Is One Dataset Enough for Evaluation? Studying Generalizability of Automated Essay Scoring Models

Sohaila Eltanbouly, Marwan Sayed, Tamer Elsayed
Computer Science and Engineering Department, Qatar University, Doha, Qatar
{se1403101, me2104862, telsayed}@qu.edu.qa

Abstract

Automated Essay Scoring (AES) has made significant advancements in writing assessment. Recently, cross-prompt AES has gained attention because of its focus on generalizing to unseen prompts. Despite the promise of these advancements, a critical question remains: how generalizable and robust are those models when applied to diverse datasets? This study assesses the generalizability of eight cross-prompt AES models across three different datasets. We employ two experimental setups: the within-dataset approach, where both training and testing occur on the same dataset, and the cross-dataset approach, which challenges the models by evaluating their performance on previously unseen datasets. The experimental results show significant performance inconsistencies, highlighting that relying on a single dataset is insufficient for building robust and generalizable AES systems.

Keywords: Automated Essay Scoring, Cross-prompt

1. Introduction

Automated Essay Scoring (AES) aims to automatically evaluate the quality of written essays, providing fast and consistent assessments. Existing AES approaches can be broadly categorized into prompt-specific models, which are trained and evaluated on essays from the same writing prompt, and cross-prompt models, which are trained on multiple prompts to enable generalization to unseen prompts. While prompt-specific AES often achieves high accuracy, it struggles to generalize to new prompts without additional training. Cross-prompt AES, on the other hand, aims to build more flexible and generalized scoring models but faces significant challenges due to variability in topics, writing styles, grade levels, and scoring rubrics across prompts.

Early work for cross-prompt AES proposed a two-stage deep neural network for essay scoring (Jin et al., 2018). Subsequent research adapted several learning approaches, including multi-task learning (Ridley et al., 2021; Do et al., 2023), contrastive learning (Chen and Li, 2023), meta-learning (Chen and Li, 2024), and mixture of experts (Wang and Liu, 2025). Interestingly, (Li and Ng, 2024b) showed that a feature-based neural model can achieve state-of-the-art (SOTA) results on holistic scoring. However, the majority of these systems have been evaluated on a single dataset, raising concerns about their generalizability to diverse proficiency levels and writing conditions.

A major limitation of current AES research is the reliance on the ASAP dataset,[1] with only a few studies exploring alternative datasets such as TOEFL11 (Lee et al., 2024; Jiang et al., 2023), ELLIPSE (Do et al., 2024b), and Persuade (Yang et al., 2024). This narrow focus on a single dataset for evaluation raises concerns about the generalizability of existing models. ASAP's limited scope, particularly the lack of English Language Learners (ELL), may lead to overfitting and reduced robustness on more diverse writing samples (Li and Ng, 2024a). Consequently, a critical question arises: what happens when new/unseen prompts emerge from a different distribution? This issue is particularly relevant in real-world scenarios, such as international assessments like IELTS and TOEFL, which are taken by diverse student populations. In such contexts, the ability of AES models to perform accurately across different scenarios is vital for ensuring fair assessments.

Our work aims to bridge this gap by evaluating the generalizability of leading cross-prompt AES models for holistic scoring. Specifically, we evaluate eight models across three AES datasets that differ in prompt topics, student populations, and essay characteristics. We adopt two experimental setups: (1) within-dataset, which is the conventional cross-prompt setup, where training and testing are performed on essays from the same dataset but from different prompts, and (2) cross-dataset, where models are trained on one dataset and tested on unseen datasets. The objective is to evaluate existing AES models in more challenging yet realistic scenarios that reflect real-world conditions, where AES systems must assess essays of varying writing quality levels and language proficiencies. Our contribution is three-fold:

1. Investigating the generalizability of current cross-prompt AES models across three datasets.

[1] https://www.kaggle.com/c/asap-aes

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 431–440

2. Examining the suitability of the datasets for developing a robust AES model.

3. Offering recommendations for future considerations within the AES community.

2. Related Work

Cross-prompt AES presents a considerable challenge due to inherent differences between writing prompts, such as prompt category and scoring criteria. Most cross-prompt AES research has relied on the ASAP and ASAP++ (Mathias and Bhattacharyya, 2018a) datasets, which have become the primary benchmarks for English AES.

Early cross-prompt AES (Jin et al., 2018) introduced a two-stage neural model (TDNN) that first predicts extreme scores on a target prompt and then fine-tunes on these examples. Ridley et al. (2020) proposed the Prompt-Agnostic Essay Scorer (PAES), a neural model trained on POS tag sequences and handcrafted features. Do et al. (2023) enhanced PAES by incorporating prompt-text features. Recent studies have explored diverse learning strategies, including multi-task learning (Li and Ng, 2024b), contrastive learning (Chen and Li, 2023), meta-learning (Chen and Li, 2024), mixture-of-experts (Wang and Liu, 2025), and LLM-based approaches (Xu et al., 2025). However, all these methods are evaluated solely on ASAP/ASAP++, raising concerns about their generalizability to other datasets and the risk of overfitting to ASAP.

