentences, reflecting how greater structural disruptions more severely impact organizational coherence. To maintain realism and precision, we integrate a tolerance margin t into the iterative swap process, halting when the achieved score falls within this margin of the target corruption score, while adhering to all defined constraints. This adaptive mechanism allows *COrE* to simulate a spectrum of organizational flaws, ranging from minor reordering to severe incoherence, mirroring real-world essay variations more effectively.

Compared to CASE, our augmentation method *COrE* introduces key improvements. First, while CASE simulates degradation, it does not account for the severity of disruption, which *COrE* explicitly models through distance-aware sentence swapping. Second, CASE allows repeated swaps of the same sentences, potentially undoing prior disruptions (e.g., swapping the same pair twice restores the original order). By addressing these limitations, *COrE* offers a more robust and targeted approach for generating synthetic Arabic essays with controlled organizational degradation.

4. Experimental Design

This section describes the experimental design for evaluating whether synthetic augmentation can improve AES performance in low-resource settings. We simulate these conditions by creating a small, balanced seed set of essays, which we then augment using three distinct approaches: naïve oversampling, corruption-based augmentation (*COrE*), and a combination of both. Details on the dataset, seed selection, and augmentation methods are provided below.

4.1. Dataset

In this study, we use *TAQAE*, a newly introduced dataset by Sayed et al. (2025), including 620 Arabic essays over 4 prompts drawn from two sources.

Source	Prompt	Type	Essays	Len
TAQEEM	P1	EXP	215	137
TAQEEM	P2	PER	210	150
QAES	P3	PER	115	500
QAES	p4	PER	80	473

Table 2: *TAQAE* dataset statistics. "EXP" and "PER" denote explanatory and persuasive prompts; "Len" is the average word count.

The first source includes 425 essays (corresponding to P1 and P2) provided by TAQEEM 2025 Shared Task (Bashendy et al., 2025). The second source is the Qatari Corpus of Argumentative Writing (QCAW) (Ahmed et al., 2024)[4], which provides 195 essays (corresponding to P3 and P4), leveraging their QAES publicly available annotations (Bashendy et al., 2024).[5] The essays from both sources are annotated across seven traits, among them is the organization trait ranging from 1 to 5. Table 2 provides a breakdown of the prompts featured in the *TAQAE* dataset.

4.2. Seed Data Selection

To simulate a realistic low-resource training setting, we intentionally limited the number of essays per score level. For each prompt, we selected up to 5 essays per score level for the organization trait, resulting in a maximum of 25 essays per prompt (5 essays × 5 score levels), which we designate as our **seed training data**. Due to class imbalance in *TAQAE*, some score levels remained underrepresented where fewer than 5 essays were available.

4.3. Synthetic Data Generation

To foster robust AES systems for Arabic and other low-resource languages, we enrich the seed training data with targeted synthetic samples using 3 approaches: naïve oversampling, corruption-based augmentation (*COrE*), and corruption-based augmentation (*COrE*) with oversampling.

Naïve Oversampling: This method involves duplicating existing original essays without introducing any modifications. While simple, it effectively increases the volume of training data. To maintain balance, we ensured an even distribution of duplicated essays across all score levels.

Corruption-Based Augmentation (*COrE*): This method utilizes the corruption strategy outlined in Section 3 to deliberately degrade essays in

[4] https://catalog.ldc.upenn.edu/LDC2022T04
[5] https://gitlab.com/bigirqu/qaes

a controlled fashion. Specifically, for each well-organized essay with an original organization score S_o (where $S_o = 4$ or 5), k corrupted variants were generated for each **lower** target score level. Here, k is a tunable hyperparameter. This corruption process simulates varying levels of organizational disruption, enabling us to generate training data that reflects a broad spectrum of structural quality while preserving topical relevance.

Corruption-Based Augmentation (*COrE*) + Oversampling: While *COrE* expands the training set by generating lower-scored variants from high-scoring essays, it results in an imbalanced score distribution. This is due to its unidirectional nature: essays of score 5 can only serve as sources, not targets, and essays of score 4 can only be generated from score 5 input essays. In contrast, essays with lower scores (1–3) can be generated from both score 4 and score 5 essays, leading to their over-representation. To restore balance, we applied post-augmentation oversampling by duplicating essays from underrepresented score levels until all levels matched the maximum count. This ensured a uniform label distribution for fair and consistent model training.

5. Experimental Setup

This section outlines the experimental configurations used to extrinsically[6] evaluate *COrE*, including the models employed, dataset splits, hyperparameter settings, and evaluation metric.

