

BIAS-AWARE LARGE LANGUAGE MODELS FOR EQUITABLE EDUCATIONAL FEEDBACK AND ASSESSMENT

QianYi Fang

University of Chichester, Chichester PO19 6PE, United Kingdom.

Abstract: Pre-trained large language models (LLMs) are increasingly used to score student writing and to produce trait-level feedback, yet their sensitivity to construct-irrelevant linguistic variation can disadvantage students who write in African American English (AAE) or who are learning English as a second language (L2). This paper presents a bias-aware scoring framework that audits and mitigates such variation without access to student demographics. Counterfactual versions of real student essays are generated by an AAE morphosyntactic rule system and by an L2 error generator whose error-category rates are estimated from the JFLEG learner corpus and whose misspellings are drawn from TOEFL-Spell. Paired representations from a frozen RoBERTa encoder define an invariance operator, and Construct-Aligned Invariance Regularization (CAIR) applies it only to score dimensions that the rubric defines as independent of language conventions, which yields a closed-form estimator. On 5,569 ASAP essays from five prompts and five random splits, CAIR reduces the mean standardized counterfactual shift of holistic content scores by 85% and the counterfactual score-flip rate by about 40%, with no significant change in quadratic weighted kappa (0.732 vs. 0.733). For trait feedback, CAIR removes the spill-over of learner errors into Ideas, Organization and Style scores while preserving the legitimate sensitivity of the Conventions score, which task-agnostic debiasing methods erase. Pairs for 2% of the training essays already halve the shift, and invariance learned from one variety transfers to the other.

Keywords: Automated essay scoring; Large language models; Algorithmic fairness; Counterfactual fairness; Dialect bias; English learners; Formative feedback

1 INTRODUCTION

Automated scoring of constructed responses has moved from research prototypes to high-volume use in classrooms and large-scale testing, a transition that the Automated Student Assessment Prize (ASAP) competition helped to accelerate [1]. Neural scorers first replaced hand-crafted feature sets with learned representations [2-3], and pre-trained transformer language models such as BERT and RoBERTa [4-5], together with much larger generative models [6], now form the default backbone for holistic scoring and trait-level feedback [7-9]. Representations learned from web-scale text encode far more than the construct a rubric intends to measure, and their use in education raises a question that agreement statistics do not answer: does a student's score depend on how the student's variety of English differs from the variety that dominates the pre-training data?

Evaluation frameworks for automated scoring have long treated consequences for subgroups as part of the validity argument [10]. Operational studies found that machine and human scores diverge for groups defined by ethnicity and country of origin [11], and the educational NLP community has proposed open tooling and multi-dimensional definitions of fairness for scoring engines [12-13], as well as broader notions of equity for language technologies in schools [14]. Reviews of algorithmic bias in education trace how disparities arise along the data and modeling pipeline [15], and group-level biases have been measured for BERT-based assessment models [16]. A recurring concern is that language technology reproduces social hierarchies among language varieties [17]. African American English (AAE) in particular is handled poorly by standard NLP systems [18-19], and content scoring models react to surface properties of their input in ways that human raters would not [20].

Two obstacles limit the use of these findings in practice. First, demographic labels are often unavailable to system developers or legally restricted, so group-level audits cannot be run where they are most needed. Second, group-level disparities mix construct-relevant and construct-irrelevant sources: a lower conventions score for English learners may correctly reflect the rubric, whereas a lower content score caused by missing articles does not. Counterfactual reasoning separates the two. We generate paired versions of real student essays that differ only in linguistic features associated with AAE or with learner English, using error statistics mined from learner corpora in the tradition of artificial error generation [21-23]. A scoring function is counterfactually fair with respect to such an intervention when its output does not change [24-25]. Existing mitigation techniques, however, enforce invariance indiscriminately: counterfactual data augmentation [26], adversarial training and projection-based removal of protected information act on every output of the model [27-29]. In writing assessment this removes legitimate information, because rubric dimensions such as conventions are defined with respect to Standard English.

This paper proposes Construct-Aligned Invariance Regularization (CAIR), a bias-aware scoring method for LLM representations that enforces invariance only where the rubric requires it. The contributions are as follows.

1) A demographic-free counterfactual audit for LLM-based essay scoring, built on AAE morphosyntactic rules and on L2 errors whose category rates and misspellings are estimated from real learner corpora. In embedding space, the synthetic

L2 displacement is closely aligned with the displacement between real learner sentences and their corrections (cosine 0.86).

2) A closed-form estimator that couples ridge scoring heads with a trait-selective counterfactual consistency penalty, together with an analysis that relates it to counterfactual data augmentation.

3) An evaluation on five ASAP prompts with holistic and trait targets, showing that CAIR cuts counterfactual score shifts by 85% and score flips by about 40% without a significant loss in agreement, and that it preserves the construct-relevant conventions feedback that task-agnostic debiasing methods remove.

2 RELATED WORK

2.1 LLM-Based Essay Scoring and Feedback

ASAP remains the most widely used public benchmark for automated essay scoring (AES) [1]. Taghipour and Ng showed that recurrent networks over word embeddings outperform feature-engineered systems [2], and Dong et al. added attention over sentence representations [3]. After BERT and RoBERTa appeared [4-5], Mayfield and Black examined when fine-tuning pays off for essay scoring relative to lighter models [7], and Yang et al. combined regression and ranking objectives for fine-tuned transformers [8]. Trait scoring, the basis of analytic feedback, has been approached with multi-task models that exploit relations among traits [9]. These studies optimize agreement with human raters; they do not examine how scores respond to construct-irrelevant variation in the language of a response.