Few studies have evaluated AES systems beyond ASAP. Jiang et al. (2023) used ASAP and TOEFL11 (Blanchard et al., 2013) to assess a prompt-aware model that disentangles prompt-invariant and prompt-specific features, while Lee et al. (2024) proposed a Multi-Trait Specialization framework with an LLM-based conversational setup, also tested on both datasets. The ELLIPSE dataset (Crossley et al., 2023a), focused on ELL students and diverse prompts, has recently gained attention. Do et al. (2024a) introduced ArTs, which reframes AES as score generation using a fine-tuned T5 model, later enhanced with reinforcement learning (Do et al., 2024b); both were evaluated on ELLIPSE and ASAP. In addition, Chen et al. (2024) proposed a Multi-Task Automated Assessment framework with dynamic learning rate decay and orthogonality constraints, evaluated on ASAP and ELLIPSE. Although these studies demonstrate some level of generalizability across datasets, they focus on prompt-specific AES.

Despite progress in AES, there is still limited evaluation of how well proposed models generalize in scenarios that closely mirror real-world deployment. To address this gap, we conduct a systematic evaluation of AES models across three diverse datasets and analyze how factors such as prompt characteristics, essay length, and student nativity affect performance. Unlike prior work that primarily focuses on cross-prompt settings within a single dataset, our study provides the first comprehensive analysis of cross-dataset transfer and underscores the importance of dataset diversity for building robust AES models.

3. Experimental Design

This study aims to assess the generalizability of existing cross-prompt AES to unseen prompts and datasets. We investigate this through *two* experimental setups: *Within-dataset* and *Cross-dataset*. In this section, we formulate the two setups and discuss the examined models and datasets.

3.1. Problem formulation

Let $M = \{m\}$ be a set of AES models. Let $D = \{d\}$ be a set of AES datasets; a dataset d consists of a set P_d of N_d prompts, each constituting a set of essays written for one prompt $p \in P_d$.

3.1.1. Within-dataset Setup

This is the *typical* cross-prompt AES evaluation scenario, where *both* source (training) and target (testing) prompts come from the *same* dataset. Formally, we train and test each model $m \in M$ on a dataset $d \in D$ using N_d-fold (leave-one-prompt-out) cross-validation, such that one fold i of unseen target prompt p_d^i is used for testing, and the rest for training, as follows:

$$m_d^i = \mathcal{F}(P_d - \{p_d^i\}) \tag{1}$$

where $\mathcal{F}$ is the training function, and m_d^i is the trained model for the i^{th} target prompt. The performance on the target prompt p_d^i is evaluated using an AES evaluation measure $\mathcal{Q}$. The final performance of model m on dataset d is typically the average score across all prompts:

$$S(m_d) = \frac{1}{N_d} \sum_{i=1}^{N_d} \mathcal{Q}(m_d^i(p_d^i)) \tag{2}$$

This setup enables us to investigate the consistency of model performance over different datasets, i.e., whether models performing well on one dataset can achieve comparable performance on another.

3.1.2. Cross-dataset Setup

It involves training the model on one dataset and testing it on an *entirely different* dataset. Formally, a model m is trained using one entire *source* dataset $d_s \in D$.

$$m_{d_s} = \mathcal{F}(P_{d_s}), \quad d_s \in D \tag{3}$$

Model	Features	Approx. Size	Arch.	Code?
Hi-att	Character-level	257K	STL	✓
AES-aug	N-gram	251K	MTL	✓
PAES	POS, R	49k	STL	✓
CTS	POS, R	460k	MTL	✓
ProTACT	POS, GloVe, R	1.7M	MTL	✓
FB-NN	R, U, L	24K	STL	✗
FB-RF	R	-	STL	✗
FB-SVR	R	-	STL	✗

Table 1: Summary of models used in our experiments. R, U, and L features are those proposed by Ridley et al. (2020), Uto et al. (2020), and Li and Ng (2024b), respectively. STL and MTL denote Single-Task and Multi-Task (with trait scoring) Learning architectures.

Then it is tested and evaluated on an *unseen target* dataset $d_t \in D, d_t \neq d_s$, i.e., on each prompt $p_{d_t} \in P_{d_t}$, before the final performance is computed:

$$S(m_{d_s}|d_t) = \frac{1}{N_{d_t}} \sum_{p_{d_t} \in P_{d_t}} \mathcal{Q}(m_{d_s}(p_{d_t})) \quad (4)$$

This setup represents a more rigorous and challenging evaluation scenario, as the model encounters prompts from unknown distributions, allowing for a deeper assessment of robustness and generalization across diverse essay collections.

