International Journal of World Economic Research
ISSN: 3080-9886 (Print), ISSN: 3080-9894 (Online) | Vol. 4, No. 1, 2026

A Comparative Study of Linguistic Bias and Fairness in Large Language Models for Automated Essay Scoring

Leila Diaz [1], Liam Sato [2], Jasmine Yoon [3, *]

[1] Graduate School of Education, Rutgers University–New Brunswick, USA

[2] College of Education, Clemson University, USA

[3] College of Education, Purdue University, USA

[*] **Corresponding author**: (Email: j.yoon29@purdue.edu)

Abstract: Large language model (LLM) representations have become the default backbone of automated essay scoring, yet the resulting scores may react to construct-irrelevant features of African American English (AAE) and learner (L2) English. We present a comparative counterfactual study of linguistic bias across five representation families—word and character n-grams, static word vectors, LLM-derived static token embeddings (WordLlama) and RoBERTa examined layer by layer—on 5,569 ASAP essays. Real essays are rewritten with eight AAE morphosyntactic rules and with learner errors whose rates and misspellings are estimated from JFLEG and TOEFL-Spell, and every edit is labelled so that score shifts can be attributed to fourteen phenomena. All unmitigated scorers lower the scores of counterfactual variants, but the mean shifts stay below 0.06 standard deviations while 4.0–9.7% of individual reported scores change, so group-level screening would accept scorers that alter individual outcomes. Within RoBERTa, sensitivity to dialect grows with depth: the AAE penalty increases from −0.015 standard deviations at the embedding output to −0.077 at layer 10, and the contextual model penalises 13 of the 14 phenomena. Construct-aligned invariance regularisation lowers counterfactual flip rates on content scores for every family (by 28–60%) with QWK changes between −0.007 and +0.004. We extend it with variety-selective construct maps and sensitivity anchoring: on the Conventions trait of RoBERTa, the extension removes 80% of the AAE penalty while retaining 98% of the learner-error penalty, whereas a uniform map retains 66%. A geometric bound and the measured overlap of dialect and learner displacement directions show why surface-identical phenomena, such as a missing third-person -s, cannot be treated differently by any linear scorer.

Keywords: Automated essay scoring, large language models, linguistic bias, counterfactual fairness, African American English, English learners, debiasing.

1. Introduction

Automated essay scoring (AES) has moved from research prototypes to high-volume use in classrooms and testing programmes, a transition accelerated by the public Automated Student Assessment Prize (ASAP) benchmark [1]. Neural scorers replaced hand-crafted feature sets with learned representations [2], and pre-trained transformer language models such as BERT and RoBERTa now provide the default backbone for holistic and trait scoring [3–5]; generative large language models (LLMs) are increasingly used as scorers as well [6]. These representations are learned from web-scale text and encode far more than the construct a rubric intends to measure. Whether a student's score depends on how that student's variety of English differs from the variety that dominates pre-training data is therefore a validity question, not only a fairness question [7].

Evidence that the question matters comes from several directions. Machine and human essay scores diverge for groups defined by ethnicity and country of origin [8], and fairness in educational applications has several non-equivalent definitions [9]. Recent comparative studies found that AES models differ in subgroup bias across economic, linguistic and psychological variables [10, 11]. Outside education, standard NLP systems misprocess African American English (AAE) [12, 13], language models hold covert dialect prejudice that shapes their decisions [14], and LLM-based detectors misclassify the writing of non-native English speakers [15]. Such failures reproduce social hierarchies among language varieties [16].

Two gaps limit what group-level comparisons can tell a scoring team. First, they report *whether* subgroups receive different scores but not *which* linguistic phenomena cause the difference or *where* in a model it enters, so they give little guidance for repair. Second, demographic labels are often unavailable or legally restricted. A recent bias-aware framework addresses the second gap with a demographic-free counterfactual audit: real essays are rewritten with AAE morphosyntactic rules and with learner (L2) errors estimated from learner corpora, and a construct-aligned invariance regulariser (CAIR) removes score sensitivity only on rubric dimensions that are independent of language conventions [17]. That study evaluated a single encoder and a binary, trait-level construct map.

This paper turns the counterfactual audit into a comparative instrument. We re-implement the dialect and learner-error generators, verify them against their published statistics, and apply them to five representation families that span the path from surface features to contextual LLM representations: word and character n-grams, static word vectors, LLM-derived static token embeddings (WordLlama), and RoBERTa examined layer by layer. Every counterfactual edit carries a phenomenon label, so score shifts can be attributed to fourteen dialect and learner phenomena. We then compare five mitigation strategies across families and extend CAIR in two directions.