Models To evaluate the effectiveness of our augmentation strategy (*COrE*), we fine-tuned two widely adopted pre-trained Arabic language models with regression heads: **AraBERTv2**[7] and **CAMeLBERT-mix**.[8] Both were previously used in Arabic AES by Ghazawi and Simpson (2024, 2025) and Qwaider et al. (2025), respectively. Both are trained on large Arabic corpora covering Modern Standard Arabic (MSA), with CAMeLBERT-mix also including dialectal data. Their Arabic-specific pretraining allows them to capture the linguistic features relevant for essay scoring. Both models adopt the BERT-base architecture and are similar in size, ensuring fair comparison.

Evaluation To assess model performance, we employ Quadratic Weighted Kappa (QWK) (Cohen, 1968), a widely adopted metric in AES that quantifies agreement between human annotators and system predictions.

Appraoch	Model	P1	P2	P3	P4
Without Augment.	AraBERT	25	22	21	19
	CAMeLB	25	22	21	19
CASE Corruption	AraBERT	$985_{(30)}$	$252_{(10)}$	$211_{(10)}$	$709_{(30)}$
	CAMeLB	$57_{(1)}$	$45_{(1)}$	$401_{(20)}$	$134_{(5)}$
Naïve Oversamp.	AraBERT	$50_{(1)}$	$352_{(15)}$	$441_{(20)}$	$589_{(30)}$
	CAMeLB	$150_{(5)}$	$132_{(5)}$	$546_{(25)}$	$589_{(30)}$
COrE Corruption	AraBERT	$470_{(15)}$	$43_{(1)}$	$496_{(25)}$	$459_{(20)}$
	CAMeLB	$55_{(1)}$	$43_{(1)}$	$211_{(10)}$	$129_{(5)}$
COrE + Oversamp.	AraBERT	$517_{(10)}$	$186_{(5)}$	$621_{(25)}$	$719_{(25)}$
	CAMeLB	$73_{(1)}$	$54_{(1)}$	$261_{(10)}$	$859_{(30)}$

Table 3: Training size per model and prompt using the optimal k value for each augmentation strategy. Numbers in subscript indicate the optimal k.

Data Splits The training set size differs based on whether naïve oversampling, *COrE*, or a combination of both is applied. In all cases, the training set begins with the seed data defined earlier, up to 25 essays per prompt, and is then expanded by adding augmented samples specific to each augmentation strategy. Table 3 summarizes the "best" training set sizes (after tuning the augmentation factor k) across experiments. However, to ensure fair comparison, the development and test sets remain consistent in all experiments. The development set is fixed at 50 essays per prompt (chosen with stratified sampling) to meet the minimum requirement of QWK, which recommends at least $2l^2$ samples for reliable estimation, where l represents the number of score levels (Cicchetti, 1981). The test set includes all remaining essays after reserving seed training and development sets: 139, 137, 44, and 11 for prompts P1-P4, respectively.

Hyper-parameters We fine-tuned all encoder layers using a learning rate of 2×10^{-5}, batch size of 16, and a total of 100 training epochs. In the synthetic data generation phase, we explored various augmentation factors $k \in [1, 5, 10, 15, 20, 25, 30]$, where k determines the number of corrupted variants generated for each original essay, thereby influencing the size of the augmented dataset. We tuned k by evaluating the performance on the development split and using the best-performing value for testing. Also, a tolerance margin of $t = \pm 0.3$ was used during the swapping process, tuned empirically after testing 0.2 and 0.4.

[6] Extrinsic evaluation measures synthetic data quality by its impact on AES system performance.

[7] AraBERTv2

[8] CAMeLBERT-mix

6. Results and Discussion

To evaluate the impact of *COrE*, we conducted experiments on the 4 prompts of *TAQAE* using the *prompt-specific* setup. To account for randomness introduced by sentence-swapping corruption, each augmentation approach was repeated *3 times*, and the average QWK was reported. For each model and augmentation approach, the value of k was tuned on the development set, and the best configuration was evaluated on the test set. The results reported in Table 4 show the performance[9] on the test set under these best-k settings. As shown in Table 4, *COrE* improves AES performance, though the gains vary by model and prompt.

For AraBERTv2, *COrE* achieved an average QWK of 0.367, marking a 9% improvement over the no-augmentation baseline. It also outperformed the CASE method, indicating that including a distance-aware variable and limiting the number of swaps per sentence effectively enhanced performance. Notably, the naïve oversampling baseline alone performed poorly (average 0.282, a 16% drop from no-augmentation baseline). However, when oversampling was combined with *COrE* corruption method, AraBERTv2 reached its peak average of 0.386, with standout improvements on P4 (0.094 $\rightarrow$ 0.391). This indicates that corruption-based augmentation coupled with oversampling can greatly benefit low-resource settings, where data scarcity and class imbalance are pronounced.