3.2. AES Models

To conduct this study, we selected eight cross-prompt AES models (compared in Table 1). Six of the models are neural network-based (NN):

- **Hi-att** (Dong et al., 2017) model is a leading prompt-specific holistic model that has been trained in a cross-prompt setting (Ridley et al., 2021). It employs a hierarchical representation approach using Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) recurrent neural network (RNN) to capture the word, sentence, and essay representations.

- **AES-aug** model is developed for prompt-specific trait scoring (Hussein et al., 2020), and has been utilized as a cross-prompt baseline. It is built upon the work of (Taghipour and Ng, 2016), transforming the approach from holistic scoring to trait scoring through multitask training. It uses a CNN for feature extraction, an LSTM for essay representation, and a linear layer for joint holistic and trait scoring.

- **PAES** model is a significant contribution to the field of cross-prompt holistic scoring (Ridley et al., 2020). Its architecture closely resembles that of Hi-att (Dong et al., 2017), with a notable distinction that the word representations are derived from POS embeddings. Additionally, PAES incorporates a set of prompt-independent features as inputs to the final linear layer. These hand-crafted features have proven advantageous and are extensively utilized in subsequent studies.

- **CTS** model pioneered the cross-prompt trait scoring task (Ridley et al., 2021). It utilizes the same core architecture as PAES, where it uses CNNs for POS-based sentence representations, followed by LSTM trait-specific representations. These are combined with hand-crafted features and refined using a trait-attention mechanism, enabling traits to leverage relevant information from each others.

- **ProTACT** (Do et al., 2023) constructs prompt-aware essay representations using CNNs and LSTMs on POS embeddings, combined with GloVe-based prompt embeddings and multi-head attention. The resulting representations, concatenated with handcrafted features, are used to predict holistic and trait-specific scores. The model further employs a trait-similarity loss alongside MSE to improve trait-based scoring.

- **Feature-based NN** is the current SOTA model for holistic scoring, proposed by Li and Ng (2024b). This model employs a simple neural network architecture consisting of two layers. This model relies entirely on engineered features, which consist of the widely used 86 features (Ridley et al., 2020), 25 features from (Uto et al., 2020), and 1,423 newly proposed features.

The selection criteria for the NN models are: 1) code availability or straightforward implementation, and 2) models trained without prior knowledge about the target prompts. Additionally, inspired by the findings of Li and Ng (2024b), where a simple NN with features outperformed more complex models, we explore 2 traditional machine-learning approaches that were prevalent in the early AES research:

- **Random Forest (FB-RF)** algorithm was previously employed by Mathias and Bhattacharyya (2018b) as an early benchmark on the ASAP dataset. Hence, we select the RF algorithm to implement a simple feature-based model using the feature set proposed by Ridley et al. (2020).

- **Support Vector Regression (FB-SVR)** algorithm was utilized in early prompt-specific AES on the ASAP dataset (Zesch et al., 2015). We also select the SVR algorithm to implement a simple feature-based model using the feature set proposed by Ridley et al. (2020).

3.3. Datasets

We selected three datasets that differ in size and essay characteristics (compared in Table 2). The three selected datasets are:

ASAP The Automated Student's Assessment Prize (ASAP) dataset is the most widely used dataset for AES development, published as part of a Kaggle competition. It contains 12,978 essays written in English in response to 8 prompts by students from 7^{th} to 10^{th} grade levels. Each of these prompts has a different number of responses. Moreover, different rubrics with different score ranges are used to score each prompt.

Persuade The Persuade 2.0[2] dataset stands out because of its larger size, as well as containing demographic attributes of the students (Crossley et al., 2023b). It contains 25k persuasive essays written by students from 6^{th} to 12^{th} grade levels for 15 different prompts. All the essays are assigned a holistic score, graded using a standardized rubric.

ELLIPSE The English Language Learner Insight, Proficiency and Skills Evaluation (ELLIPSE)[3] dataset has been recently published as a strong benchmark for AES evaluation (Crossley et al., 2023a). The dataset comprises about 6,500 essays written by ELL for 44 distinct prompts, covering 8^{th} to 12^{th} grade levels. ELLIPSE contains a diverse student population, making it suitable for AES evaluation. All essays are evaluated using a standardized rubric that assesses 6 traits, in addition to a holistic score. It is noteworthy that we used 37 out of the 44 available prompts in our experiments, where the remaining 7 overlap with the Persuade dataset. Thus, we removed these prompts from ELLIPSE to avoid data contamination in the cross-dataset evaluation. The excluded prompts are: Distance learning, Grades for extracurricular activities, Community service, Seeking multiple opinions, Cell phones at school, Mandatory extracurricular activities, and Summer projects.

4. Experimental Setup

In this section, we outline the setup and implementation details used to conduct our experiments, including the data splits, hyperparameter tuning, model training, and the evaluation measure.

