# How reliable is that explanation? Intrinsic Evaluation of XAI methods in Automated Essay Scoring models

Daniel Mora[1,2,*], Andrea Horbach[1,2]

[1]Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, Kiel, 24118, Germany
[2]Kiel University, Christian-Albrechts-Platz 4, Kiel, 24118, Germany

**Abstract**

The need of transparent and trustworthy Automated Essay Scoring (AES) models in the educational field is of paramount importance. Even though there are an increasing number of post-hoc local explainability (XAI) methods that can be applied to different model architectures and input features, it is not clear which ones are more reliable because different methods often give conflicting explanations for the same prediction. In this study, we carry out an intrinsic evaluation of different token- and feature-level attribution-based post-hoc XAI methods considering their faithfulness –does the explanation correspond to the model's scoring criteria?– and robustness –does the explanation provide similar explanations to similar texts?– on AES models for both classification and regressions tasks, and on two writing datasets (MEWS and ASAP) that differ in the writer's first language (English as L1 or L2). Main results indicate that the most effective explainability method varies by task: Saliency is the best performing method overall for classification, while Integrated Gradients is more suitable for regression. Additionally, the performance of XAI methods is also sensitive to dataset characteristics, revealing that models assign attributions to the input text in the ASAP dataset (English L1) more evenly, whereas in the MEWS dataset (English L2), models rely on fewer and more salient input features.

**Keywords**

explainability, post-hoc local explanation, faithfulness, robustness, automated essay scoring

## 1. Introduction

Assessing open-ended writing at scale remains a challenge in educational settings. Instructors struggle to provide timely, consistent, and high-quality feedback. Automated Essay Scoring (AES) systems are increasingly being explored to address these challenges, yet their adoption critically depends on whether their scoring processes are transparent and trustworthy for educators, students, and administrators.

This broader concern also aligns with regulatory initiatives, as the use of machine learning models expands into socially sensitive contexts. For example, the European Union has introduced the AI Act [1], which establishes regulatory requirements for systems deployed in areas considered high risk, such as Education (Art. 6 Annex III). Machine learning models used to assess learning outcomes or determine access to educational opportunities must provide a level of explainability that enables stakeholders (e.g. developers, teachers and school administrators, students and parents) to interpret model outputs. AES systems must therefore meet this regulatory framework in order to be deployed in real-world scenarios.

Despite growing interest in AES models with a focus on explainability [2, 3, 4], prediction justification, and more detailed feedback [5, 6, 7], and a continuously increasing availability of XAI methods [8, 9], there is still no quantitative comparison of XAI methods tailored to AES models, and their degree of usability is still an open issue in the educational field [10, 11, 12]. Main concerns relate to explanations that lack pedagogical value (i.e. construct-irrelevant features), appear arbitrary, or conflict with human judgement. For instance, Table 1 shows an example output of two token-level attribution methods, which assign a contribution to each token towards the final prediction determining if it increases or decreases the final score, and its magnitude. Even though both methods are explaining the same model's

*XAI-Ed 2026: Demystifying AI in Education and Learning Analytics through Explainability, Agency, and Transparency Workshop (XAI-Ed@LAK26), 27 April, 2026, Bergen, Norway*
*Corresponding author.
mora@leibniz-ipn.de (D. Mora); horbach@leibniz-ipn.de (A. Horbach)
0000-0002-2758-5198 (D. Mora); 0009-0004-3680-3304 (A. Horbach)


prediction, they produce conflicting attributions. In this regard, the explanations assigned to models deployed in educational applications should not only satisfy the requirements of AI experts, but also be coherent for end users by reflecting how the model actually arrives at its decisions, and by behaving consistently across similar inputs.

In this study, we focus on the quantitative evaluation of different XAI methods regarding two criteria, faithfulness and robustness [13, 14], as they are initial indicators that can be intrinsically evaluated. In educational assessment, faithfulness ensures that explanations genuinely reflect the model's scoring criteria, rather than artifacts or spurious correlations. Robustness ensures that explanations remain stable across similar texts, which is essential to avoid confusing or inconsistent feedback for students or instructors.

The XAI methods considered in this study are post-hoc local feature attribution methods drawn from two major categories: gradient- and SHAP-based methods. These approaches provide an intuitive explanation of how each input feature (e.g. tokens from the student's text or linguistic representations) contribute towards the model's prediction, and can be applied to different model architectures without requiring the model to be inherently interpretable.

As already suggested [15, 16], the evaluation of XAI methods is model- and task-dependent, therefore we compare two of the most widely used machine learning paradigms in AES: token-based XAI methods on transformer-based models for AES as a classification task, and linguistic feature-based models used for AES as a regression task. We test these methods on two writing datasets containing texts written in English, produced by writers with different first languages (L1 vs. L2 English): the ASAP dataset (prompts 1, 2, 7 and 8) for L1 English essays, and the MEWS dataset [17] for L2 English essays.

The contributions of this study are twofold. First, we present a quantitative comparison of post-hoc local feature attribution methods applied to AES models in two settings: token-level explanations for transformer-based models in a classification task, and feature-level explanations for linguistic feature-based models in a regression task. Second, we evaluate these methods across L1 and L2 writing datasets to examine their stability under differences in linguistic proficiency and writing style, given that AES models are commonly used to assess both populations. The results of this evaluation will highlight which XAI methods offer more reliable explanations considering different tasks and datasets, and are therefore suitable for further assessment regarding their complexity, that is, how understandable they are and the extent of their pedagogical value for end users, such as students or teachers.

**Table 1**
Comparison of two XAI methods that assign conflicting attribution scores for the same prediction. Green and red cells indicate positive and negative contributions, respectively.

| **Token** | dear | local | newspaper | , | i | think | effects | comp. |
|---|---|---|---|---|---|---|---|---|
| **Saliency** | 0.09 | 0.07 | 0.07 | 0.06 | 0.05 | 0.04 | 0.05 | 0.05 |
| **IG** | -0.07 | 0.00 | -0.03 | -0.17 | 0.06 | -0.01 | 0.00 | 0.01 |

## 2. Related Work

The implementation of AES models in educational settings has brought greater attention to transparency, explainability, and interpretability [18, 19]. Their integration into classrooms not only forces students and teachers to develop the skills to understand, manage, and critically engage with AI-based tools [20], but also creates a demand for transparency in AES systems [21].

To this regard, recent research has shown two common approaches on how to increase the interpretability of models. One direction is using an inherently interpretable model [22], whose internal working structure can be directly examined, such as linear or tree-based models [23, 24, 25].

These approaches rely on hand-engineered linguistic features, where model coefficients can be inspected to understand how particular textual attributes influence predicted scores [26]. While such models offer a higher degree of intrinsic transparency and sometimes achieve performance competitive

with state-of-the-art models [27], most research focuses on deep learning models, as they typically adapt better to large and heterogeneous datasets.

The other direction relies on post-hoc local explanations. In contrast to intrinsically interpretable models, post-hoc methods enable insight into otherwise opaque, black-box models, such as deep neural networks, without requiring them to be transparent by design. Instead, these explainability techniques are applied after model training to reveal how a prediction was produced for a particular sample [9].