2.2 Fairness in Automated Scoring

Williamson et al. list the impact on population subgroups among the evidence required before automated scores are used [10], and recommend flagging standardized mean differences between machine and human scores larger than 0.15. Bridgeman et al. reported that e-rater and human raters diverged for test takers from particular countries and ethnic groups [11]. Madnani et al. released tools for subgroup analysis of scoring models [12], and Loukina et al. distinguished overall score accuracy, overall score difference and conditional score difference as separate dimensions of fairness [13]. Mayfield et al. argued that equity requires attention to how technologies shape learning opportunities, beyond statistical parity [14], and Baker and Hawn surveyed where algorithmic bias enters educational systems [15]. Kwako et al. applied item response theory to measure gender and racial bias in a BERT-based speaking assessment [16]. These analyses depend on demographic labels. The present work complements them with an intervention-based audit that needs no labels and separates construct-relevant from construct-irrelevant sensitivity.

2.3 Dialect and Learner-Language Variation

Blodgett et al. documented that AAE text is misprocessed by standard NLP tools [18], and later framed such failures as questions of power and representation [17]. VALUE transformed benchmark data with linguistically validated AAE rules and exposed performance gaps of pre-trained models [19]. On the learner side, JFLEG provides fluency corrections of learner sentences [21], TOEFL-Spell annotates misspellings in essays written by TOEFL test takers [22], and artificial error generation produces synthetic learner errors from observed patterns [23]. Ding et al. showed that content scoring systems accept input that human raters would reject [20], which motivates systematic robustness testing of scoring engines. We combine these strands into transformations that preserve content while changing only dialectal or learner features.

2.4 Counterfactual Fairness and Representation Debiasing

Counterfactual fairness requires a prediction to remain unchanged when a protected attribute and its consequences are altered [24]. Garg et al. adapted the idea to text through counterfactual token substitution and proposed counterfactual logit pairing [25]. Counterfactual data augmentation (CDA) trains on rewritten examples with copied labels [26]. Adversarial training removes attribute information from encoders [27], iterative nullspace projection (INLP) repeatedly projects out directions that linearly predict the attribute [28], and SENT-DEBIAS removes a bias subspace estimated from counterfactual sentence pairs [29]. These methods treat the attribute as irrelevant to every downstream decision. CAIR differs in three respects: it acts on the scoring head instead of a shared representation, it applies invariance selectively according to a construct map derived from the rubric, and it has a closed-form solution.

3 PROBLEM FORMULATION

Let x be a student response, f a frozen LLM encoder and $\hat{s}(x) = w_t^\top f(x) + b_t$ the predicted score for target t , which is either a holistic score or a rubric trait. A linguistic variety $v \in \mathcal{V} = \{\text{AAE}, \text{L2}\}$ is represented by a stochastic transformation T_v that alters dialectal or learner features of x and leaves its content intact. The counterfactual shift of a response and its standardized mean over N responses are

$$\Delta_v(x) = \hat{s}(T_v(x)) - \hat{s}(x), \quad \bar{\Delta}_v = \frac{1}{\sigma_y N} \sum_{i=1}^N \Delta_v(x_i) \quad (1)$$

where σ_y is the standard deviation of human scores in the training split. Because students receive rounded scores, we also report the counterfactual flip rate

$$\text{FR}_v = \frac{1}{N} \sum_{i=1}^N \mathbf{1} \left[\text{rd}(\hat{s}(T_v(x_i))) \neq \text{rd}(\hat{s}(x_i)) \right] \quad (2)$$

where $\text{rd}(\cdot)$ rounds and clips a prediction to the score scale. The mean shift captures systematic disadvantage of a variety, while the flip rate is the proportion of individual students whose reported score would change.

A construct map assigns each target to one of two classes. For an invariant target, such as the content score of a source-based response or the Ideas trait of a narrative, the rubric makes no reference to the conventions of Standard English, and an equitable scorer should satisfy $\Delta_v \approx 0$ with a low FR_v . For a construct-relevant target, such as the Conventions trait, the rubric explicitly evaluates grammar, usage and spelling against Standard English, so a shift under T_v is expected and should be kept as feedback. The goal is to reduce counterfactual shifts on invariant targets without lowering agreement with human raters and without suppressing sensitivity on construct-relevant targets.

4 METHODOLOGY

4.1 Counterfactual Generation Grounded in Learner and Dialect Data

Figure 1 summarizes the framework. The L2 transformation injects ten error categories whose conditional probabilities are estimated from JFLEG [21]. For each of the 1,501 JFLEG source sentences and each of its four corrections, the tokens of the correction and of the learner sentence are aligned, and single-token edits are classified by part of speech and lemma. The rate of a category is the number of learner deviations divided by the number of eligible sites in the corrections, for example the proportion of articles in the corrected text that the learner omitted. The estimated rates are 7.6% for article omission, 0.9% for article substitution, 2.0% and 2.7% for preposition omission and substitution, 2.1% for *be*-agreement, 6.4% and 1.1% for plural-to-singular and singular-to-plural noun errors, 9.0% for omission and 1.3% for insertion of third-person *-s*, and 8.0% for past-to-base tense errors. Misspellings are drawn from TOEFL-Spell [22]. We keep orthographic misspellings of words with at least four letters that share the first letter with the intended word and lie within edit distance two, which yields 1,905 target words, and set the per-word probability to 0.072 so that the expected density matches the TOEFL-Spell average of about 1.8 misspellings per 100 words.