[2]https://github.com/scrosseye/persuade_corpus_2.0

[3]https://github.com/scrosseye/ELLIPSE-Corpus

[4]7-prompts are excluded due to their overlap with Persuade

	ASAP	ELLIPSE	Persuade
Essays	13k	6.5k	25k
Prompts	8	44[4]	15
Essay category	P, S, N	P	P, S
Grade Levels	7-10	8-12	6-12
Language	Native	ELL	Native+ELL
Score Range	Varies	1-5 (+0.5)	1-6
Avg. Essay Length	281	427	418
Avg. Essays/Prompt	1622	147	1733

Table 2: Comparison over the datasets. Last 2 rows are averages. Essay length is in words. Essay categories are persuasive (P), narrative (N), and source-dependent (S).

4.1. Data Splits

For the within-dataset setup, all models were evaluated using leave-one-prompt-out cross-validation to ensure that the target prompt remains unseen during training. The training data was further divided into 85% for training and 15% for validation. For the ASAP and Persuade datasets, we employed 8-fold and 15-fold cross-validation, respectively, corresponding to the number of prompts in each dataset. For ELLIPSE, we divided the 37 prompts into 9 folds, each containing 4 prompts except one fold containing 5.[5] For the cross-dataset setup, models were trained on one dataset and evaluated on a completely different dataset. In cases where validation sets were required for early stopping, 15% of the training dataset was set aside for validation.

4.2. Hyperparameter Tuning

To tune the hyperparameters, we used the Tree-structured Parzen estimator (TPE), which is a Bayesian optimization approach for hyperparameter tuning (Bergstra et al., 2011). We used the TPESampler from the optuna library.[6] We set the number of trials to 20 with 5 startup trials. We also used the MedianPruner[7] to early-stop unpromising trials. For all the NN-based models, we used a batch size of 16 and a maximum number of 50 epochs. We also used early stopping based on the QWK score on the validation set, with a patience of 10, to prevent overfitting. We set the random seed to 12 to ensure reproducibility.

For the within-dataset setup, the models are

[5]The folds used in our experiments for ELLIPSE dataset are available at https://drive.google.com/drive/folders/1k0r3dTQTdGTkVQCLO8J249ia2b8bYdkO

[6]https://optuna.readthedocs.io/en/stable/reference/samplers/generated/optuna.samplers.TPESampler.html

[7]https://optuna.readthedocs.io/en/stable/reference/generated/optuna.pruners.MedianPruner.html

Model	Hyperparameter	Value
All NN models•	Learning rate	[0.01,0.001,0.0001]
	Dropout	[0.3, 0.4, 0.5]
5-NN models*	CNN kernal	[3, 5]
	CNN filters	[32, 64, 128]
	LSTM units	[32, 64, 128]
ProTACT	TS loss threshold δ	[0.5,0.6,0.7]
	Loss function weighting λ	[0.5,0.6,0.7]
FB-NN	Feature selection threshold	[0.1, 0.2, 0.3, 0.4, 0.5, 0.6]
	Hidden units	[32, 64, 128]
FB-RF	Max depth	[3, 4, 5, 6, 7, 8, 9, 10]
	Max features	[0.25, 0.5, 0.75, 1.0]
	Max samples	[0.25, 0.5, 0.75, 1.0]
FB-SVR	C	[0.1, 1, 10]
	Max iter	[500, 1000, 1500, -1]
	Gamma	[1, 0.1, 0.01, 0.001, 0.0001,scale,auto]

Table 3: Hyperparameter search spaces for the eight models. •All NN refers to hyperparameters common to all six NN-based models. *5-NN indicates common hyperparameters for Hi-att, AES-aug, PAES, CTS, and ProTACT.

tuned separately for each target fold, and the model with the best hyperparameters based on the validation set is used to evaluate the performance on the unseen target fold. To train the cross-dataset models on the entire training dataset, we selected the most frequent hyperparameters from the cross-validation folds of the within-dataset models. The same procedures are applied across all models and datasets. Table 3 presents the hyperparameters search space for all models.

For the ASAP dataset, the six selected NN models were re-trained using the same hyperparameter tuning procedure, as the original studies provided limited information on how hyperparameter values were determined. This approach ensures a fair comparison across all models and datasets. Consequently, the obtained results differ from those reported in the original studies.

4.3. Model Training

In this section, we describe the training of the 8 AES models. All the models are trained on an Azure VM equipped with 2 NVIDIA A10 GPUs and an AMD EPYC 74F3 24-Core Processor.

Models with Existing Code For the following four models: Hi-att (Dong et al., 2017), AES-aug (Hussein et al., 2020), PAES (Ridley et al., 2020), and CTS (Ridley et al., 2021), we utilized the implementation provided by Ridley et al. (2021).[8] For ProTACT, we used the official implementation released by the authors.[9] The MTL models are not trained on the Persuade dataset due to the unavailability of trait scores.