In the AES context, these methods are particularly relevant because they can be used as feedback for learners and teachers, as they reflect the inner workings of the model. Examples include attention-based approaches [28, 29], which highlight the tokens a model focuses on during scoring and offer a proxy for identifying influential parts of an essay; gradient-based techniques [30], which estimate how small changes in the input affect the output, indicating which tokens most strongly impact the final prediction; and SHAP-based methods, which compute contribution scores for linguistic features specifying how much each feature accounts for the final prediction [11]. Recent work provides a comparison of AES models beyond performance [12], discussing also XAI methods (e.g. SHAP, LIME, and LLM-generated explanations), albeit without offering an objective evaluation of their reliability.

Even though commercial generative LLMs have been recently used for AES [31, 32], they were not included in our experiments due to their lack of transparency. While self-explanations have been proposed as an interpretability mechanism, they cannot be immediately trusted [16]. Similarly, for chain-of-thought, there is no independent way of determining whether the generated reasoning is faithful to the model's internal workings [33] or are potentially misleading [34].

## 3. Methodology

This study compares feature-level attribution methods applied to AES models in two settings (classification and regression), and on two writing datasets (MEWS and ASAP). For the classification task, we apply token-level XAI methods to a transformer-based model that is directly fed with the text essay, and evaluate regarding faithfulness using comprehensiveness [35], sufficiency [35], and Kendall's Tau correlation [36]. For the regression task, we apply feature-level XAI methods to a neural network that uses linguistic features as essay representations, and evaluate on faithfulness using the infidelity metric [37], and on robustness using the sensitivity metric [37]. Figure 1 depicts the overall experimental setup, highlighting the attribution-based XAI methods and their corresponding evaluation metrics.[1] We train and test the performance of the models in a 5-fold cross-validation setting to account for model variability on different test sets. Similarly, we evaluate the XAI methods on two of these five folds, reducing the dependence of the results on any specific data partition.

We next describe the datasets and XAI methods used in this study, and then specify the preprocessing procedure, model architectures, and evaluation metrics of the XAI methods specific to the classification and regression settings.

### 3.1. Datasets

In our experiments, we use two writing datasets commonly used in AES research: MEWS [17] and ASAP (prompts 1, 2, 7 and 8).[2] Both datasets consist of English essays written in response to a debatable prompt that can be answered without external resources, but they differ in the writers' first language. MEWS contains 4,495 essays written by L2 English (German-speaking EFL learners) grouped into two prompts (prompts TE and AD), while ASAP (prompts 1, 2, 7 and 8) includes 5,875 texts written by L1 English speakers. All texts are manually scored at a holistic level following a prompt-specific evaluation rubric. Figure 2 shows the score distributions per prompt and dataset for both task setups (classification and regression) after preprocessing raw values (see 3.3 and 3.4). In both cases, most texts fall into the middle of the scoring range, with relatively few samples at the extreme scores or classes. We split the

[1] Source-code available at https://github.com/melanchthon19/xai-aes-intrinsic-evaluation.git
[2] https://www.kaggle.com/c/asap-aes/data

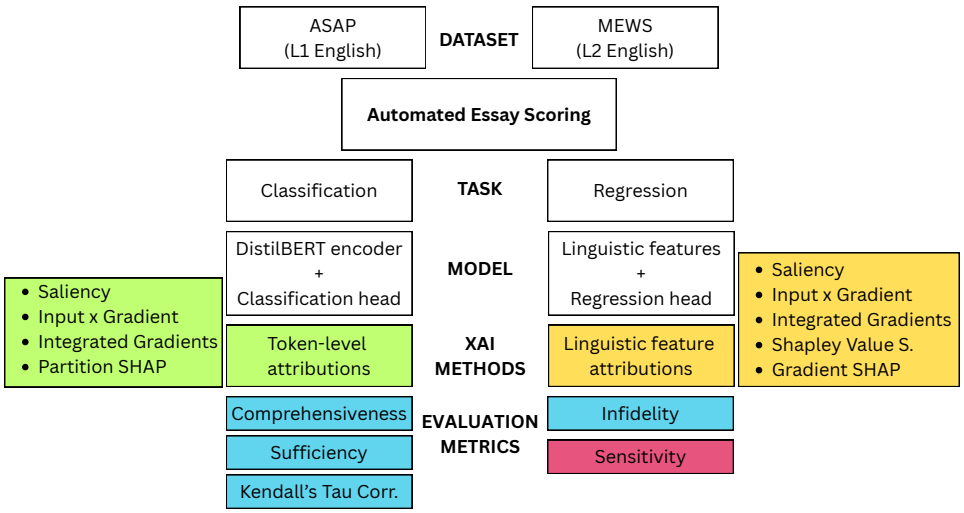



**Figure 1:** Overview of the experimental setup, highlighting the XAI methods applied to each task setting and their corresponding evaluation metrics. Token-level attribution methods are shown in light green, linguistic-feature attribution methods in yellow, faithfulness metrics in light blue, and the robustness metric in red.

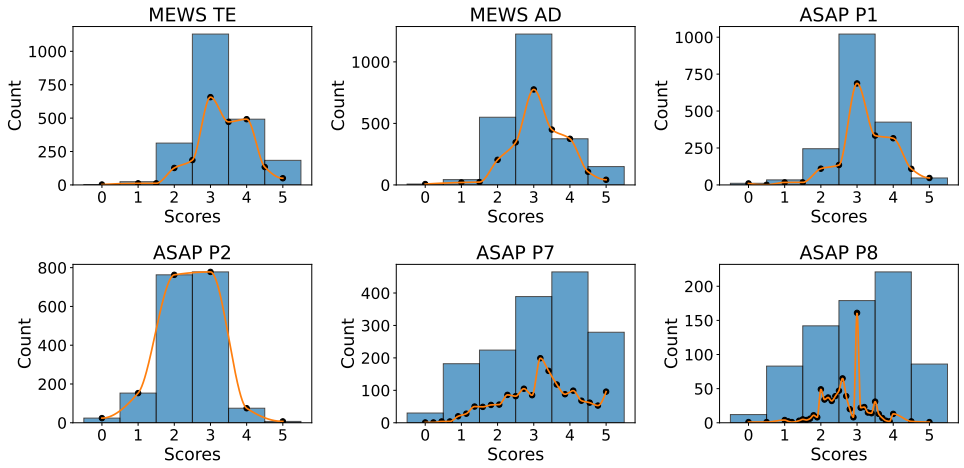



**Figure 2:** Distribution of processed scores for the classification task (blue bars) and z-normalized continuous scores (shifted on the x-axis for visualization purposes) for the regression task (orange line) per dataset and prompt.

datasets into training (60%), validation (20%), and test (20%) sets in a stratified manner, and evaluate the XAI methods on the test sets only.