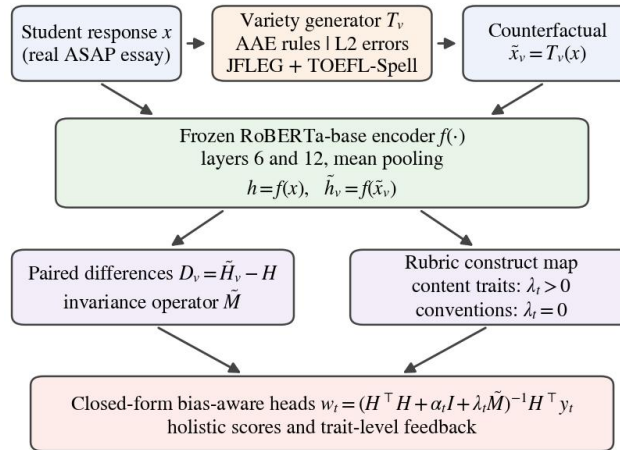


Figure 1 Overview of the Bias-Aware Scoring Framework

Note: Counterfactual variants of each essay are encoded by a frozen LLM; paired differences define an invariance operator that CAIR applies only to rubric dimensions that are independent of language conventions.

The AAE transformation applies eight morphosyntactic features described in the linguistic literature and used in dialect robustness studies: copula and auxiliary absence (*she happy*), zero third-person *-s* (*he walk*), negative concord (*don't have no*), *ain't* for negated *be* and *have*, *was*-levelling (*they was*), existential *it* (*it's many problems*), possessive *they* (*they house*) and plural *-s* absence after numerals (*two dog*) [18-19]. The rules operate on dependency parses and respect known constraints; for instance, the copula is retained with first-person subjects, with *it*, *that* and *what*, and in clause-final position. Each eligible site is transformed with probability 0.7 to reflect variable use. Anonymization tokens and acronyms are never altered, and whitespace is normalized identically for originals and variants, so every difference within a pair is a linguistic edit. For example, in a response to prompt 5 the sentence *The author gave a loving mood in the memoir* becomes *The author give loving mood in the memoir* in the L2 variant, and *he did not talk badly about anything* becomes *he did not talk badly about nothing* in the AAE variant, where the misspelling *badily* is the student's own.

4.2 LLM Representation

The frozen encoder is RoBERTa-base with 12 layers and 125M parameters, taken from the checkpoint distributed with the spaCy `en_core_web_trf` 3.6 pipeline, whose transformer was further trained on tagging, parsing and entity recognition [5]. A response is truncated at 512 subword tokens, and the hidden states of layers 6 and 12 are mean-pooled over non-special tokens and concatenated into a 1,536-dimensional vector h . Intermediate layers carry transferable lexical and

syntactic information that complements the top layer [30]. Keeping the encoder frozen makes all mitigation methods directly comparable and allows a new scoring head to be trained in seconds.

4.3 Construct-Aligned Invariance Regularization

Let $H \in \mathbb{R}^{n \times d}$ contain the standardized training representations, y_t the centered scores of target t , and $D_v = \tilde{H}_v - H$ the paired differences between variant and original representations of the training responses that have counterfactual pairs. CAIR solves

$$\min_w \|Hw - y_t\|_2^2 + \alpha_t \|w\|_2^2 + \lambda_t \sum_{v \in \mathcal{V}} \|D_v w\|_2^2 \quad (3)$$

The third term penalizes the component of the scoring direction w that lies along variety-induced displacements, so predictions for x and $T_v(x)$ are pulled together. The minimizer has the closed form

$$w_t^* = (H^T H + \alpha_t I + \lambda_t \tilde{M})^{-1} H^T y_t, \quad \tilde{M} = \kappa \sum_{v \in \mathcal{V}} D_v^T D_v \quad (4)$$

where $\kappa = \text{tr}(H^T H) / \text{tr}(\sum_v D_v^T D_v)$ makes λ_t comparable across prompts and pair budgets. As $\lambda_t \rightarrow \infty$ the estimator becomes ridge regression restricted to the orthogonal complement of the variety subspace; a finite λ_t balances invariance against fit along directions that also carry score information.

The relation to CDA clarifies the design. CDA adds the variant representations $\tilde{H}_v$ with copied labels, and for each variety its loss decomposes as

$$\|\tilde{H}_v w - y\|^2 = \|Hw - y\|^2 + 2(Hw - y)^T D_v w + \|D_v w\|^2 \quad (5)$$

The augmented objective therefore contains the CAIR penalty, a duplicate of the fit term, and a cross term $2(Hw - y)^T D_v w$ that couples the variety direction with the residuals. This cross term can be lowered by moving predictions along variety directions whenever residuals correlate with them, which reintroduces the dependence that the penalty is meant to remove. CAIR keeps only the consistency term. Projection methods such as SENT-DEBIAS and INLP instead delete a subspace from h for every head, and thereby impose the same invariance on all targets.

4.4 Trait-Selective Invariance

CAIR applies $\lambda_t > 0$ only to targets that the construct map labels invariant. For ASAP prompts 3–6, the rubrics score the understanding and use of a source text, so the holistic score is invariant. For prompt 7, the Ideas, Organization and Style traits are invariant, whereas Conventions is construct-relevant and receives $\lambda_t = 0$. Because each target has its own head and closed-form solution, a single representation serves all traits without compromise.