The contributions are as follows. (i) A phenomenon-resolved counterfactual audit of 5,569 ASAP essays across

five representation families and all thirteen RoBERTa layers, which separates group-level mean shifts from individual score changes. (ii) A variety-selective construct map with sensitivity anchoring (VS-CAIR+SA), which generalises CAIR from a trait-level binary map to a target-by-variety map in which construct-relevant counterfactual pairs regress onto reference shifts instead of zero, still in closed form. (iii) A geometric bound showing that selective invariance cannot retain sensitivity to a phenomenon whose displacement lies in the span of the phenomena that must be ignored, together with empirical evidence that surface-identical dialect features and learner errors are entangled in exactly this way.

2. Related Work

2.1. LLM Representations for Essay Scoring

ASAP remains the most widely used public AES benchmark [1]. Recurrent networks over word embeddings outperformed feature-engineered systems [2]; fine-tuned and frozen transformer encoders were subsequently compared with lighter models [5], and prompted generative LLMs have been evaluated as zero-shot or feature-producing scorers [6]. Probing studies show that transformer layers encode different linguistic information, with morphosyntax concentrated in lower and middle layers [18, 19]. These findings motivate a layer-wise view of bias, which agreement-oriented AES studies do not provide.

2.2. Fairness in Automated Scoring

Williamson et al. list subgroup impact among the evidence required before automated scores are used and recommend flagging standardised mean differences above 0.15 [7]. Bridgeman et al. reported machine–human differences by gender, ethnicity and country [8], and Loukina et al. distinguished overall score accuracy, overall score difference and conditional score difference [9]. Yang et al. compared nine AES methods on the PERSUADE corpus and found that prompt-specific models were more biased with respect to economic status [10]; Schaller et al. compared shallow, deep and LLM-based scorers on German learner essays [11]. These studies rely on demographic labels and measure disparities at the group level. FairPromote demonstrates the methodological value of measuring predictive usefulness together with subgroup fairness and interpretable feature attribution in consequential automated assessment [31].

2.3. Dialect and Learner Language in NLP

Blodgett et al. documented that AAE text is misprocessed by standard tools [12] and framed such failures as questions of power [16]. VALUE transformed benchmark data with linguistically validated AAE rules and exposed performance gaps of pre-trained models [13]; Hofmann et al. showed that dialect prejudice persists in language models, including RoBERTa, and affects downstream decisions [14]. For learner English, JFLEG provides fluency corrections of learner sentences [20] and TOEFL-Spell annotates authentic misspellings in TOEFL essays [21]; both ground the error statistics used here.

2.4. Counterfactual Fairness and Debiasing

A prediction is counterfactually fair when it is unchanged under an intervention on a protected attribute and its consequences [22]; Garg et al. adapted the notion to text through counterfactual token substitution [23]. Counterfactual data augmentation (CDA) trains on rewritten examples with copied labels [24], SENT-DEBIAS removes a bias subspace estimated from counterfactual pairs [25], and iterative nullspace projection (INLP) repeatedly removes directions that linearly predict the attribute [26]. These methods treat the attribute as irrelevant to every decision. Construct-aligned invariance regularisation (Section IV) instead acts on individual scoring heads and applies invariance only where a rubric requires it; the present study extends this selectivity to individual language varieties and to the preservation of legitimate sensitivity.

3. Counterfactual Audit Framework

3.1. Problem Formulation and Metrics

Let x be a student response, $f(x)$ a frozen representation and $\hat{s}(x) = w^{\top} f(x) + b$ the predicted score for a target t, either a holistic score or a rubric trait. A language variety $v \in \{\text{AAE}, \text{L2}\}$ is represented by a stochastic transformation T_v that alters dialectal or learner features of x and leaves its content intact. Following the audit protocol we adopt, the counterfactual shift of a response and its standardised mean over N responses are

$$\Delta_v(x) = \hat{s}(T_v(x)) - \hat{s}(x), \qquad \bar{\Delta}_v = (N\sigma_y)^{-1} \Sigma_i \Delta_v(x_i) \tag{1}$$

Where σ_y is the standard deviation of human scores in the training split. Because students receive rounded scores, the counterfactual flip rate counts students whose reported score would change,

$$FR_v = N^{-1} \Sigma_i \, 1[\mathrm{rd}(\hat{s}(T_v(x_i))) \neq \mathrm{rd}(\hat{s}(x_i))] \tag{2}$$

Where $\mathrm{rd}(\cdot)$ rounds and clips to the score scale. Positive and negative shifts of individual essays cancel in $\bar{\Delta}_v$, so we additionally report the mean absolute individual shift

$$MAIS_v = (N\sigma_y)^{-1} \Sigma_i \, |\Delta_v(x_i)| \tag{3}$$

Which measures how strongly individual scores react to the intervention regardless of direction.