## 3.2. Attribution-based XAI Methods

Attributions are computed using the `captum` [38] and `ferret` [39] libraries. The overall goal is to identify which input features (e.g. tokens or linguistic features) contribute the most to the model's prediction. We compare different XAI implementations from three attribution methods that differ in how the influence of each input feature on the final prediction is measured: Saliency [40], Integrated Gradients [41], and SHAP [42]. The following section presents their implementation details.

### 3.2.1. Gradient-based attributions

**Saliency** [40] estimates input attributions by taking the gradient of the logits for a particular class with respect to the input text, treating the network locally as a linear function. Given that deep models are

highly non-linear, this can be approximated by a first-order Taylor expansion:

$$S_c(I) \approx w^\top I + b \quad \text{with} \quad w = \left.\frac{\partial S_c}{\partial I}\right|_{I_0},$$

where $S_c(I)$ denotes the logits assigned by the model to class $c$ for an input $I$. The absolute value of each component of the resulting gradient vector $w$ reflects how strongly the corresponding input token influences the class score around $I_0$. Inputs with large gradients are those to which the model's prediction is most responsive.

**Input × Gradient** [43] is a variant of saliency that rescales the local gradient by the value of each input feature. As a result, it amplifies features with larger magnitudes and attenuates those with smaller magnitudes.

**Integrated Gradients** (IG) [41] is derived from two axioms that an attribution method should satisfy: *sensitivity*, which requires that any feature that changes the model's prediction receives a non-zero attribution, and *implementation invariance*, which requires that functionally equivalent models (i.e. predict the same output) yield identical attributions. To satisfy these axioms, IG compares the prediction at the actual input to that at a baseline input $I'$, chosen to represent the absence of informative content (e.g. a zero embedding vector). Rather than relying on the gradient at a single point, IG evaluates how the logits of the class $S_c$ change as the model moves along the straight-line path from the baseline $I'$ to the input $I$. This path is parameterised by $\alpha \in [0, 1]$, where $\alpha = 0$ corresponds to the baseline and $\alpha = 1$ corresponds to the actual input. Intuitively, the method estimates the relevance of an input feature by tracking how the prediction changes as the feature moves from a non-informative baseline (a zero embedding obtained when $\alpha = 0$) to its actual value (reached when $\alpha = 1$).

Formally, the attribution for the $i$-th input dimension is defined as

$$\mathrm{IG}_i(I) = (I_i - I_i') \int_0^1 \frac{\partial S_c(I' + \alpha(I - I'))}{\partial I_i}\, d\alpha,$$

where the integral accumulates the gradients of the class logit along this entire path. The inclusion of the $(I - I')$ term also guarantees *completeness*, meaning that the attributions sum to the difference in logits between the input and the baseline.

We add a local variant of IG obtained by omitting the $(I_i - I_i')$ multiplier in the IG formula, leaving only the path-integrated gradients, and lacking, therefore, the *completeness* property. In the experiments, we refer to the standard formulation as **IG Global**, and its path-integrated gradients-only variant as **IG Local**.

### 3.2.2. SHAP-based attributions

**SHAP** [42] assigns a contribution to every input feature by computing its Shapley value, which quantifies how much that feature changes the model's output when it is added to different subsets of the remaining features. This value is obtained by evaluating the feature's average marginal contribution across all possible subsets of features. Let $F$ denote the full set of input features. The Shapley value for feature $i \in F$ is defined as the weighted average change in model output when $i$ is added to every coalition $S \subseteq F \setminus \{i\}$ of the remaining features:

$$\phi_i \;=\; \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[ f(x_{S \cup \{i\}}) - f(x_S) \right]$$

SHAP expresses these contributions through an additive explanation model, and the explanation for instance $x$ is defined as

$$f(x) \;=\; \phi_0 + \sum_{i=1}^{M} \phi_i,$$

where $\phi_0$ is the expected model output (base value), and $\phi_i$ the contribution of the input feature $i$. SHAP also maintains three properties: *local accuracy* (the sum of the attributions and the baseline reconstruct the prediction), *missingness* (features that are not present do not have any contribution), and *consistency* (an attribution cannot decrease if a feature's effect increases).

Given the combinatorial cost of computing Shapley values at all subsets $S \subseteq F$, approximation strategies are applied. We compute **Partition SHAP** (hierarchical clustering of features), and **Shapley Value Sampling** (Monte Carlo sampling of random feature permutations). We also compute **Gradient SHAP**, an extension of the IG methodology inspired by Shapley principles.

## 3.3. Classification Setting

This section describes the implementation details for the classification task: data preprocessing, model architecture, and evaluation metrics of XAI methods. We train one model per prompt.

### 3.3.1. Score Preprocessing

For the classification task, continuous raw scores are mapped to six ordinal classes (0–5) using a binning procedure that normalizes score ranges across prompts and datasets. For each prompt, six intervals of approximately equal width are computed based on the minimum and maximum scores, preserving the overall score distribution. This reduces the number of possible classes that would result if each unique score were treated as an independent class.

### 3.3.2. Model Architecture

The model is a transformer-based AES classifier built on top of the DistilBERT architecture [44]. Texts are tokenized and encoded by the pretrained encoder *distilbert-base-uncased*, and a linear classification head maps the pooled representation to the six classes. We fine-tune the classification head, while the encoder remains frozen. Training uses cross-entropy loss with class weighting to reduce class imbalance. We evaluate the performance of the model using the weighted F1-score.

### 3.3.3. Evaluation of Token-level XAI Methods

We evaluate token-level XAI methods on faithfulness. Faithfulness evaluates if an explanation truly reflects how the model functions, that is, whether the explanation corresponds to the model's actual prediction [13, 14]. We use three metrics for evaluating faithfulness on the classification task: comprehensiveness, sufficiency, and Kendall's Tau correlation.

**Comprehensiveness** measures the degree to which all relevant features needed to make a prediction were selected by the XAI method [35]. In order to measure this, the model's predicted probability for the original input is compared with its prediction after the relevant features are removed following the formula below. A larger decrease in probability indicates that the identified features were indeed important for the model's decision.

$$\text{comprehensiveness} = m(x_i)_j - m(x_i \setminus r_i)_j$$

This score captures the change in the prediction for class $j$ when the input $x_i$ is evaluated without the feature set $r_i$. A high value indicates that the removed features contributed to the prediction, while a negative value indicates that the prediction increased without them, suggesting that the explanation was not faithful.

Given that identifying relevant features requires selecting a relevancy threshold for continuous attribution values, *comprehensiveness* is computed multiple times by taking the top $k\%$ most relevant tokens and then averaging the resulting scores. This aggregated measure is known as the Area Over the Perturbation Curve (AOPC):

$$\text{AOPC}_{\text{comp}} = \frac{1}{|K|}\left(\sum_{k \in K}\left[m(x_i)_j - m(x_i \setminus r_{i,k})_j\right]\right)$$

Following *ferret's* implementation, $k$ ranges from 10% to 100% (step of 10%) after removing all tokens with negative attributions.