4.5 Model Selection

The ridge strength α_t is selected by the development-set quadratic weighted kappa (QWK) of the unmitigated model and then fixed for all methods. The invariance strength is chosen as

$$\lambda_t^* = \arg\min_{\lambda \in \Lambda} |\bar{\Delta}^{\text{dev}}(\lambda)| \quad \text{s. t.} \quad \text{QWK}^{\text{dev}}(\lambda) \geq \text{QWK}^{\text{dev}}(0) - \delta \quad (6)$$

with $\delta = 0.01$ and $\Lambda = \{0, 0.03, 0.1, 0.3, 1, 3, 10, 30, 100, 300\}$: among the values whose development QWK stays within δ of the unmitigated model, the one with the smallest development shift is chosen. The same rule selects the hyperparameters of all baselines. Training cost is dominated by one $d \times d$ linear solve per candidate λ , which takes well under a second for $d = 1,536$. Algorithm 1 summarizes the procedure.

Algorithm 1. Training of CAIR scoring heads

Input: essays $\{\mathbf{x}_i, \mathbf{y}_i\}$, varieties $\mathcal{V}$, construct map, grids for α and Λ , tolerance δ

- 1: generate counterfactuals $\tilde{\mathbf{x}}_{v,i} = T_v(\mathbf{x}_i)$ for every $v \in \mathcal{V}$
- 2: encode $\mathbf{h}_i = f(\mathbf{x}_i)$, $\tilde{\mathbf{h}}_{v,i} = f(\tilde{\mathbf{x}}_{v,i})$; standardize with training statistics
- 3: form $D_v = \tilde{H}_v - H$ and the operator $\tilde{M} = \kappa \sum_v D_v^T D_v$
- 4: for each target t do
- 5: select α_t by development QWK of the ridge head ($\lambda = 0$)
- 6: if t is invariant then select λ_t by (6) else $\lambda_t \leftarrow 0$
- 7: $\mathbf{w}_t \leftarrow (H^T H + \alpha_t I + \lambda_t \tilde{M})^{-1} H^T \mathbf{y}_t$
- 8: end for

Output: scoring heads $\{\mathbf{w}_t\}$ for holistic scores and trait feedback

5 EXPERIMENTAL SETUP

5.1 Data

Experiments use the ASAP essay data [1]. Prompts 3–6 are source-dependent responses written in grades 8 and 10 and scored for content on a 0–3 (prompts 3 and 4) or 0–4 (prompts 5 and 6) scale. From each of these prompts, 1,000 responses are sampled with stratification by score. Prompt 7 is a grade-7 narrative with four traits (Ideas, Organization, Style and Conventions), each scored 0–3 by two raters; the trait target is the sum of both ratings (0–6), and all 1,569 essays are used. Table 1 reports corpus statistics, human–human agreement and the density of counterfactual edits. Between 82% and 95% of the AAE variants and between 93% and 98% of the L2 variants differ from their originals.

Table 1 Data and Counterfactual Statistics

Prompt	Type	N	Scale	Words	hQWK	AAE	L2
3	Source	1000	0–3	107.3	0.766	3.06	4.06
4	Source	1000	0–3	93.1	0.855	4.32	4.22
5	Source	1000	0–4	120.1	0.754	4.36	4.37
6	Source	1000	0–4	152.0	0.767	1.59	4.26
7	Narrative	1569	4×(0–6)	159.9	0.596	1.71	3.53

Note: hQWK: agreement between the two human raters (mean over the four traits for prompt 7). AAE and L2: counterfactual edits per 100 words.

5.2 Protocol and Metrics

Each prompt is split into 60% training, 20% development and 20% test data, with five random stratified splits. Counterfactual variants of training responses provide the pairs used by CAIR and by all baselines; development variants serve only for model selection; results are reported on test responses and their variants. The metrics are QWK on original and variant essays, the standardized shift Δ_v of (1) and the flip rate FR_v of (2). Significance is assessed with the Wilcoxon signed-rank test over the 20 prompt–split combinations of prompts 3–6.

5.3 Compared Methods

TF-IDF is a ridge regressor on word unigram and bigram TF-IDF features and serves as a non-neural reference. LLM-Ridge is the unmitigated ridge head on RoBERTa representations. CDA augments the training data with both variants and copied labels, with weight $\gamma \in \{0.25, 0.5, 1, 2\}$ [26]. SENT-DEBIAS removes the top- k principal directions of the paired differences, $k \in \{1, 2, 4, 8, 16, 32\}$ [29]. INLP removes the directions found by $T \in \{1, 2, 4, 8, 16\}$ iterations of linear classifiers that separate originals from variants [28]. CAIR-uniform is an ablation that applies the CAIR penalty to every trait, Conventions included.

5.4 Implementation

Representations were computed with PyTorch and Hugging Face Transformers in bfloat16 precision on a single CPU core with AMX support, at about 2,300 tokens per second. All scoring heads and baselines use closed-form solvers implemented in NumPy, and the transformations rely on spaCy dependency parses.

6 RESULTS

6.1 Holistic Content Scoring

Table 2 averages the results over prompts 3–6. The unmitigated LLM scorer reaches a QWK of 0.733, far above the TF-IDF reference (0.607), which confirms the value of LLM representations for content scoring. Its counterfactual behavior exposes the problem that motivates this work: AAE variants lower the predicted score by 0.023 standard deviations on average and L2 variants by 0.013, and the rounded score changes for 4.5% (AAE) and 5.6% (L2) of the students. These mean shifts are small compared with the 0.15 criterion used for subgroup screening [10], so a group-level check would accept the model even though roughly one student in twenty receives a different score because of dialect or learner features alone.