3.2. Counterfactual Generation

We adopted the dialect and learner-error generators of the bias-aware scoring framework of Fang [17] and re-implemented them on spaCy dependency parses [27]. The AAE transformation applies eight morphosyntactic features (copula and auxiliary absence, zero third-person -s, negative concord, *ain't*, *was*-levelling, existential *it*, possessive *they* and plural -s absence after numerals) with probability 0.7 per eligible site and respects the published constraints, for example the retention of the copula with *I*, *it*, *that* and *what* and in clause-final position. The L2 transformation injects ten grammatical error categories with the conditional rates that the framework estimated from the 1,501 JFLEG source sentences and their four corrections [20], and replaces words with authentic misspellings from TOEFL-Spell [21] with per-word probability 0.072. Anonymisation tokens and acronyms are never altered, and whitespace is normalised identically for originals and variants, so every difference within a pair is a linguistic edit.

Our re-implementation reproduces the published generator statistics. Re-estimating the category rates from JFLEG with our own token alignment gives values within 0.6 percentage points of the published rates for 9 of the 10 categories (largest deviation: past-to-base tense, 5.4% vs. 8.0%), and the TOEFL-Spell filter yields the same 1,905 target words; we

therefore use the published rates. On the 5,569 essays the generators produce 18,268 AAE and 28,272 L2 edits. Zero third-person -s (42.8%) and copula absence (29.7%) dominate the AAE edits, misspellings (45.1%) and article errors (19.2%) dominate the L2 edits, and 79–95% of the AAE and 93–98% of the L2 variants differ from their originals.

3.3. Phenomenon-Level Decomposition

Each sampled edit is labelled with its phenomenon: the eight AAE features and six L2 groups (misspelling, article, noun number, preposition, tense and agreement, the latter combining *be*-agreement and third-person -s errors). A phenomenon-level counterfactual $T_v^{(f)}(x)$ renders only the edits of phenomenon f from the *same* sampled edit set that defines $T_v(x)$. The decomposition therefore attributes the full variety-level shift without resampling, and the sum of phenomenon-level shifts can be compared with the full shift to test additivity. To compare phenomena that occur at very different rates, we report the per-edit potency

$$\pi_f = \Sigma_i \, \Delta_f(x_i) \,/\, (\sigma_y \, \Sigma_i \, n_f(x_i)) \tag{4}$$

Where $n_f(x_i)$ is the number of edits of phenomenon f in essay i; π_f is the average standardised shift caused by a single edit.

4. Scorers and Construct-Aligned Mitigation

4.1. Representation Families

Five families of essay representations are compared, each followed by a linear ridge head [28]. Word n-gram TF-IDF (unigrams and bigrams, 1,000 features) and character n-gram TF-IDF (2–5 characters within word boundaries, 1,000 features) are fitted on the training essays of each fold. Static word vectors are the 300-dimensional vectors distributed with spaCy en_core_web_lg [27], averaged over in-vocabulary tokens. WordLlama (configuration l2_supercat) provides 256-dimensional token embeddings distilled from the token-embedding codebooks of LLMs that share the Llama 2 vocabulary [29], [30]; it represents the input-embedding layer of large generative models without contextualisation. The contextual family is RoBERTa-base [4] with the weights distributed in the spaCy en_core_web_trf 3.8.0 pipeline, exported to a standard transformer implementation; the export reproduces the hidden states of the original implementation to within 10^{-5}. Responses are truncated at 512 subword tokens and each layer is mean-pooled over non-special tokens. The main configuration concatenates layers 6 and 12 (1,536 dimensions), the configuration used by CAIR; all thirteen layer outputs (embedding layer included) are used for the layer-wise analysis.