**Sufficiency** measures whether the features identified by the XAI method are sufficient for the model's prediction [35]. In contrast to *comprehensiveness*, the model's predicted probability for the original input is compared with its prediction when only the selected relevant features $r_i$ are retained, following the formula below:

$$\text{sufficiency} = m(x_i)_j - m(r_i)_j$$

A smaller value indicates that the chosen features are sufficient to support the original prediction, while a larger value suggests that the model relies on additional information beyond the selected features. As with *comprehensiveness*, the AOPC is computed by varying $k$ from 10% to 100%:

$$\text{AOPC}_{\text{suff}} = \frac{1}{|K|}\left(\sum_{k \in K}\left[m(x_i)_j - m(r_{i,k})_j\right]\right)$$

**Kendall's Tau correlation** is computed between the relevance scores assigned by the XAI method and the *leave-one-out* (LOO) importance scores, which measure the change in prediction when each token is removed [36]. The ranking generated by the LOO procedure assumes that token contributions are independent of each other [13]. Given two rankings $R_{\text{exp}}$ and $R_{\text{LOO}}$, Kendall's $\tau$ is defined as:

$$\tau = \frac{C - D}{\frac{1}{2}n(n-1)},$$

where $C$ is the number of concordant pairs, $D$ is the number of discordant pairs, and $n$ is the number of tokens. A value of $\tau$ closer to 1 indicates a stronger agreement between the explanation and the LOO importance ranking.

## 3.4. Regression Setting

This section describes the implementation details for the regression task: data preprocessing, model architecture, and evaluation metrics of XAI methods. We train one model per prompt.

### 3.4.1. Score Preprocessing

For the regression task, scores are z-normalized into the range [0,1] for training stability. Predictions are transformed back to the original scoring scales for evaluation.

### 3.4.2. Model Architecture

We use a shallow neural regressor that mirrors the overall structure of the classification pipeline. For the regression task, however, we use hand-crafted linguistic features rather than transformer-based representations. Each essay is represented through a 220-dimensional feature vector extracted using the linguistic feature extractor from Lohmann et al. [45], which includes a set of indicators related to writing quality: vocabulary, syntactic structure, cohesion, spelling errors, and various frequency- and length-derived measures and ratios. A full description of the features can be found in Lohmann et al. [45].

The model consists of two fully connected layers with a ReLU non-linear activation function and dropout regularization, mapping the linguistic feature vector to a continuous score bounded to 0 and 1.

Training is performed on z-normalized targets to stabilize optimization, and predictions are transformed back to the original scoring scale for evaluation. We evaluate the performance of the model using Quadratic Weighted Kappa (QWK) [46] given that it accounts for the magnitude of the difference between predictions and true scores.

#### 3.4.3. Evaluation of Feature-level XAI Methods

The feature-level XAI methods are evaluated on two dimensions: faithfulness and robustness. Robustness evaluates if the XAI method provides similar explanations to similar input samples and, similarly, if irrelevant slight perturbations to the input sample significantly change the explanation [13, 14]. We use infidelity for measuring faithfulness and sensitivity for measuring robustness.

**Infidelity** measures the extent to which an explanation aligns with the model's actual output changes under meaningful input perturbations [37]. Following the completeness axiom in attribution methods [41], which requires that the sum of the feature importances at a given input equals the difference between the model's prediction at that input and its prediction at a chosen baseline, the notion of infidelity extends this principle to arbitrary perturbations of the input.

Given a model $f$, an XAI method $\Phi(f, x)$ that assigns an attribution value to each input feature, and a meaningful perturbation vector $I$, the infidelity score of the explanation is defined as:

$$\text{INFD}(\Phi, f, x) \;=\; \mathbb{E}_I\Big[\big(I^\top \Phi(f, x) \;-\; \big(f(x) - f(x - I)\big)\big)^2\Big]$$

The infidelity score is the expected mean squared error between the change *predicted* by the explanation (the dot product between $I$ and $\Phi(f, x)$) and the *actual* change in the model output when the input is significantly perturbed. A lower score therefore indicates that the explanation is consistent with the model's behaviour under the chosen perturbations.

In our implementation, the perturbation vector $I$ is sampled from a Gaussian distribution with a varying standard deviation $\sigma \in \{0.001, 0.005, 0.01, 0.05, 0.1\}$. For a given input $x$, we construct the perturbed input as $x - I$. This perturbation allows us to evaluate if the XAI method can correctly predict the model's behaviour under significant input variations.

**Sensitivity** or sensitivity-max measures the largest possible change in the explanation when the input is slightly perturbed [37]. The main intuition behind sensitivity is that small, insignificant perturbations should not cause large changes in the explanation. High variability under such minor perturbations indicates an unstable and fragile explanation method, which in turn may behave unpredictably when the input is altered in ways that are not semantically meaningful.

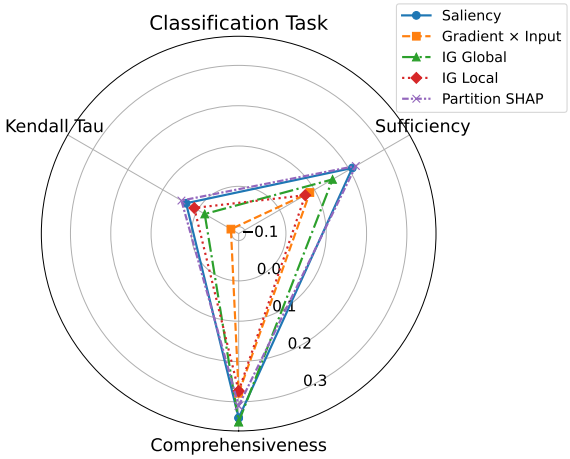
Formally, it is defined as:

$$\text{SENSMAX}(\Phi, f, x, r) = \max_{\|y - x\| \le r} \|\Phi(f, y) - \Phi(f, x)\|,$$

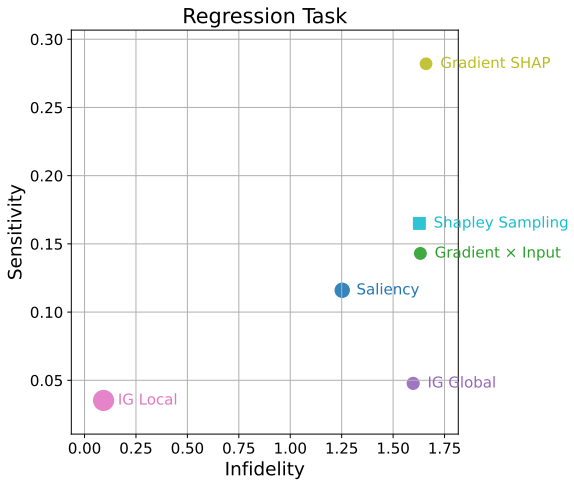
where the constraint $\|y - x\| \le r$ denotes all inputs $y$ lying inside an $L_\infty$ ball of radius $r$, meaning that no individual feature of the input is perturbed by more than $r$. Since there are infinitely many possible perturbations, the exact maximum cannot be computed directly. Therefore, sensitivity-max is approximated using Monte Carlo sampling by drawing multiple perturbed inputs from the $L_\infty$ ball and taking the largest observed change in the explanation. A lower sensitivity-max value indicates that the explanation is stable under small input perturbations, whereas higher values signal fragility.