Table 2 Holistic Content Scoring on Prompts 3–6 (Mean $\pm$ SD over Five Splits)

Method	QWK	Δ AAE	Δ L2	FRAAE (%)	FRL2 (%)
TF-IDF	0.607 $\pm$ 0.024	−0.021 $\pm$ 0.008	−0.025 $\pm$ 0.003	5.22 $\pm$ 1.60	4.88 $\pm$ 0.70
LLM-Ridge	0.733 $\pm$ 0.005	−0.023 $\pm$ 0.013	−0.013 $\pm$ 0.012	4.45 $\pm$ 0.54	5.62 $\pm$ 0.41
CDA	0.733 $\pm$ 0.006	−0.006 $\pm$ 0.005	−0.005 $\pm$ 0.006	3.55 $\pm$ 0.72	4.65 $\pm$ 0.21
SENT-DEBIAS	0.732 $\pm$ 0.007	−0.001 $\pm$ 0.002	0.001 $\pm$ 0.004	3.32 $\pm$ 0.68	5.20 $\pm$ 0.73
INLP	0.733 $\pm$ 0.006	−0.020 $\pm$ 0.011	−0.011 $\pm$ 0.013	4.38 $\pm$ 0.44	5.55 $\pm$ 0.24
CAIR (ours)	0.732 $\pm$ 0.006	−0.001 $\pm$ 0.003	0.000 $\pm$ 0.002	2.62 $\pm$ 0.57	3.40 $\pm$ 0.62

Note: Δ : standardized counterfactual shift, Eq. (1); FR: counterfactual flip rate, Eq. (2). QWK on AAE and L2 variants of the CAIR scorer is 0.733 and 0.734.

CAIR reduces the mean absolute standardized shift from 0.021 to 0.003, a reduction of 85% ($p < 0.001$), and lowers the flip rates to 2.6% for AAE and 3.4% for L2, relative reductions of 41% and 40%. QWK on the original essays is unchanged within noise (0.732 vs. 0.733, $p = 0.97$). CDA reduces the mean shift less (0.009) and leaves higher flip rates (3.6% and 4.7%). SENT-DEBIAS nearly cancels the mean shift, but its flip rates remain at 3.3% and 5.2%: deleting a few principal directions removes the average displacement without making individual predictions stable. INLP barely changes the scorer, because the directions that best separate originals from variants are not the directions the scoring head relies on, a gap between information removal and decision invariance. The absolute shift of CAIR is significantly lower than that of every baseline ($p = 0.048$ against SENT-DEBIAS and $p < 0.001$ against the others).

Table 3 breaks the comparison down by prompt. CAIR keeps both shifts within 0.005 standard deviations on every prompt and lowers the flip rate on every prompt and variety. The largest gain is on prompt 3, where the L2 flip rate falls from 6.4% to 2.3% while QWK rises from 0.632 to 0.640; the largest change in agreement occurs on prompt 6 (0.770 to 0.756).

Table 3 Per-Prompt Results of the Unmitigated Scorer and CAIR

Prompt	Method	QWK	Δ AAE	Δ L2	FRAAE	FRL2
3	LLM-Ridge	0.632	-0.026	-0.031	3.9	6.4
	CAIR	0.640	0.000	-0.001	1.2	2.3
4	LLM-Ridge	0.757	-0.029	-0.012	6.1	6.4
	CAIR	0.760	-0.002	-0.002	4.3	4.4
5	LLM-Ridge	0.771	-0.032	0.009	4.4	3.3
	CAIR	0.771	-0.005	0.002	2.4	2.7
6	LLM-Ridge	0.770	-0.005	-0.016	3.4	6.4
	CAIR	0.756	0.002	0.000	2.6	4.2

Note: Flip rates in percent; mean over five splits.

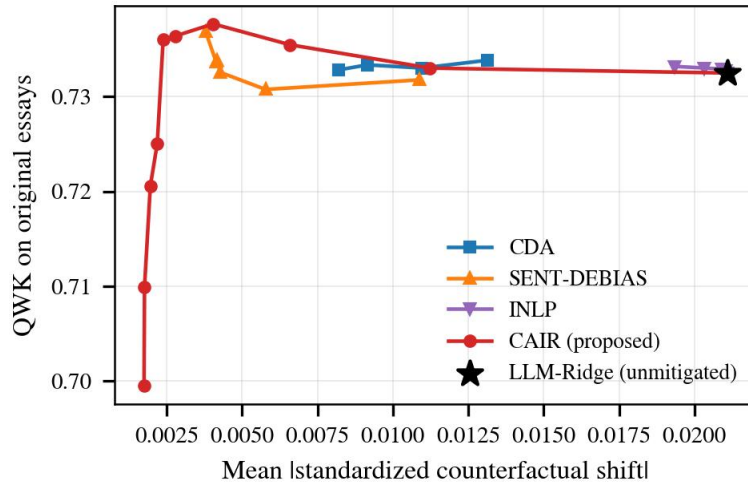


Figure 2 Accuracy–Invariance Frontier on Prompts 3–6 (Mean over Prompts and Five Splits), Obtained by Sweeping λ (CAIR), γ (CDA), k (SENT-DEBIAS) and T (INLP)

The frontier in Figure 2 explains why the selection rule favors CAIR. Moderate invariance costs no accuracy: at $\lambda = 0.3$ CAIR attains a mean absolute shift of 0.004 with a QWK of 0.738, better on both axes than the strongest CDA setting (0.008, 0.733), and at $\lambda = 1$ it reaches 0.003 with a QWK of 0.736, below the smallest shift obtained by SENT-DEBIAS (0.004). The slight gain in agreement at moderate λ indicates that the penalty also regularizes the head against surface cues. Agreement declines only for $\lambda \geq 10$ and reaches 0.700 at $\lambda = 300$, while the shift saturates near 0.002. INLP never falls below 0.019, because its projections target the detection of a variety rather than the scoring direction.