FB-NN Model Due to the unavailability of the model's code, we re-implemented it based on the paper's description (Li and Ng, 2024b). For the *Feature Extraction*, the model utilizes three distinct feature sets. The first set comprises 86 features (Ridley et al., 2020), with their extraction code publicly available. The second set includes 25 features (Uto et al., 2020), all extracted using the Natural Language Toolkit (NLTK).[10] The third feature set is proposed by Li and Ng (2024b) and categorized into four groups. First, part-of-speech (POS) features are extracted using the NLTK POS tagset.[11] Second, prompt adherence features are generated using the essay and prompt embeddings using the all-mpnet-base-v2 model from the sentence-transformers library.[12] Finally, pronoun features are extracted using an extensive list of pronouns categorized into seven groups: first person, second person, third person, demonstrative, interrogative, relative, and indefinite. The model architecture follows the one described in the corresponding study (Li and Ng, 2024b), consisting of two hidden layers with ReLU activation, and an output layer with a sigmoid activation function, implemented using the PyTorch library.

FB-RF & FB-SVR We used the RandomForestRegressor and Support Vector Regression mod-

[8] https://github.com/robert1ridley/cross-prompt-trait-scoring

[9] https://github.com/doheejin/ProTACT

[10] https://www.nltk.org/

[11] The POS tagset used is the Penn Treebank tagset

[12] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

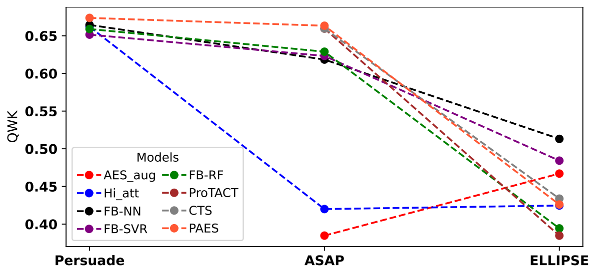

Figure 1: Within-dataset performance of the eight AES models.

Figure 2: Average absolute difference of model performance across datasets.

els from the sklearn library[13] to train the FB-RF and FB-SVR models, respectively. The models are trained using the 86 features proposed by Ridley et al. (2020), normalized to the [0, 1] range.

4.4. Evaluation Measure

For evaluation, we employ Quadratic Weighted Kappa (QWK) (Cohen, 1968), the standard metric in AES for measuring the agreement between human and system-assigned scores. QWK accounts for both the agreement and the degree of disagreement by assigning higher penalties to larger differences between predicted and true scores, making it suitable for ordinal scoring tasks.

5. Results and Discussion

In this section, we discuss our experimental results addressing two research questions: **RQ1**: How effectively do existing cross-prompt models generalize across different datasets? and **RQ2**: Are existing datasets suitable for developing robust generalizable models?, followed by a discussion and some recommendations.

5.1. Within-dataset Results (RQ1)

Figure 1 illustrates the within-dataset performance of models, showing interesting observations.

Inconsistent Model Performance Overall, the models exhibit *inconsistent* performance across the datasets. Among the NN models, Hi-att and AES-aug models show comparable performance over ASAP and ELLIPSE datasets; however, all the other models show clear discrepancy: higher QWK values on ASAP, but much lower on ELLIPSE. Showing strong competitiveness, the FB models outperform the other models on ELLIPSE, with FB-NN achieving the highest QWK, followed by FB-SVR. Moreover, they show comparable performance to the top-performing models over the ASAP and Persuade datasets.

Different Task Difficulty Generally, the differences in performance across models on ELLIPSE and Persuade are relatively smaller (with standard deviations of 0.04 and 0.01, respectively) than on ASAP (with a standard deviation of 0.11). However, tasks of ELLIPSE prove to be significantly more challenging to score, with QWK values ranging from 0.38 to 0.51. Conversely, Persuade is the least challenging (likely due to the larger size), with all models achieving high QWK values > 0.65.

Uncorrelated Model Ranking We further examine the ordinal associations between model rankings across datasets by computing the Kendall's-τ correlation between rankings of models (based on QWK) produced over each pair of datasets. The results indicate low correlations, with τ values ranging from -0.3 to 0.2, indicating that the different datasets assess different aspects of essay scoring that are supported in different ways by different AES models, leading to completely-different relative performance.

These findings emphasize that one dataset is insufficient for assessing the model's generalizability.

5.2. Cross-dataset Results (RQ2)

To capture the effect of cross-dataset evaluation, we marginalize the different models by computing, for each pair of datasets d_x and d_y, the average absolute difference between the performance of models in cross-dataset (trained on d_x but tested on d_y) and within-dataset scenarios (trained and tested on d_y). Figure 2 illustrates this across all dataset pairs. Furthermore, Figure 3 shows the performance of individual models with that setup.