4.2. Construct-Aligned Invariance Regularisation

Let H contain the standardised training representations, y_t the centred scores of target t, and $D_v = \tilde{H}_v - H$ the paired differences between variant and original representations. CAIR [17] penalises the component of the scoring direction that lies along variety-induced displacements and has the closed form

$$\hat{w}_t = (H^\top H + \alpha I + \lambda_t \kappa \, \Sigma_v \, D_v^\top D_v)^{-1} \, H^\top y_t \tag{5}$$

With $\kappa = \mathrm{tr}(H^\top H)/\mathrm{tr}(\Sigma_v D_v^\top D_v)$, which makes λ_t comparable across prompts and families. Under the trait-level construct map of CAIR, $\lambda_t > 0$ for targets whose rubric is independent of language conventions and $\lambda_t = 0$ otherwise. We use the same estimator for all five families, which allows the regulariser to be compared across representations rather than on one encoder.

4.3. Variety-Selective Maps and Sensitivity Anchoring

Eq. (5) is the ridge solution on an augmented design in which the training rows are stacked with the scaled difference rows $(\lambda_t \kappa)^{1/2} D_v$ whose regression targets are zero. Two consequences follow. First, invariance can be assigned per target *and* per variety: a construct map $\mathbf{\Lambda}$ with entries $\lambda_{t,v}$ replaces the scalar λ_t, so that, for example, a Conventions head may be required to ignore dialect features while remaining responsive to learner errors (VS-CAIR). Second, the zero targets can be replaced by *reference shifts*. For the varieties R_t that are construct-relevant for target t, we anchor the difference rows to the shifts of the unmitigated head w_0, which yields the objective

$$\|Hw - y_t\|^2 + \alpha\|w\|2 + \lambda\kappa \, \Sigma_{v\in I_\{t\}} \, \|D_v w\|^2 + \lambda\kappa \, \Sigma_{v\in R_\{t\}} \, \|D_v(w - w_0)\|^2 \tag{6}$$

Where I_t contains the varieties to which target t must be invariant. Its minimiser remains closed form,

$$\hat{w}_t = (G + \lambda\kappa \, \Sigma_{v\in I_\{t\}\cup R_t\}} \, D_v^\top D_v)^{-1} (H^\top y_t + \lambda\kappa \, \Sigma_{v\in R_\{t\}} \, D_v^\top D_v w_0) \tag{7}$$

With $G = H^\top H + \alpha I$. Without the anchor, a construct-relevant variety is merely unconstrained; because difference directions of different varieties overlap, the invariance imposed for I_t can still erase sensitivity to R_t. The anchor term (SA) counteracts this loss explicitly, at the cost of one additional matrix product.

4.4. A Geometric Limit to Selective Invariance

Let μ_f denote the mean displacement produced by phenomenon f and let U span the mean displacements of the phenomena that must be ignored. **Proposition 1.** If a linear head satisfies $w^\top u = 0$ for all $u \in U$, then for every phenomenon b

$$|w^\top \mu_b| \le \|w\| \, \|(I - \Pi_U)\mu_b\| = \|w\| \, \|\mu_b\| \, \rho_b \tag{8}$$

Where Π_U is the orthogonal projector onto U and $\rho_b = \|(I - \Pi_U)\mu_b\|/\|\mu_b\|$. Proof. Since $w \perp U$, $w^\top \mu_b = w^\top (I - \Pi_U)\mu_b$, and the Cauchy–Schwarz inequality gives the bound. The expected penalty that an invariant head can still assign to phenomenon b thus scales with the residual fraction ρ_b. Phenomena whose surface realisation coincides with an ignored dialect feature, such as a missing third-person -s, have small ρ_b, and no linear head can treat them differently from the dialect feature.

4.5. Model Selection

The ridge strength α is chosen by development quadratic weighted kappa (QWK) of the unmitigated head and then fixed for all methods. The strength of every mitigation method, including zero, is chosen with the selection rule of CAIR,

$$\lambda^\star = \mathrm{argmin}_{\lambda\in\Lambda} \, |\bar{\Delta}|^{\mathrm{dev}}(\lambda) \quad \text{s.t.} \quad \mathrm{QWK}^{\mathrm{dev}}(\lambda) \ge \mathrm{QWK}^{\mathrm{dev}}(0) - \delta \tag{9}$$

With $\delta = 0.01$, $\Lambda = \{0, 0.03, 0.1, 0.3, 1, 3, 10, 30, 100, 300\}$ and $|\bar{\Delta}|^{\mathrm{dev}}$ the mean absolute development shift over the varieties in scope (both varieties, or AAE only for the VS-

CAIR Conventions head).