On the development data, the selection rule most often chose $\lambda = 10$ for the holistic heads, $\lambda = 1$ for Ideas and Organization and $\lambda = 0.1$ for Style. CDA selected its largest weight ($\gamma = 2$), SENT-DEBIAS $k = 8$, and INLP $T = 16$ for holistic scoring and $T = 1$ for traits, where additional projections reduced development QWK beyond the tolerance.

6.2 Trait-Level Feedback

Table 4 and Figure 3 report the four traits of prompt 7. Without mitigation, L2 variants lower the Conventions prediction by 0.138 standard deviations, which is the expected, rubric-consistent response to additional grammar and spelling errors. The same variants also lower Organization by 0.037 and Style by 0.052, and they change the rounded Ideas, Organization and Style scores of 7.7%, 7.5% and 8.7% of the students. This spill-over produces misleading feedback: a learner would be told that the organization of a narrative is weak because of article or agreement errors.

Table 4 Trait Results on Prompt 7: QWK and L2 Shift Δ L2

Method	Ideas		Organization		Style		Conventions	
	QWK	Δ	QWK	Δ	QWK	Δ	QWK	Δ
LLM-Ridge	0.774	-0.014	0.692	-0.037	0.691	-0.052	0.707	-0.138
CDA	0.773	-0.009	0.687	-0.022	0.693	-0.037	0.704	-0.097
SENT-DEBIAS	0.772	-0.002	0.688	-0.001	0.685	0.000	0.681	-0.004
INLP	0.774	-0.014	0.692	-0.037	0.692	-0.052	0.707	-0.138
CAIR-uniform	0.771	-0.006	0.688	-0.010	0.691	-0.017	0.695	-0.037
CAIR (ours)	0.768	-0.004	0.683	-0.005	0.703	-0.017	0.707	-0.138

Note: Ideas, Organization and Style are invariant targets; Conventions is construct-relevant, so a negative shift under L2 variants is desired.

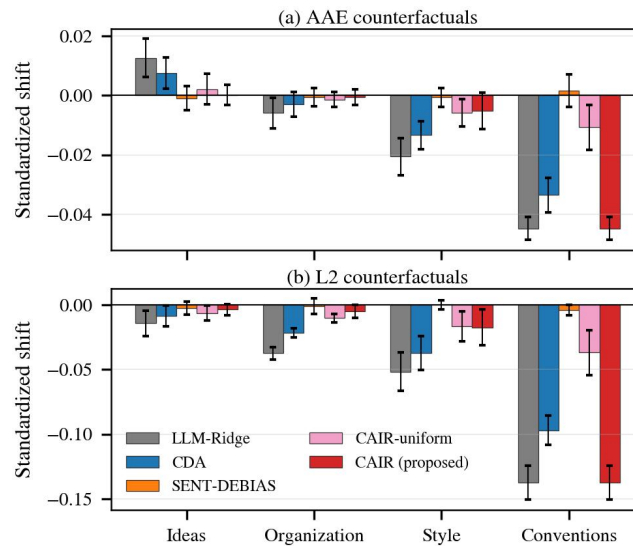


Figure 3 Standardized Counterfactual Shift of Trait Predictions on Prompt 7 (Mean $\pm$ SD over Five Splits)
Note: Negative values indicate lower predicted scores for the variant.

CAIR reduces the mean absolute L2 shift on the three invariant traits from 0.034 to 0.009 (-75%) and their flip rates to 5.5%, 4.2% and 5.1%. Because Conventions receives $\lambda = 0$, its sensitivity (-0.138) and its QWK (0.707) are untouched, and Style agreement improves from 0.691 to 0.703. The ratio between the Conventions shift and the mean invariant-trait shift rises from 4.0 to 15.9, so the effect of learner errors is concentrated on the dimension where the rubric places it. Task-agnostic methods cannot produce this separation. SENT-DEBIAS drives every trait, Conventions included (-0.004), to near-zero shift and lowers Conventions QWK to 0.681; CAIR-uniform keeps only 27% of the Conventions sensitivity and lowers its QWK to 0.695; CDA retains 70% of the sensitivity but leaves larger spill-over on Organization (-0.022) and Style (-0.037); INLP leaves the trait scores essentially unchanged. AAE variants show the same pattern at a smaller scale: CAIR limits the invariant-trait shifts to at most 0.005 in absolute value, while the Conventions shift of -0.045 is preserved.

6.3 Pair Budget and Edit Density

Figure 4(a) varies the fraction of training essays for which counterfactual pairs are available (three splits). Pairs for 2% of the training essays, 12 per prompt, already cut the mean absolute shift from 0.025 to 0.011, and pairs for 10% reduce it to 0.006 (-77%) at unchanged QWK. The flip rate declines steadily from 5.0% to 2.7% as the budget grows to the full training set. Figure 4(b) groups test essays by the density of counterfactual edits. The penalty imposed by the unmitigated scorer grows with density: for AAE variants the mean shift moves from -0.005 in essays with fewer than two edits per 100 words to -0.059 in essays with eight or more, a twelvefold increase, and for L2 variants from -0.005 to -0.027 at six to eight edits per 100 words. Under CAIR the mean shift stays within ± 0.003 for AAE in every bin and within ± 0.011 for L2, so the students whose writing contains the most dialect or learner features, and who are therefore most exposed to bias, benefit the most.