Within- vs. Cross-dataset Setups Figure 2 shows that testing on ASAP becomes much harder when the models are trained on the other 2 datasets, likely due to the diversity of essay categories in ASAP, whereas the other datasets consist solely of persuasive essays. This is more prevalent when training on ELLIPSE, with an average

[13]https://scikit-learn.org/

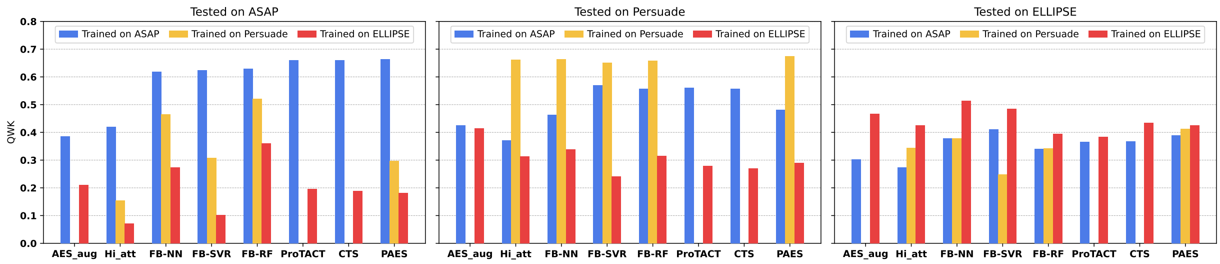

Figure 3: Cross-dataset performance of AES models. Each sub-figure shows the performance of the models trained on each dataset (color-coded) but tested on one *test* dataset (left to right: ASAP, Persuade, ELLIPSE, respectively).

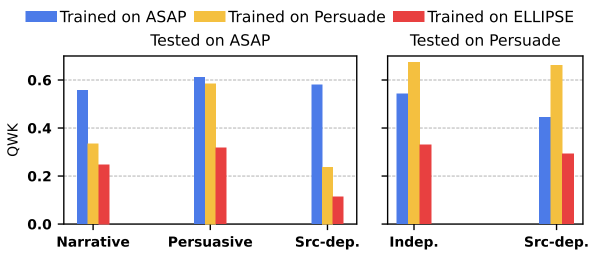

Figure 4: The average QWK over all models for each essay category in ASAP and Persuade datasets.

performance gap of 0.38. Similarly, an average gap of 0.36 is observed with training on ELLIPSE and testing on Persuade. Despite having the most prompts, ELLIPSE has the fewest essays, indicating that increasing the size of training data in terms of essays is more important than prompts for robust performance. In contrast, testing on ELLIPSE was less challenging for models trained on the other datasets, with average performance gaps of 0.1 and 0.09 only, indicating the greatest robustness.

Datasets & Models As expected, Figure 3 shows, across all models, that the cross-dataset setup yields lower performance than the within-dataset setup. However, performance patterns vary by dataset. Specifically, training on ELLIPSE consistently results in the worst performance, whereas training on Persuade produces the best performance in most evaluation scenarios. This reinforces our earlier observation about the size of training data. For the models, FB models demonstrate better generalization in the cross-dataset setup, with at least one FB model consistently ranked in the top 2 across all scenarios.

5.3. Dataset Analysis

Performance of cross-dataset models is indeed influenced by the characteristics of the training dataset. We discuss three key factors affecting the performance.

Essay Category Figure 4 demonstrates the critical influence of *essay category* on the performance of cross-dataset evaluation. The important role of the essay category is evident in the superior performance on the persuasive essays compared to other essay categories in the ASAP dataset. Notably, training on Persuade results in only a 3-point reduction in performance compared to training on ASAP. Interestingly, when trained on the Persuade dataset, the FB-NN and FB-RF models demonstrated approximately 10-point improvements on ASAP prompts 1 and 2 (persuasive prompts) compared to models trained and tested exclusively on ASAP. Furthermore, the FB-NN model, when trained on Persuade, achieves the SOTA performance on ASAP prompt 2, outperforming all other FB and NN models across evaluation scenarios. This enhanced performance can likely be attributed to the Persuade dataset's large size and its exclusive focus on persuasive essays, enabling the models to better capture features specific to persuasive writing. Moreover, the independent prompts in Persuade yield higher QWK than the source-dependent ones in the cross-dataset evaluation scenarios, likely because all persuasive prompts in both ASAP and ELLIPSE are independent. These findings highlight the need for further research into advanced techniques to generalize across essay categories. This is particularly crucial given the distinct rubrics and scoring criteria for different essay types, where scores depend not only on general writing quality but also on adherence to category-specific requirements (e.g., argumentative coherence in persuasive essays).