5. Experimental Setup

5.1. Data

Experiments use the ASAP essays [1] under the sampling protocol of the CAIR study [17]. Prompts 3–6 are source-dependent responses of grades 8 and 10 scored for content on 0–3 (prompts 3, 4) or 0–4 (prompts 5, 6); 1,000 responses per prompt were sampled with stratification by score. Prompt 7 is a grade-7 narrative with four traits (Ideas, Organization, Style, Conventions) scored 0–3 by two raters; the target is the sum of both ratings (0–6), and all 1,569 essays are used. Table I reports corpus statistics, human agreement and edit densities. Ideas, Organization, Style and the content scores of prompts 3–6 are treated as invariant targets; Conventions, which the rubric defines with respect to Standard English grammar, usage and spelling, is construct-relevant.

Table I. Data and Counterfactual Statistics

Prompt	Type	N	Scale	Words	hQWK	AAE	L2	Changed (%)
3	Source	1,000	0–3	108.7	0.783	2.71	3.76	81/95
4	Source	1,000	0–3	95.1	0.848	4.09	4.03	91/93
5	Source	1,000	0–4	124.5	0.766	4.19	4.16	95/96
6	Source	1,000	0–4	155.6	0.778	1.44	4.16	86/98
7	Narrative	1,569	4×(0–6)	168.4	0.596	1.51	3.30	79/96

hQWK: agreement between the two human raters (mean over the four traits for prompt 7). AAE, L2: counterfactual edits per 100 words. Changed: share of AAE/L2 variants that differ from the original.

5.2. Protocol, Metrics and Baselines

Each prompt is split into five stratified folds (seed 2024); fold k serves as test set, fold k+1 as development set and the remaining three folds as training set (60/20/20), so that every essay is tested exactly once. Counterfactual variants of training essays provide the pairs used by all mitigation methods; development variants serve only for model selection; all metrics are computed on test essays and their variants. We report QWK, $\bar{\Delta}_v$, FR_v and $MAIS_v$; significance is assessed with the Wilcoxon signed-rank test over the 20 prompt–fold units of prompts 3–6. Baselines are CDA with weight $\gamma \in \{0.25, 0.5, 1, 2\}$ [24], SENT-DEBIAS with $k \in \{1, 2, 4, 8, 16, 32\}$ removed principal directions [25] and INLP with $T \in \{1, 2, 4, 8, 16\}$ projections of least-squares linear classifiers that separate originals from variants [26]. On prompt 7 we further compare construct-map policies for the Conventions head: the trait-level map of CAIR (Conventions unconstrained), a uniform map (Conventions invariant to both varieties), VS-CAIR (invariant to AAE only) and VS-CAIR+SA (invariant to AAE, anchored to the unmitigated L2 shifts).

5.3. Implementation

Representations were computed with PyTorch in bfloat16 on a single CPU core with AMX support at about 2,100 subword tokens per second. All heads and mitigation methods use closed-form solvers in NumPy with Gram matrices computed once per fold, and the generators rely on spaCy parses and lemminflect inflections. Random seeds are fixed for sampling, counterfactual generation (one seed per essay and variety) and fold assignment; code, configurations and per-essay predictions are retained for verification.

6. Results

6.1. Comparative Audit Across Representation Families

Table II. Counterfactual Audit of Unmitigated Scorers (Mean over Prompts and Folds)

Family	QWK	$\bar{\Delta}_{AAE}$	$\bar{\Delta}_{L2}$	FR_{AAE}	FR_{L2}	$MAIS_{AAE}$	$MAIS_{L2}$
Content scores, prompts 3–6							
Word TF-IDF	0.671	−0.027	−0.030	6.6	6.9	0.073	0.073
Char TF-IDF	0.718	−0.021	−0.004	4.0	4.3	0.040	0.044
Static vectors	0.489	−0.029	−0.056	6.0	9.2	0.059	0.099
WordLlama	0.452	−0.012	−0.025	4.7	9.7	0.055	0.119
RoBERTa	0.733	−0.047	−0.021	6.2	6.0	0.066	0.065
Ideas/Organization/Style, prompt 7							
Word TF-IDF	0.638	−0.016	−0.026	5.8	7.8	0.043	0.061
Char TF-IDF	0.701	−0.004	−0.007	3.0	4.8	0.025	0.039
Static vectors	0.569	+0.014	−0.073	4.4	12.8	0.035	0.093
WordLlama	0.524	+0.007	−0.036	3.7	9.1	0.026	0.069
RoBERTa	0.719	−0.010	−0.041	3.5	8.3	0.027	0.064

$\bar{\Delta}$: standardised mean shift, Eq. (1); FR: flip rate in %, Eq. (2); MAIS: mean absolute individual shift, Eq. (3). Negative $\bar{\Delta}$ means that the variant receives a lower score.