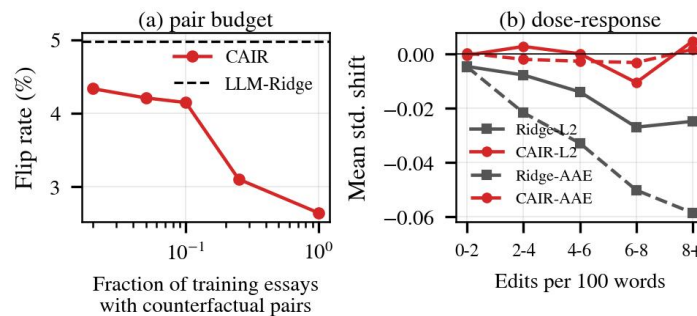


Figure 4 (a) Counterfactual Flip Rate of CAIR versus the Fraction of Training Essays with Counterfactual Pairs (Prompts 3–6, Three Splits). (b) Mean Standardized Shift of Test Essays Binned by Edit Density for the Unmitigated Scorer and CAIR

Table 5 lists the composition of the edits over all 5,569 essays. Misspellings and article errors make up almost two thirds of the L2 edits, and zero third-person *-s* and copula absence account for 71% of the AAE edits. Both varieties therefore concentrate on frequent function-word, morphological and orthographic changes that leave the content of a response untouched.

Table 5 Composition of Counterfactual Edits (5,569 Essays)

L2 category	Edits	%	AAE feature	Edits	%
Misspelling	12,842	44.5	Zero 3sg -s	8,190	42.6
Article	5,592	19.4	Copula absence	5,453	28.3
Noun number	3,094	10.7	Was-levelling	1,508	7.8
Preposition	2,828	9.8	Possessive they	1,333	6.9
Tense	2,647	9.2	Existential it	1,156	6.0
Agreement	1,867	6.5	Plural -s absence	959	5.0
			Ain't	329	1.7
			Negative concord	316	1.6
Total	28,870	100	Total	19,244	100

6.4 Transfer Across Varieties and Representation Analysis

Table 6 estimates the invariance operator from one variety and evaluates it on both. Invariance learned from L2 pairs transfers to AAE, reducing the AAE shift from -0.026 to -0.008 and the AAE flip rate from 4.4% to 2.8%. The reverse direction reduces the L2 shift from -0.020 to -0.009 . The mean displacements of the two varieties have a cosine similarity of 0.55, so the transfer is partial, and pairs from both varieties give the best result.

Table 6 Cross-Variety Transfer of CAIR (Prompts 3–6, Three Splits)

Pairs $\rightarrow$ evaluated	Δ^- (CAIR)	FR (CAIR)	Δ^- (Ridge)	FR (Ridge)
AAE $\rightarrow$ AAE	0.000	2.2	-0.026	4.4
AAE $\rightarrow$ L2	-0.009	4.3	-0.020	5.5
L2 $\rightarrow$ AAE	-0.008	2.8	-0.026	4.4
L2 $\rightarrow$ L2	0.000	4.2	-0.020	5.5

Note: Ridge denotes the unmitigated LLM-Ridge scorer; flip rates in percent.

The representation itself encodes the variety of a response. A linear probe distinguishes AAE and L2 variants from their originals with 81.0% and 74.3% accuracy, although the members of each pair have identical content and identical human scores. The mean displacement produced by synthetic L2 errors has a cosine of 0.86 with the mean displacement between real JFLEG learner sentences and their corrections, and its ten leading directions capture 26.3% of the energy of the real learner–correction differences, compared with 0.65% for a random ten-dimensional subspace and 22.1% for the ten leading content directions of the essays. The generator therefore moves representations in the same direction as authentic learner language.

7 DISCUSSION

Three findings matter for equitable assessment. First, group-level mean differences and individual-level stability diverge. All unmitigated mean shifts in Table 2 are far below the thresholds used in operational subgroup screening, yet one score in twenty changes under a dialect or learner-language counterfactual. Counterfactual flip rates therefore belong in the evidence reported for automated scores, next to agreement and subgroup statistics [10, 13]. Second, equity in feedback depends on the construct. Removing variety information everywhere protects the content score but also silences the conventions feedback that students and teachers rely on. A construct map derived from the rubric turns this tension into an explicit design decision, and the per-trait closed-form estimator implements the decision at negligible cost. Third, the audit needs no demographic labels. Counterfactual pairs can be generated for any essay collection, a few dozen of them already reduce most of the shift, and they can be regenerated when prompts or rubrics change, which suits institutions that cannot or should not collect sensitive attributes. The framework extends directly to further varieties, such as regional Englishes or specific first-language backgrounds, by adding transformation rules and pairs.

A practical audit protocol follows from these results. Before deployment, a scoring team generates variants for a development sample, reports Δ^- and FR_v per variety next to QWK, and fixes δ as the accuracy budget that the testing program accepts. Retraining the heads with CAIR takes seconds, so the audit can be repeated whenever the rubric, the prompt pool or the encoder changes. Because the construct map is expressed in terms of rubric dimensions, it can be reviewed by assessment specialists without inspecting the model.