## 4. Results and Discussion

This section presents the results for the two settings introduced above. First, it summarizes the evaluation of the XAI methods, which are tested on two independent test folds to account for data




(a) Evaluation of faithfulness (comprehensiveness, sufficiency, and Kendall's Tau correlation). Sufficiency scores, where lower is better, are inverted for visualization purposes. Higher scores are better.




(b) Evaluation of faithfulness (infidelity) and robustness (sensitivity). Infidelity scores are rescaled for visualization purposes. Lower scores are better.

**Figure 3:** Comparison of explanation quality for classification and regression tasks.

**Table 2**
Performance of models per prompt on the MEWS and ASAP datasets. Regression is evaluated using QWK, and classification using weighted F1-score. Results shown as mean (std) across 5 folds.

| | MEWS | | ASAP | | | |
|---|---|---|---|---|---|---|
| Task | TE | AD | P1 | P2 | P7 | P8 |
| Regression | 0.59 (0.09) | 0.61 (0.04) | 0.67 (0.05) | 0.62 (0.03) | 0.66 (0.07) | 0.46 (0.06) |
| Classification | 0.58 (0.02) | 0.56 (0.04) | 0.63 (0.05) | 0.58 (0.03) | 0.49 (0.04) | 0.41 (0.03) |

**Table 3**
AOPC Comprehensiveness scores (↗) per prompt and dataset. Values shown as mean (std). Best performing method in bold.

| | MEWS | | ASAP | | | | Mean |
|---|---|---|---|---|---|---|---|
| Method | TE | AD | Prompt 1 | Prompt 2 | Prompt 7 | Prompt 8 | (std) |
| Saliency | 0.35 (0.07) | 0.37 (0.06) | 0.34 (0.02) | **0.36 (0.07)** | 0.38 (0.08) | 0.24 (0.03) | 0.34 (0.05) |
| Input × Gradient | 0.34 (0.06) | 0.34 (0.06) | 0.22 (0.02) | 0.31 (0.09) | 0.23 (0.08) | 0.23 (0.03) | 0.28 (0.06) |
| IG Local | 0.32 (0.07) | 0.33 (0.07) | 0.22 (0.02) | 0.33 (0.08) | 0.21 (0.08) | 0.23 (0.04) | 0.27 (0.06) |
| IG Global | **0.38 (0.08)** | **0.39 (0.04)** | **0.34 (0.01)** | 0.36 (0.09) | **0.39 (0.09)** | **0.24 (0.02)** | **0.35 (0.06)** |
| Partition SHAP | 0.38 (0.09) | 0.30 (0.07) | 0.32 (0.03) | 0.32 (0.06) | 0.32 (0.04) | 0.20 (0.01) | 0.31 (0.06) |
| Mean (std) | 0.35 (0.03) | **0.35 (0.04)** | 0.29 (0.06) | 0.34 (0.02) | 0.31 (0.08) | 0.23 (0.02) | |

partition variability, along with the overall performance of the models, assessed using a 5-fold cross-validation procedure. Then, detailed observations per task are provided, highlighting differences between XAI methods and datasets.

Table 2 depicts the performance of the classification and regression model for each task per prompt and dataset. The wide performance range, varying by almost 0.2 across tasks, provides a meaningful basis for comparing XAI methods across baseline models that differ in how well they capture the nuances and task-specific variations of each prompt (cf. [2, 45]).

Figure 3a summarizes the results of the evaluation of XAI methods for the classification task. The most faithful XAI method overall is Saliency. It consistently ranks among the best in comprehensiveness, sufficiency, and Kendall's Tau correlation, while also exhibiting low variance across datasets, indicating greater stability and reliability across different prompts.

Figure 3b summarizes the overall results for the regression task. In this case, IG Local performs

**Table 4**
AOPC Sufficiency scores (↘) per prompt and dataset. Values shown as mean (std). Best performing method in bold.

| | MEWS | | ASAP | | | | Mean |
|---|---|---|---|---|---|---|---|
| Method | TE | AD | Prompt 1 | Prompt 2 | Prompt 7 | Prompt 8 | (std) |
| Saliency | 0.37 (0.06) | **0.35 (0.04)** | **0.27 (0.01)** | **0.32 (0.07)** | **0.23 (0.05)** | 0.21 (0.02) | 0.29 (0.07) |
| Input × Gradient | 0.45 (0.07) | 0.45 (0.04) | 0.41 (0.02) | 0.45 (0.09) | 0.44 (0.08) | 0.28 (0.02) | 0.41 (0.07) |
| IG Local | 0.47 (0.08) | 0.47 (0.02) | 0.43 (0.05) | 0.46 (0.09) | 0.44 (0.09) | 0.28 (0.02) | 0.43 (0.07) |
| IG Global | 0.42 (0.08) | 0.39 (0.07) | 0.31 (0.01) | 0.41 (0.08) | 0.29 (0.05) | 0.27 (0.03) | 0.35 (0.07) |
| Partition SHAP | **0.34 (0.07)** | 0.35 (0.05) | 0.30 (0.02) | 0.30 (0.07) | 0.23 (0.09) | **0.16 (0.03)** | **0.28 (0.07)** |
| Mean (std) | 0.41 (0.05) | 0.40 (0.06) | 0.34 (0.07) | 0.39 (0.07) | 0.33 (0.11) | **0.24 (0.05)** | |

**Table 5**
Kendall's Tau correlation scores (↗) per prompt and dataset. Values shown as mean (std). Strongest correlations in bold.

| | MEWS | | ASAP | | | | Mean |
|---|---|---|---|---|---|---|---|
| Method | TE | AD | Prompt 1 | Prompt 2 | Prompt 7 | Prompt 8 | (std) |
| Saliency | 0.01 (0.03) | 0.02 (0.01) | 0.02 (0.01) | 0.06 (0.01) | 0.02 (0.02) | **0.07 (0.01)** | 0.03 (0.03) |
| Input × Gradient | **-0.12 (0.03)** | **-0.09 (0.02)** | **-0.11 (0.01)** | **-0.08 (0.04)** | **-0.11 (0.00)** | -0.06 (0.00) | **-0.10 (0.02)** |
| IG Local | 0.01 (0.00) | 0.01 (0.00) | 0.01 (0.00) | 0.00 (0.00) | 0.02 (0.00) | 0.01 (0.00) | 0.01 (0.01) |
| IG Global | 0.00 (0.02) | -0.03 (0.01) | -0.06 (0.00) | -0.04 (0.02) | 0.05 (0.02) | -0.04 (0.00) | -0.02 (0.04) |
| Partition SHAP | 0.03 (0.01) | 0.03 (0.01) | 0.02 (0.02) | 0.05 (0.00) | 0.10 (0.01) | 0.05 (0.00) | 0.05 (0.03) |
| Mean (std) | -0.01 (0.06) | -0.01 (0.05) | **-0.02 (0.06)** | 0.00 (0.06) | 0.02 (0.08) | 0.01 (0.06) | |