Table II reports the unmitigated scorers. On the content scores of prompts 3–6, RoBERTa reaches the highest agreement (QWK 0.733), followed by character n-gram TF-IDF (0.718), word n-gram TF-IDF (0.671), static word vectors (0.489) and WordLlama (0.452). In all five families both varieties lower the predicted score on average, and L2 variants move scores more than AAE variants in 3 of 5 families. The mean shifts are small in group-level terms: the largest, 0.056 standard deviations (static word vectors), is far below the 0.15 threshold used to flag subgroup differences [7]. Individual outcomes tell a different story. Between 4.0% (character n-gram TF-IDF) and 6.6% (word n-gram TF-IDF) of the students would receive a different rounded score under

the AAE rewrite, and between 4.3% and 9.7% under the L2 rewrite, although the content of the response is unchanged.

Group-level and individual-level indicators also rank the families differently. Ordered by the summed absolute mean shift the families rank character n-gram TF-IDF < WordLlama < word n-gram TF-IDF < RoBERTa < static word vectors, whereas ordered by the summed flip rate they rank character n-gram TF-IDF < RoBERTa < word n-gram TF-IDF < WordLlama < static word vectors. WordLlama, for instance, has the second-smallest mean shifts but the second-largest flip rates, because its individual shifts are large and partly cancel. For RoBERTa the mean absolute individual shift exceeds the absolute mean shift by a factor of 1.4 for AAE and 3.1 for L2: individual scores move in both directions and partly cancel in the group mean. The invariant traits of prompt 7 show the same picture at a larger scale, with L2 flip rates between 4.8% (character n-gram TF-IDF) and 12.8% (static word vectors). Auditing a scorer only through mean subgroup differences would therefore accept models that change the reported scores of a substantial minority of students for reasons the rubric excludes. Horizon-aware group-conditional conformal prediction provides an informative statistical perspective on auditing reliability separately across groups rather than relying exclusively on aggregate performance [32].

6.2. Where Bias Enters the Contextual Model

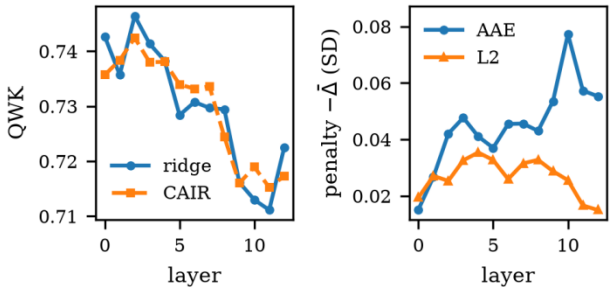


Figure 1. Layer-wise profile of RoBERTa (content scores, prompts 3–6, mean over prompts and folds). Left: QWK of ridge and CAIR heads trained on single layers (0 = embedding output). Right: penalty $-\bar{\Delta}$ of the unmitigated heads for AAE and L2 variants.

Fig. 1 trains a head on each RoBERTa layer separately. Agreement is highest in the lower layers (QWK 0.746 at layer 2 and 0.743 at the embedding output) and declines to 0.711–0.722 in the top three layers. Sensitivity to dialect moves in the opposite direction: the AAE penalty grows from −0.015 at the embedding output to −0.077 at layer 10, and the mean flip rate from 2.9% to 8.4% at layer 10. The L2 penalty instead peaks in the middle layers (−0.035 at layer 4) and falls to −0.015 at layer 12. Contextualisation therefore amplifies the sensitivity of the scorer to dialect features rather than washing it out, in line with the covert dialect prejudice reported for contextual language models [14]. CAIR lowers the flip rate at 13 of 13 layers with per-layer QWK changes between −0.007 and +0.006, so the regulariser is effective wherever the head is placed.