For CAMeLBERT-mix, *COrE* also showed clear benefits, outperforming both the no-augmentation baseline (17% improvement) and the CASE baseline (4% improvement). However, these gains were less consistent across prompts; for instance, *COrE* excels on P2 (0.570, a 42% jump from the no augmentation baseline) but lags on P4 (0.139 vs. 0.214 for the no augmentation baseline). Also, oversampling did not improve performance beyond *COrE*, and interestingly, naïve oversampling alone produced the highest overall average for CAMeLBERT-mix model (0.346), largely due to its strong performance on P4. Despite these variations, both corruption-based variants, *COrE* alone and *COrE* combined with oversampling, show that our trait-specific augmentation method remains beneficial for this model. The relatively smaller improvements compared to AraBERTv2 suggest that model architecture may influence the effectiveness of augmentation strategies, with CAMeLBERT-mix appearing less sensitive to corruption-induced perturbations. This may further reflect pre-training differences: AraBERTv2, trained primarily on Modern Standard Arabic, aligns closely with formal essay scoring, whereas CAMeLBERT-mix includes a broader dialectal diversity.

[9]Average performance is reported for completeness but should be interpreted cautiously, as prompts differ substantially in test size.

Approach	Model	P1	P2	P3	P4	Avg.
Without Augment.	AraBERT	0.644	**0.601**	0.009	0.094	0.337
	CAMeLB	0.584	0.402	-0.162	0.214	0.259
CASE Corruption	AraBERT	0.650	0.505	-0.048	0.238	0.336
	CAMeLB	0.596	0.518	-0.092	0.142	0.291
Naïve Oversamp.	AraBERT	0.602	0.509	-0.015	0.033	0.282
	CAMeLB	0.558	0.513	-0.076	0.389	0.346
COrE Corruption	AraBERT	0.710	0.550	**0.030**	0.179	0.367
	CAMeLB	0.623	0.570	-0.123	0.139	0.302
COrE + Oversamp.	AraBERT	**0.712**	0.503	-0.064	**0.391**	**0.386**
	CAMeLB	0.589	0.549	-0.074	0.111	0.294

Table 4: Average QWK performance of **three** runs on the test set. **Bold** values indicate the best performance per prompt, and underlined values represent the second best.

Notably, performance on prompts 3 and 4 remained lower than on prompts 1 and 2 across both models, likely due to smaller test sets, which increase variance, and longer essays with more complex structures, which challenge scoring and model generalization. Despite these difficulties, data augmentation, whether via naïve oversampling or *COrE*+oversampling, yielded improvements, with the most gains observed on prompt 4. This underscores the effectiveness of trait-specific augmentation in mitigating the impact of limited and complex data.

Overall, these findings confirm that *COrE* consistently outperforms the no-augmentation and CASE baselines for both models, highlighting *COrE*'s ability to generate more semantically relevant augmentations tailored to organizational flow. By injecting trait-specific disruptions, *COrE* enhances the model's ability to capture fine-grained distinctions, offering a practical solution to data scarcity in low-resource AES settings. These insights validate our approach and open directions for refining augmentations to handle under-resourced languages.

7. Conclusion and Future Work

This work proposes *COrE*, a corruption-based data augmentation method for Arabic AES, targeting the **organization** trait, to address the challenging problem of training data scarcity. By applying controlled distance-aware sentence-level swapping, we simulate varying degrees of disorganization while preserving linguistic integrity. Results across both AraBERTv2 and CAMeLBERT-mix confirms that *COrE* improves model perfor-

mance in low-resource scenarios. These findings position corruption-based augmentation methods as a promising and effective solution to the data scarcity problem in Arabic AES.

Future work includes evaluating the approach with other Arabic-specific pre-trained models, extending augmentation to other traits through trait-specific strategies, applying hyperparameter tuning for improved AES performance, and assessing the augmentation performance in a cross-prompt setup to test its generalizability.

8. Acknowledgments

This work was made possible by NPRP grant# NPRP14S-0402-210127 from the Qatar Research Development and Innovation (QRDI) Council. The statements made herein are solely the responsibility of the authors.

9. Limitations

This preliminary study explores data augmentation for Arabic, a low-resource language, with several acknowledged limitations. First, findings are restricted to the organization trait, leaving other writing dimensions unexplored. Second, *COrE* intentionally degrades essay quality, making it unsuitable for generating high-scoring essays (e.g., score 5), and thus limiting coverage across the full score range. However, we believe that well-written essays can be generated by LLMs as they shown effectiveness in this task. Third, the relatively small test set sizes of prompts 3 and 4 constrain evaluation stability; empirical gains are primarily observed on prompts with sufficiently sized test sets (P1 and P2), while results for smaller test sets (P3 and P4) remain unstable and inconclusive. Finally, the task-specific experimental design may constrain generalizability across diverse prompts. These boundaries reflect the deliberate focus of this initial work but represent key areas where the approach could be expanded.