**Table 6**
Infidelity scores (↘) per prompt and dataset. Values are reported as mean ($\times 10^{-4}$) and standard deviation ($\times 10^{-5}$). Best performing method in bold.

| | MEWS | | ASAP | | | | Mean |
|---|---|---|---|---|---|---|---|
| Explainer | TE | AD | Prompt 1 | Prompt 2 | Prompt 7 | Prompt 8 | (std) |
| Saliency | 1.44 (1.57) | 1.36 (1.36) | 1.32 (0.16) | 0.89 (0.29) | 1.59 (1.18) | 1.13 (0.09) | 1.29 (0.23) |
| Input × Gradient | 1.72 (0.69) | 2.13 (2.20) | 1.58 (1.36) | 1.10 (1.09) | 1.72 (1.34) | 1.15 (0.01) | 1.57 (0.36) |
| IG Local | **0.10 (0.02)** | **0.10 (0.06)** | **0.10 (0.07)** | **0.07 (0.02)** | **0.09 (0.02)** | **0.08 (0.02)** | **0.09 (0.01)** |
| IG Global | 1.70 (0.07) | 2.06 (2.24) | 1.53 (1.31) | 1.10 (1.27) | 1.67 (1.26) | 1.13 (1.01) | 1.53 (0.35) |
| Gradient SHAP | 1.77 (0.54) | 2.22 (2.11) | 1.54 (1.45) | 1.11 (1.32) | 1.75 (1.24) | 1.14 (0.73) | 1.59 (0.38) |
| Shapley Sampling | 1.75 (0.57) | 2.14 (2.78) | 1.54 (1.21) | 1.08 (0.90) | 1.68 (1.00) | 1.22 (0.58) | 1.57 (0.33) |
| Mean (std) | 1.41 (0.63) | 1.67 (0.77) | 1.27 (0.47) | **1.06 (0.35)** | 1.42 (0.52) | 1.14 (0.43) | |

the best for both infidelity and sensitivity, which suggests that aggregating gradients along the path between the baseline (non-informative input) up to the current input yields more stable and consistent attributions, despite not satisfying the completeness property.

Considering both datasets, which differ in whether English is the writer's first or second language, XAI methods show worse comprehensiveness, but better sufficiency and infidelity on the ASAP dataset (English L1), implying that the relevance of tokens in this dataset is more evenly spread and not highly focalised on a few input features. Kendall's Tau correlation is not an informative metric for the classification task, as its values remain close to 0 for all XAI methods, indicating no meaningful (positive or negative) correlation. The lack of correlation could be due to the main assumption of the LOO procedure when generating the ranking, which, in this case, translates to transformer-based models being stable against individual token perturbations.

## 4.1. Classification Task

The three evaluation metrics applied in the classification task measure faithfulness, and they reveal differences across datasets and XAI methods. Regarding comprehensiveness (Table 3), there is a clear distinction between the two datasets when comparing the mean of the XAI methods across prompts:

**Table 7**
Sensitivity scores (↘) across datasets and prompts. Values shown as means. All standard deviations are ≤ 0.01. Best performing method in bold.

| | MEWS | | ASAP | | | | Mean |
|---|---|---|---|---|---|---|---|
| Explainer | TE | AD | Prompt 1 | Prompt 2 | Prompt 7 | Prompt 8 | (std) |
| Saliency | 0.12 | 0.10 | 0.12 | 0.13 | 0.11 | 0.15 | 0.12 (0.02) |
| Input × Gradient | 0.14 | 0.14 | 0.14 | 0.15 | 0.13 | 0.16 | 0.14 (0.01) |
| IG Local | **0.04** | **0.03** | **0.04** | **0.04** | **0.04** | **0.06** | **0.04 (0.01)** |
| IG Global | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | **0.06** | 0.05 (0.00) |
| Gradient SHAP | 0.29 | 0.28 | 0.27 | 0.28 | 0.27 | 0.32 | 0.29 (0.02) |
| Shapley Sampling | 0.17 | 0.16 | 0.16 | 0.17 | 0.16 | 0.12 | 0.16 (0.02) |
| Mean (std) | 0.14 (0.09) | 0.13 (0.09) | **0.13 (0.08)** | 0.14 (0.08) | **0.13 (0.08)** | 0.15 (0.09) | |

comprehensiveness is 0.6 higher in MEWS than in ASAP. This indicates that, across methods, removing the most important features results in a larger drop in prediction confidence for MEWS. We hypothesize that the model is able to capture more salient tokens in the MEWS dataset, whereas in the ASAP dataset it distributes the relevancy of tokens in a more even manner, making the removal of individual tokens less discriminatory.

Considering each XAI method, IG Global is the best performing method, closely followed by Saliency, which shows almost the same performance across prompts with lower standard deviation. In contrast, IG Local and Input × Gradient perform 6 and 8 points worse than their variants, respectively. This suggests that scaling gradients by the input values tends to amplify irrelevant variations, and removing the $(I - I')$ term in the Integrated Gradients formulation reduces the ability to properly track feature contributions as the model goes from the baseline input (i.e. non-informative embedding) to the actual input.

There is a marked difference between the datasets regarding sufficiency (Table 4) when averaging across methods, and this trend is opposite to the trend observed when evaluating comprehensiveness. In this case, MEWS has a mean score of 0.40, whereas ASAP has a lower, thus better, score of 0.32. This suggests that although both metrics are complementary in nature, the model's confidence does not necessarily follow a linear relationship between them. In MEWS, the methods are able to detect relevant tokens, leading to better performance in comprehensiveness. However, keeping only a few top-ranked tokens is not sufficient to maintain the model's confidence, which leads to weaker sufficiency performance. In ASAP, on the other hand, the methods are able to capture a more distributed predictive signal, resulting in lower comprehensiveness, but stronger sufficiency, since the model is able to maintain a similar prediction even when based on a small subset of tokens compared to when using the entire essay. When considering individual methods, Saliency and Partition SHAP perform the best.

Regarding Kendall's Tau correlation (Table 5), the results show a consistently weak alignment between the explanation rankings and the model's perturbation-based importance order. Absolute values are not higher than 0.12, and two methods (Input × Gradient and IG Global) even show negative correlations across prompts. This suggests that, regardless of the overall performance obtained by the model as seen in Table 2, the ranked attributions provided by the XAI method do not align with the ranking of tokens produced by the leave-one-out procedure. This result can be explained by the fact that transformer-based models make use of contextualized input embeddings, and removing a single token from a long sequence does not produce a linear effect on the prediction, which is the main assumption in this metric. Despite the overall weakness of the correlations, Saliency and Partition SHAP are the only ones achieving positive and relatively higher correlations compared to the other methods.