Essay Length Another factor influencing model performance is *essay length*. Persuade and ELLIPSE datasets have similar average essay lengths, whereas ASAP essays are generally shorter. Figure 5 shows the average QWK across all models for each ASAP prompt, indicated by its average essay length and category. The results indicate that cross-dataset models achieve higher accuracy when the test essays have lengths similar to those in the training data. Prompts with an aver-

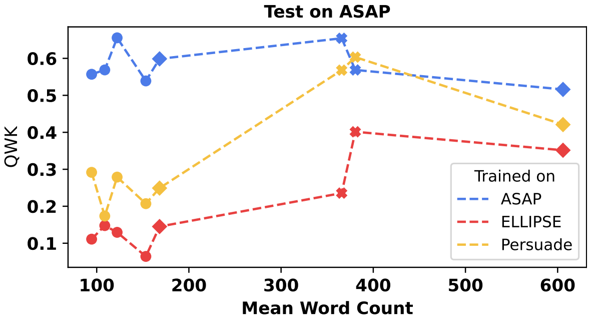

Figure 5: The average QWK over all models for each prompt in ASAP, sorted by the average essay length. Prompts annotated by • are source-dependent, ⋄ are narrative, and × are persuasive.

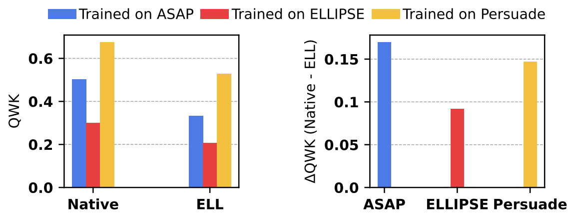

Figure 6: Performance of models when tested on Persuade essays written by native and ELL students. Left: average QWK for each group. Right: performance drop (ΔQWK) between native and ELL students.

age length < 200 words exhibit lower performance in cross-dataset settings, likely because Persuade and ELLIPSE contain few essays in this range. In contrast, prompts with longer essays demonstrate better performance. Notably, ASAP prompt 8 (narrative prompt with average essay length of 600 words) shows a smaller performance gap of 16 points when trained on ELLIPSE and 9 points when trained on Persuade, compared to the much larger gaps observed for source-dependent prompts (averaging 47 points with ELLIPSE and 36 points with Persuade), despite the fact that narrative essays have a different writing style than persuasive essays. This finding highlights that essay length is a key factor influencing AES performance, even when the essay category differs between training and testing sets.

Native vs. ELL A key limitation of the ASAP dataset is its exclusive focus on native speakers. We investigate the performance of the models for native and ELL students in the Persuade dataset. Figure 6 presents the average performance of all models on the two student groups and the performance drop between the two groups. Overall, scoring essays written by ELL is more challenging, which is evident in the lower QWK observed with the ELL group when training and testing on Persuade. This can also explain the relatively lower performance generally observed with the ELLIPSE dataset, which contains only ELL students. In the cross-dataset setup, models trained on ASAP show a 17-point QWK gap between native and ELL groups, whereas models trained on ELLIPSE show a smaller 9-point gap. Although performance on ELL students is lower for both datasets, ELLIPSE-trained models exhibit more balanced performance across the two groups. These findings emphasize the importance of including students from diverse backgrounds and proficiency levels when building robust AES models.

These findings further confirm that existing datasets alone are insufficient for developing generalizable AES systems for real-world scenarios. Relying on a single dataset may lead to biased models that fail to generalize across diverse writing styles and proficiency levels.

5.4. Discussion and Recommendations

Developing robust and generalizable AES systems remains a significant challenge. Based on our findings, we highlight three important considerations:

One dataset is not sufficient Most AES studies focus on optimizing performance on the ASAP dataset. However, our findings suggest these models may “overfit” to ASAP, leading to inconsistent performance on other datasets. This shows that no single dataset can reliably assess AES model generalizability and underscores the need for evaluation on multiple datasets to ensure generalizability across diverse prompts, writing conditions, and student populations.

Dataset characteristics affect generalization The diversity of prompts, essay types, and student populations is crucial for building generalizable AES systems. Constructing an AES system based on a narrow population or specific essay category limits its applicability beyond the training conditions. Training datasets should be sufficiently diverse to capture a wide range of writing styles and proficiencies.

Complex models ⇏ better generalizability Complex models might be outperformed by simple models that are more efficient, interpretable, and generalizable. While improving performance measures is important, it is even more important to understand how models contribute to these improvements and in what scenarios.

6. Conclusion and Future Work

Cross-prompt AES research has largely relied on the ASAP dataset, raising concerns about generalizability. This study evaluates the generalizability of existing cross-prompt models and datasets using two training setups: within-dataset and cross-dataset. The findings highlight the fact that the performance of the models is not consistent across the datasets, as well as training on one dataset is not enough to build a generalizable and robust AES. Future studies should explore broader datasets, diverse model architectures, other languages, and the generalizability of trait scoring.

Acknowledgment

The work of Sohaila Eltanbouly was supported by GSRA grant# GSRA12-L-0413-250111, and the work of the other authors was supported by NPRP grant# NPRP14S-0402-210127, both from the Qatar Research Development and Innovation (QRDI) Council. The statements made herein are solely the responsibility of the authors.

7. Bibliographical References

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.

Shigeng Chen, Yunshi Lan, and Zheng Yuan. 2024. A multi-task automated assessment system for essay scoring. In *International Conference on Artificial Intelligence in Education*, pages 276–283. Springer.