8 CONCLUSION

This paper introduced a bias-aware scoring framework for LLM-based educational assessment and feedback. Counterfactual essays grounded in AAE grammar and in learner-error statistics reveal how a frozen RoBERTa scorer changes scores for reasons that the rubric excludes. Construct-Aligned Invariance Regularization removes most of this sensitivity for content scores and content-related traits in closed form, maintains agreement with human raters, and preserves the conventions feedback that the rubric intends. Future work will extend the construct map to argumentative prompts with multi-dimensional rubrics, combine counterfactual invariance with group-conditional error analysis where demographic data are available, and apply the approach to generative feedback models.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

- [1] Shermis M D. State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, 2014, 20: 53-76.
- [2] Taghipour K, Ng H T. A neural approach to automated essay scoring. In: *Proceedings of Conference Empirical Methods in Natural Language Processing (EMNLP)*, 2016: 1882-1891.
- [3] Dong F, Zhang Y, Yang J. Attention-based recurrent convolutional neural network for automatic essay scoring. In: *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL)*, 2017: 153-162.
- [4] Devlin J, Chang M-W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of Conference North American Chapter of the ACL: Human Language Technologies (NAACL-HLT)*, 2019: 4171-4186.
- [5] Liu Y, Ott M, Goyal N, et al. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*, 2019.
- [6] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. In: *Advances in Neural Information Processing Systems*, 2020, 33: 1877-1901.
- [7] Mayfield E, Black A W. Should you fine-tune BERT for automated essay scoring? In: *Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications*, 2020: 151-162.
- [8] Yang R, Cao J, Wen Z, et al. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020: 1560-1569.
- [9] Kumar R, Mathias S, Saha S, et al. Many hands make light work: Using essay traits to automatically score essays. In: *Proceedings of NAACL-HLT*, 2022: 1485-1495.
- [10] Williamson D M, Xi X, Breyer F J. A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 2012, 31(1): 2-13.
- [11] Bridgeman B, Trapani C, Attali Y. Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, 2012, 25(1): 27-40.
- [12] Madnani N, Loukina A, von Davier A, et al. Building better open-source tools to support fairness in automated scoring. In: *Proceedings of First ACL Workshop on Ethics in Natural Language Processing*, 2017: 41-52.
- [13] Loukina A, Madnani N, Zechner K. The many dimensions of algorithmic fairness in educational applications. In: *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications*, 2019: 1-10.
- [14] Mayfield E, Madaio M, Prabhunoye S, et al. Equity beyond bias in language technologies for education. In: *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications*, 2019: 444-460.
- [15] Baker R S, Hawn A. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 2022, 32(4): 1052-1092.
- [16] Kwako A, Wan Y, Zhao J, et al. Using item response theory to measure gender and racial bias of a BERT-based automated English speech assessment system. In: *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications*, 2022: 1-7.
- [17] Blodgett S L, Barocas S, Daumé III H, et al. Language (technology) is power: A critical survey of ‘bias’ in NLP. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020: 5454-5476.
- [18] Blodgett S L, Green L, O’Connor B. Demographic dialectal variation in social media: A case study of African-American English. In: *Proceedings of EMNLP*, 2016: 1119-1130.
- [19] Ziems C, Chen J, Harris C, et al. VALUE: Understanding dialect disparity in NLU. In: *Proceedings of the 60th Annual Meeting of the ACL (Volume 1: Long Papers)*, 2022: 3701-3720.
- [20] Ding Y, Riordan B, Horbach A, et al. Don’t take ‘nswvtnvakgxp’ for an answer – The surprising vulnerability of automatic content scoring systems to adversarial input. In: *Proceedings of the 28th International Conference Computational Linguistics (COLING)*, 2020: 882-892.
- [21] Napoles C, Sakaguchi K, Tetreault J. JFLEG: A fluency corpus and benchmark for grammatical error correction. In: *Proceedings of the 15th Conference European Chapter of the ACL (EACL)*, 2017, 2: 229-234.
- [22] Flor M, Fried M, Rozovskaya A. A benchmark corpus of English misspellings and a minimally-supervised model for spelling correction. In: *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications*, 2019: 76-86.
- [23] Felice M, Yuan Z. Generating artificial errors for grammatical error correction. In: *Proceedings of Student Research Workshop at the 14th Conference European Chapter of the ACL*, 2014: 116-126.
- [24] Kusner M J, Loftus J, Russell C, et al. Counterfactual fairness. In: *Advances in Neural Information Processing Systems*, 2017, 30.
- [25] Garg S, Perot V, Limtiaco N, et al. Counterfactual fairness in text classification through robustness. In: *Proceedings of AAAI/ACM Conference AI, Ethics, and Society (AIES)*, 2019: 219-226.
- [26] Zmigrod R, Mielke S J, Wallach H, et al. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In: *Proceedings of the 57th Annual Meeting of the ACL*, 2019: 1651-1661.
- [27] Elazar Y, Goldberg Y. Adversarial removal of demographic attributes from text data. In: *Proceedings of EMNLP*, 2018: 11-21.

- [28] Ravfogel S, Elazar Y, Gonen H, et al. Null it out: Guarding protected attributes by iterative nullspace projection. In: Proceedings of the 58th Annual Meeting of the ACL, 2020: 7237-7256.
- [29] Liang P P, Li I M, Zheng E, et al. Towards debiasing sentence representations. In: Proceedings of the 58th Annual Meeting of the ACL, 2020: 5502-5515.
- [30] Liu N F, Gardner M, Belinkov Y, et al. Linguistic knowledge and transferability of contextual representations. In: Proceedings of NAACL-HLT, 2019: 1073-1094.