## 4.2. Regression Task

The regression setting considers the evaluation of faithfulness (infidelity) and robustness (sensitivity). Regarding infidelity (Table 6), all methods show comparably small values, suggesting that they properly predict the change in prediction when the input text is significantly perturbed. IG Local, nevertheless,

shows the best performance by a relatively wide margin, indicating that aggregating gradients along a path yields attributions that better approximate the model's response to perturbations. When averaging across methods, ASAP exhibits a lower infidelity score (1.22) than MEWS (1.54), with lower variance, which suggests that the XAI methods are more capable of predicting the change in prediction when the perturbed input contains a more evenly distributed predictive signal.

Regarding sensitivity (Table 7), IG Local achieves the best performance, closely followed by its variant IG Global, indicating that the Integrated Gradients XAI method is the most stable under small input perturbations. When averaging across methods, both datasets obtain similar sensitivity scores, suggesting that the explanations are neither susceptible to dataset characteristics, nor to the overall performance of the models.

## 5. Conclusion

In this study, we evaluated a set of commonly used XAI methods, i.e., token- and feature-level attribution methods, for AES models regarding their faithfulness and robustness. We evaluated the explanations on two settings: transformer-based AES models in a classification task, and linguistic feature-based models in a regression task. Training and evaluation was performed on two writing datasets that differ in the writer's first language (English L1 or L2).

Our main findings indicate that Saliency [40] performs better than the other XAI methods considering faithfulness metrics in the classification setting as it is more stable across prompts and datasets, while having high performance overall. Integrated Gradients [41], for the regression setting, performs the best for both faithfulness and robustness. The difference between comprehensiveness and sufficiency across datasets suggests that even though these metrics are related, they should be considered in parallel as they highlight complementary aspects of feature relevancy that differ across datasets. Lastly, Kendall's Tau correlation shows no correlation between the attributions assigned by the XAI methods and the ranking produced when leaving one token out at a time, indicating that models are generally stable against these perturbations.

Our results shed light on the reliability of XAI methods applied to AES models, providing insights on the faithfulness and robustness of methods already in use and found, for instance, in [11] and [12]. As machine learning models become part of the learning process in classroom settings, the need for faithful and robust XAI methods is an essential step toward meeting not only legislations already imposed by the EU AI Act [1], but also deciding which methods should be further tested considering qualitative aspects, such as their complexity or degree of interpretability with end users (e.g. students and teachers). Improving the performance of the models, alongside their degree of interpretability, will help in building trustworthy models. This alignment between technical performance, regulatory compliance, and faithful and robust interpretability methods is critical for the responsible integration of AES systems.

## Limitations

Obtained results cannot be generalized to different text genres and model architectures, given that our approach tests on just one model per each task setting (classification and regression) on argumentative/-narrative writings. Results may vary depending on the genre, size of the model, pretrained encoder, and set of input features used.

## Ethical Consideration

This study uses publicly available datasets that adhere to established ethical guidelines. No additional bias or fairness risks are introduced through the evaluation of XAI methods explored in this study. We highlight the need for more explanations that are indeed faithful and robust, and we hope that our findings contribute to developing more transparent and reliable AES systems.

## Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to: Grammar and spelling check. After using these tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

## References


[1] E. Parliament, E. Council, Regulation (eu) 2024/1689 laying down harmonised rules on artificial intelligence, 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng, official Journal L 2024/1689 (12 July 2024).

[2] D. Ramesh, S. K. Sanampudi, An automated essay scoring systems: A systematic literature review, Artificial Intelligence Review 55 (2022) 2495–2527. doi:10.1007/s10462-021-10068-2.

[3] W. Xu, R. Mahmud, W. Lam Hoo, A Systematic Literature Review: Are Automated Essay Scoring Systems Competent in Real-Life Education Scenarios?, IEEE Access 12 (2024) 77639–77657. doi:10.1109/ACCESS.2024.3399163.

[4] J. Sun, T. Song, W. Peng, J. Song, A survey of automated essay scoring: Challenges, advances, and future, Neurocomputing 650 (2025) 130916. doi:10.1016/j.neucom.2025.130916.

[5] T. Mizumoto, H. Ouchi, Y. Isobe, P. Reisert, R. Nagata, S. Sekine, K. Inui, Analytic Score Prediction and Justification Identification in Automated Short Answer Scoring, in: H. Yannakoudakis, E. Kochmar, C. Leacock, N. Madnani, I. Pilán, T. Zesch (Eds.), Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, Florence, Italy, 2019, pp. 316–325. doi:10.18653/v1/W19-4433.

[6] S. Parekh, Y. K. Singla, C. Chen, J. J. Li, R. R. Shah, My Teacher Thinks The World Is Flat! Interpreting Automatic Essay Scoring Mechanism, 2020. doi:10.48550/arXiv.2012.13872.

[7] W. Yupei, H. Renfen, A prompt-independent and interpretable automated essay scoring method for Chinese second language writing, in: S. Li, M. Sun, Y. Liu, H. Wu, K. Liu, W. Che, S. He, G. Rao (Eds.), Proceedings of the 20th Chinese National Conference on Computational Linguistics, Chinese Information Processing Society of China, Huhhot, China, 2021, pp. 1202–1217. URL: https://aclanthology.org/2021.ccl-1.107/.

[8] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, P. Sen, A Survey of the State of Explainable AI for Natural Language Processing, in: K.-F. Wong, K. Knight, H. Wu (Eds.), Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2020, pp. 447–459. doi:10.18653/v1/2020.aacl-main.46.

[9] A. Madsen, S. Reddy, S. Chandar, Post-hoc Interpretability for Neural NLP: A Survey, ACM Comput. Surv. 55 (2022) 155:1–155:42. doi:10.1145/3546577.

[10] H. Chefer, S. Gur, L. Wolf, Transformer Interpretability Beyond Attention Visualization, 2021. doi:10.48550/arXiv.2012.09838. arXiv:2012.09838.

[11] V. Kumar, D. Boulanger, Explainable automated essay scoring: Deep learning really has pedagogical value, Frontiers in Education 5 (2020). doi:10.3389/feduc.2020.572367.

[12] Y. Plasencia-Calaña, Operationalizing automated essay scoring: A human-aware approach, 2025. URL: https://arxiv.org/abs/2506.21603. arXiv:2506.21603.

[13] A. Jacovi, Y. Goldberg, Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4198–4205. doi:10.18653/v1/2020.acl-main.386.

[14] S. Sithakoul, S. Meftah, C. Feutry, BEExAI: Benchmark to Evaluate Explainable AI, in: L. Longo, S. Lapuschkin, C. Seifert (Eds.), Explainable Artificial Intelligence, Springer Nature Switzerland, Cham, 2024, pp. 445–468. doi:10.1007/978-3-031-63787-2_23.

[15] P. Atanasova, J. G. Simonsen, C. Lioma, I. Augenstein, A diagnostic study of explainability techniques for text classification, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 3256–3274. doi:10.18653/v1/2020.emnlp-main.263.

[16] A. Madsen, S. Chandar, S. Reddy, Are self-explanations from large language models faithful?, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 295–337. doi:10.18653/v1/2024.findings-acl.19.