6.3. Phenomenon-Level Attribution

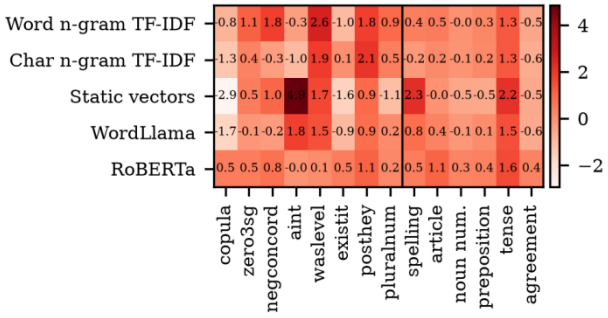


Figure 2. Per-edit potency $-100 \cdot \pi f$ of each phenomenon, Eq. (4), for the unmitigated heads of each family (prompt 7, mean over the invariant traits Ideas, Organization and Style). Larger values mean that a single edit lowers the score more; the vertical line separates the eight AAE features from the six L2 groups.

Fig. 2 decomposes the shifts of the invariant prompt-7 traits into the fourteen phenomena. The most potent phenomena differ by family: was-levelling and possessive they for word n-gram TF-IDF; possessive they and was-levelling for character n-gram TF-IDF; ain't and misspellings for static word vectors; ain't and tense errors for WordLlama; tense errors and possessive they for RoBERTa. Averaged over phenomena, one L2 edit costs $0.71 \cdot 10^{-2}$ and one AAE edit $0.45 \cdot 10^{-2}$ standard deviations in RoBERTa, against $0.14 \cdot 10^{-2}$ and $0.29 \cdot 10^{-2}$ for character n-grams. The decomposition is close to additive for RoBERTa: the sum of the phenomenon-level shifts of an essay correlates with its full-variety shift at $r = 0.98$ (AAE) and $r = 0.95$ (L2), and recovers 100% and 98% of the mean shift. The families also differ in how consistently they react. RoBERTa penalises 13 of the 14 phenomena, whereas the n-gram, static-vector and WordLlama heads penalise only 7–9 and reward some edits: copula absence, for example, raises their scores, an artefact of averaged and bag-of-words representations rather than a judgement about quality. Phenomenon-level potencies are thus a usable diagnostic: they identify which features a given representation penalises, which a group-level comparison cannot do.

6.4. Mitigation Across Families

Table III. Mitigation of Holistic Content Scores (Prompts 3–6)

Family	Method	QWK	$\lvert\bar{\Delta}\rvert$	FR (%)	MAIS
Word TF-IDF	Ridge	0.671	0.030	6.8	0.073
	CDA	0.678	0.017	6.9	0.076
	SENT-DEBIAS	0.671	0.013	5.6	0.062
	INLP	0.670	0.027	6.7	0.072
	CAIR	0.666	0.007	2.8	0.031
Char TF-IDF	Ridge	0.718	0.016	4.2	0.042
	CDA	0.715	0.013	4.1	0.044
	SENT-DEBIAS	0.721	0.003	2.4	0.028
	INLP	0.717	0.012	3.8	0.040
	CAIR	0.711	0.002	1.7	0.018
Static vectors	Ridge	0.489	0.044	7.6	0.079
	CDA	0.493	0.028	7.4	0.077
	SENT-DEBIAS	0.485	0.016	5.3	0.061
	INLP	0.495	0.028	6.4	0.070
	CAIR	0.484	0.016	4.9	0.052
WordLlama	Ridge	0.452	0.026	7.2	0.087
	CDA	0.453	0.023	7.3	0.088
	SENT-DEBIAS	0.454	0.016	6.3	0.075
	INLP	0.453	0.023	7.0	0.086
	CAIR	0.456	0.017	5.2	0.061
RoBERTa	Ridge	0.733	0.038	6.1	0.065
	CDA	0.730	0.017	5.3	0.061
	SENT-DEBIAS	0.731	0.004	4.2	0.047
	INLP	0.731	0.034	5.9	0.063
	CAIR	0.735	0.006	3.2	0.034

$\lvert\bar{\Delta}\rvert$, FR and MAIS are averaged over the AAE and L2 varieties and over the 20 prompt–fold units.