10. Bibliographical References

Abdelhamid M. Ahmed, Xiao Zhang, Lameya M. Rezk, and Wajdi Zaghouani. 2024. Building an Annotated L1 Arabic/L2 English Bilingual Writer Corpus: The Qatari Corpus of Argumentative Writing (QCAW). *Corpus-based Studies across Humanities*, 1(1):183–215.

Mansour Alghamdi, Mohamed Alkanhal, Mohamed Al-Badrashiny, Abdulaziz Al-Qabbany, Ali Areshey, and Abdulaziz Alharbi. 2014. A Hybrid Automatic Scoring System for Arabic Essays. *AI Communications*, 27(2):103–111.

Aqil M. Azmi, Maram F. Al-Jouie, and Muhammad Hussain. 2019. AAEE – Automated Evaluation of Students' Essays in Arabic Language. *Information Processing Management*, 56(5):1736–1752.

May Bashendy, Salam Albatarni, Sohaila Eltanbouly, Walid Massoud, Houda Bouamor, and Tamer Elsayed. 2025. TAQEEM 2025: Overview of the First Shared Task for Arabic Quality Evaluation of Essays in Multi-dimensions. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, China.

May Bashendy, Salam Albatarni, Sohaila Eltanbouly, Eman Zahran, Hamdo Elhuseyin, Tamer Elsayed, Walid Massoud, and Houda Bouamor. 2024. QAES: First Publicly-Available Trait-Specific Annotations for Automated Scoring of Arabic Essays. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 337–351, Bangkok, Thailand. Association for Computational Linguistics.

Domenic V Cicchetti. 1981. Testing the Normal Approximation and Minimal Sample Size Requirements of Weighted Kappa When the Number of Categories is Large. *Applied psychological measurement*, 5(1):101–104.

Jacob Cohen. 1968. Weighted kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit. *Psychological Bulletin*, 70(4):213–220.

S.A. Crossley, Y. Tian, P. Baffour, A. Franklin, M. Benner, and U. Boser. 2024. A Large-Scale Corpus for Assessing Written Argumentation: PERSUADE 2.0. *Assessing Writing*, 61:100865.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data Augmentation Using Large Language Models: Data Perspectives, Learning

Paradigms and Challenges. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1679–1705, Bangkok, Thailand. Association for Computational Linguistics.

Ahmed Adel ElSabagh, Shahira Shaaban Azab, and Hesham Ahmed Hefny. 2025. A Comprehensive Survey on Arabic Text Augmentation: Approaches, Challenges, and Applications. *Neural Computing and Applications*, pages 1–34.

Rayed Ghazawi and Edwin Simpson. 2024. Automated Essay Scoring in Arabic: A Dataset and Analysis of a BERT-based system. *arXiv preprint arXiv:2407.11212*.

Rayed Ghazawi and Edwin Simpson. 2025. How Well Can LLMs Grade Essays in Arabic? *Computers and Education: Artificial Intelligence*, 9:100449.

Kshitij Gupta. 2023. Data Augmentation for Automated Essay Scoring using Transformer Models. In *International Conference on Artificial Intelligence and Smart Communication (AISC)*, pages 853–857. IEEE.

Nizar Habash and David Palfreyman. 2022. ZAEBUC: An Annotated Arabic-English Bilingual Writer Corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 79–88, Marseille, France. European Language Resources Association.

Bohan Li, Yutai Hou, and Wanxiang Che. 2022. Data Augmentation Approaches in Natural Language Processing: A Survey. *AI Open*, 3:71–90.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. 2024. Best Practices and Lessons Learned on Synthetic Data. *arXiv preprint arXiv:2404.07503*.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. ASAP++: Enriching the ASAP Automated Essay Grading Dataset with Essay Attribute Scores. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Chatrine Qwaider, Bashar Alhafni, Kirill Chirkunov, Nizar Habash, and Ted Briscoe. 2025. Enhancing Arabic Automated Essay Scoring with Synthetic Data and Error Injection. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 549–563, Vienna, Austria. Association for Computational Linguistics.

Marwan Sayed, Sohaila Eltanbouly, May Bashendy, and Tamer Elsayed. 2025. Feature Engineering is not Dead: A Step Towards State of the Art for Arabic Automated Essay Scoring. In *Proceedings of The Third Arabic Natural Language Processing Conference*, pages 231–245, Suzhou, China. Association for Computational Linguistics.

Haneul Yoo, Jieun Han, So-Yeon Ahn, and Alice Oh. 2025. DREsS: Dataset for Rubric-based Essay Scoring on EFL Writing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13439–13454, Vienna, Austria. Association for Computational Linguistics.