[17] S. D. Keller, J. Fleckenstein, M. Krüger, O. Köller, A. A. Rupp, English writing skills of students in upper secondary education: Results from an empirical study in switzerland and germany, Journal of Second Language Writing 48 (2020) 100700. doi:https://doi.org/10.1016/j.jslw.2019.100700.

[18] E. Hall, M. Seyam, D. Dunlap, Exploring explainability and transparency in automated essay scoring systems: A user-centered evaluation, in: P. Zaphiris, A. Ioannou (Eds.), Learning and Collaboration Technologies, Springer Nature Switzerland, Cham, 2024, pp. 266–282.

[19] Y. Feldman-Maggor, M. Cukurova, C. Kent, G. Alexandron, The impact of explainable ai on teachers' trust and acceptance of ai edtech recommendations: The power of domain-specific explanations, International Journal of Artificial Intelligence in Education (2025). doi:10.1007/s40593-025-00486-6.

[20] M. Bearman, R. Ajjawi, Learning to work with the black box: Pedagogy for a world with artificial intelligence, British Journal of Educational Technology 54 (2023) 1160–1173. doi:https://doi.org/10.1111/bjet.13337.

[21] H. Khosravi, S. B. Shum, G. Chen, C. Conati, Y.-S. Tsai, J. Kay, S. Knight, R. Martinez-Maldonado, S. Sadiq, D. Gašević, Explainable artificial intelligence in education, Computers and Education: Artificial Intelligence 3 (2022) 100074. doi:https://doi.org/10.1016/j.caeai.2022.100074.

[22] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206–215. doi:10.1038/s42256-019-0048-x.

[23] F. Urrutia, C. Buc, R. Araya, V. Barriere, Unsupervised automatic short answer grading and essay scoring: A weakly supervised explainable approach, in: E. Kochmar, B. Alhafni, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, Z. Yuan (Eds.), Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 38–54. doi:10.18653/v1/2025.bea-1.4.

[24] Y. Kumar, S. Aggarwal, D. Mahata, R. R. Shah, P. Kumaraguru, R. Zimmermann, Get it scored using autosas — an automated system for scoring short answers, AAAI'19/IAAI'19/EAAI'19, AAAI Press, 2019. doi:10.1609/aaai.v33i01.33019662.

[25] S. Eltanbouly, S. Albatarni, T. Elsayed, TRATES: Trait-specific rubric-assisted cross-prompt essay scoring, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 20528–20543. doi:10.18653/v1/2025.findings-acl.1054.

[26] C. Molnar, Interpretable Machine Learning, 3 ed., 2025. URL: https://christophm.github.io/interpretable-ml-book.

[27] S. Li, V. Ng, Conundrums in Cross-Prompt Automated Essay Scoring: Making Sense of the State of the Art, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 7661–7681. doi:10.18653/v1/2024.acl-long.414.

[28] F. Dong, Y. Zhang, J. Yang, Attention-based recurrent convolutional neural network for automatic essay scoring, in: R. Levy, L. Specia (Eds.), Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 153–162. doi:10.18653/v1/K17-1017.

[29] R. Yang, J. Cao, Z. Wen, Y. Wu, X. He, Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking, in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 1560–1569. doi:10.18653/v1/2020.findings-emnlp.141.

[30] D. Alikaniotis, H. Yannakoudakis, M. Rei, Automatic text scoring using neural networks, in: K. Erk, N. A. Smith (Eds.), Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 715–725. doi:10.18653/v1/P16-1068.

[31] A. Mizumoto, M. Eguchi, Exploring the potential of using an AI language model for automated essay scoring, Research Methods in Applied Linguistics 2 (2023) 100050. doi:10.1016/j.rmal.2023.100050.

[32] G.-G. Lee, E. Latif, X. Wu, N. Liu, X. Zhai, Applying large language models and chain-of-thought for automatic scoring, Computers and Education: Artificial Intelligence 6 (2024) 100213. doi:10.1016/j.caeai.2024.100213.

[33] T. Lanham, A. Chen, A. Radhakrishnan, B. Steiner, C. Denison, D. Hernandez, D. Li, E. Durmus, E. Hubinger, J. Kernion, K. Lukošiūtė, K. Nguyen, N. Cheng, N. Joseph, N. Schiefer, O. Rausch, R. Larson, S. McCandlish, S. Kundu, S. Kadavath, S. Yang, T. Henighan, T. Maxwell, T. Telleen-Lawton, T. Hume, Z. Hatfield-Dodds, J. Kaplan, J. Brauner, S. R. Bowman, E. Perez, Measuring faithfulness in chain-of-thought reasoning, 2023. URL: https://arxiv.org/abs/2307.13702. arXiv:2307.13702.

[34] M. Turpin, J. Michael, E. Perez, S. R. Bowman, Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Curran Associates Inc., Red Hook, NY, USA, 2023.

[35] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, B. C. Wallace, ERASER: A benchmark to evaluate rationalized NLP models, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4443–4458. doi:10.18653/v1/2020.acl-main.408.

[36] S. Jain, B. C. Wallace, Attention is not Explanation, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 3543–3556. doi:10.18653/v1/N19-1357.

[37] C.-K. Yeh, C.-Y. Hsieh, A. S. Suggala, D. I. Inouye, P. Ravikumar, On the (In)fidelity and Sensitivity for Explanations, 2019. doi:10.48550/arXiv.1901.09392. arXiv:1901.09392.

[38] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, O. Reblitz-Richardson, Captum: A unified and generic model interpretability library for PyTorch, 2020. doi:10.48550/arXiv.2009.07896. arXiv:2009.07896.

[39] G. Attanasio, E. Pastor, C. Di Bonaventura, D. Nozza, Ferret: A Framework for Benchmarking Explainers on Transformers, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 256–266. doi:10.18653/v1/2023.eacl-demo.29.

[40] K. Simonyan, A. Vedaldi, A. Zisserman, Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, 2014. doi:10.48550/arXiv.1312.6034. arXiv:1312.6034.

[41] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: D. Precup, Y. W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 3319–3328. URL: https://proceedings.mlr.press/v70/sundararajan17a.html.

[42] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran

Associates Inc., Red Hook, NY, USA, 2017, p. 4768–4777.

[43] A. Shrikumar, P. Greenside, A. Shcherbina, A. Kundaje, Not Just a Black Box: Learning Important Features Through Propagating Activation Differences, 2017. doi:10.48550/arXiv.1605.01713. arXiv:1605.01713.

[44] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. URL: https://arxiv.org/abs/1910.01108. arXiv:1910.01108.

[45] J. F. Lohmann, F. Junge, J. Möller, J. Fleckenstein, R. Trüb, S. Keller, T. Jansen, A. Horbach, Neural networks or linguistic features? - comparing different machine-learning approaches for automated assessment of text quality traits among l1- and l2-learners' argumentative essays, International Journal of Artificial Intelligence in Education (2024). doi:10.1007/s40593-024-00426-w.

[46] J. Cohen, Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit, Psychological bulletin 70 (1968) 213—220. doi:10.1037/h0026256.