Yuan Chen and Xia Li. 2023. PMAES: Prompt-mapping contrastive learning for cross-prompt automated essay scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1503, Toronto, Canada. Association for Computational Linguistics.

Yuan Chen and Xia Li. 2024. PLAES: Prompt-generalized and level-aware learning framework for cross-prompt automated essay scoring. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12775–12786, Torino, Italia. ELRA and ICCL.

Jacob Cohen. 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Scott Crossley, Yu Tian, Perpetual Baffour, Alex Franklin, Youngmeen Kim, Wesley Morris, Meg Benner, Aigner Picou, and Ulrich Boser. 2023a. The english language learner insight, proficiency and skills evaluation (ellipse) corpus. *International Journal of Learner Corpus Research*, 9(2):248–269.

Scott Andrew Crossley, Perpetual Baffour, Yu Tian, Alex Franklin, Meg Benner, and Ulrich Boser. 2023b. A large-scale corpus for assessing written argumentation: Persuade 2.0. *Available at SSRN 4795747*.

Heejin Do, Yunsu Kim, and Gary Lee. 2024a. Autoregressive score generation for multi-trait essay scoring. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1659–1666, St. Julian's, Malta. Association for Computational Linguistics.

Heejin Do, Yunsu Kim, and Gary Geunbae Lee. 2023. Prompt- and trait relation-aware cross-prompt essay trait scoring. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1538–1551, Toronto, Canada. Association for Computational Linguistics.

Heejin Do, Sangwon Ryu, and Gary Lee. 2024b. Autoregressive multi-trait essay scoring via reinforcement learning with scoring-aware multiple rewards. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16427–16438, Miami, Florida, USA. Association for Computational Linguistics.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 153–162, Vancouver, Canada. Association for Computational Linguistics.

Mohamed A. Hussein, Hesham A. Hassan, and Mohammad Nassef. 2020. A trait-based deep learning automated essay scoring system with adaptive feedback. *International Journal of Advanced Computer Science and Applications*, 11(5).

Zhiwei Jiang, Tianyi Gao, Yafeng Yin, Meng Liu, Hua Yu, Zifeng Cheng, and Qing Gu. 2023. Improving domain generalization for prompt-aware essay scoring via disentangled representation learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12456–12470, Toronto, Canada. Association for Computational Linguistics.

Cancan Jin, Ben He, Kai Hui, and Le Sun. 2018. TDNN: A two-stage deep neural network for prompt-independent automated essay scoring. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1088–1097, Melbourne, Australia. Association for Computational Linguistics.

Sanwoo Lee, Yida Cai, Desong Meng, Ziyang Wang, and Yunfang Wu. 2024. Unleashing large language models' proficiency in zero-shot essay scoring. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 181–198, Miami, Florida, USA. Association for Computational Linguistics.

Shengjie Li and Vincent Ng. 2024a. Automated essay scoring: A reflection on the state of the art. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17876–17888, Miami, Florida, USA. Association for Computational Linguistics.

Shengjie Li and Vincent Ng. 2024b. Conundrums in cross-prompt automated essay scoring: Making sense of the state of the art. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7661–7681, Bangkok, Thailand. Association for Computational Linguistics.

Sandeep Mathias and Pushpak Bhattacharyya. 2018a. 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).

Sandeep Mathias and Pushpak Bhattacharyya. 2018b. 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)*.

Robert Ridley, Liang He, Xin-yu Dai, Shujian Huang, and Jiajun Chen. 2021. Automated cross-prompt scoring of essay traits. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13745–13753.

Robert Ridley, Liang He, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2020. Prompt agnostic essay scorer: a domain generalization approach to cross-prompt automated essay scoring. *arXiv preprint arXiv:2008.01441*.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.

Masaki Uto, Yikuan Xie, and Maomi Ueno. 2020. Neural automated essay scoring incorporating handcrafted features. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6077–6088, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jiong Wang and Jie Liu. 2025. T-MES: Trait-aware mix-of-experts representation learning for multi-trait essay scoring. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1224–1236, Abu Dhabi, UAE. Association for Computational Linguistics.

Jiangsong Xu, Jian Liu, Mingwei Lin, Jiayin Lin, Shenbao Yu, Liang Zhao, and Jun Shen. 2025. Epcts: Enhanced prompt-aware cross-prompt essay trait scoring. *Neurocomputing*, 621:129283.

Kaixun Yang, Mladen Raković, Yuyang Li, Quanlong Guan, Dragan Gašević, and Guanliang Chen. 2024. Unveiling the tapestry of automated essay scoring: a comprehensive investigation of accuracy, fairness, and generalizability. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Torsten Zesch, Michael Wojatzki, and Dirk Scholten-Akoun. 2015. Task-independent features for automated essay grading. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 224–232, Denver, Colorado. Association for Computational Linguistics.