Table III compares the mitigation methods on the content scores. CAIR attains the lowest flip rate in 5 of 5 families, the lowest mean absolute individual shift in 5 and the lowest absolute mean shift in 3. Relative to the unmitigated heads it lowers the flip rate by 58% (word n-gram TF-IDF), 60% (character n-gram TF-IDF), 35% (static word vectors), 28% (WordLlama), 47% (RoBERTa), while QWK changes by between −0.007 and +0.004. For RoBERTa, the flip rate of CAIR is significantly lower than that of the unmitigated head ($p < 0.001$), CDA ($p < 0.001$), SENT-DEBIAS ($p = 0.005$), INLP ($p < 0.001$) over the 20 prompt–fold units, and its QWK does not differ significantly from the unmitigated head ($p = 0.084$). The task-agnostic methods behave differently across families: SENT-DEBIAS removes much of the mean shift where a few directions carry the variety signal, but leaves larger individual shifts, whereas INLP removes directions that separate originals from variants without necessarily touching the scoring direction. Acting on the head rather than on the representation is what makes the effect of CAIR consistent across representations that encode variation very differently.

6.5. Construct-Map Policies for Trait Feedback

Table IV. Construct-Map Policies on Prompt 7 (RoBERTa)

Method	QWK_{inv}	$\lvert\bar{\Delta}_{L2}\rvert_{inv}$	$FR_{L2,\,inv}$	QWK_C	$\bar{\Delta}_{AAE,\,C}$	$\bar{\Delta}_{L2,\,C}$	Ret. (%)
Ridge	0.719	0.041	8.3	0.696	−0.053	−0.134	100
CDA	0.719	0.030	9.0	0.692	−0.042	−0.104	78
SENT-DEBIAS	0.716	0.006	6.2	0.695	−0.040	−0.108	81
INLP	0.719	0.040	8.2	0.692	−0.050	−0.128	96
CAIR, uniform map	0.716	0.009	4.3	0.692	−0.030	−0.089	66
CAIR, trait map	0.716	0.009	4.3	0.696	−0.053	−0.134	100
VS-CAIR (ours)	0.716	0.009	4.3	0.691	−0.013	−0.119	89
VS-CAIR+SA (ours)	0.716	0.009	4.3	0.693	−0.011	−0.131	98

inv: mean over the invariant traits Ideas, Organization and Style; C: Conventions. Ret.: Conventions L2 shift as a percentage of the unmitigated shift (construct-relevant sensitivity retained).

Table IV turns to the trait heads of prompt 7. Without mitigation the Conventions head lowers L2 variants by 0.134 standard deviations, the rubric-consistent response to additional grammar and spelling errors, and AAE variants by 0.053. On the invariant traits CAIR reduces the mean absolute L2 shift from 0.041 to 0.009 and the L2 flip rate from 8.3% to 4.3%, at a mean QWK of 0.716 versus 0.719. For Conventions, the trait-level map leaves the head unconstrained and keeps both penalties; the uniform map retains only 66% of the learner-error penalty; and the task-agnostic baselines retain 78% (CDA), 81% (SENT-DEBIAS) and 96% (INLP). VS-CAIR removes 75% of the AAE penalty

but keeps only 89% of the L2 penalty, and sensitivity anchoring restores it to 98% while removing 80% of the AAE penalty, with Conventions QWK of 0.693 against 0.696 for the unmitigated head.

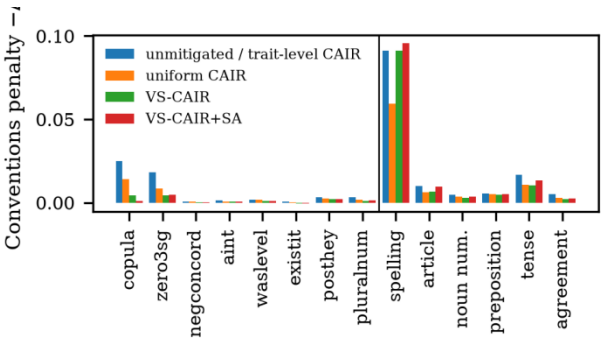


Figure 3. Penalty $-\bar{\Delta}$ of the RoBERTa Conventions head under single-phenomenon counterfactuals (prompt 7). The trait-level CAIR map leaves Conventions unconstrained and coincides with the unmitigated head.

Fig. 3 shows where the remaining conflict lies. VS-CAIR suppresses the Conventions penalty for the AAE features, and the learner-error penalty that shrinks most is the one whose surface form coincides with a dialect feature. In RoBERTa space the mean displacement of zero third-person -s has a cosine of 0.83 with that of L2 agreement errors, and plural -s absence after numerals one of 0.63 with noun-number errors (Fig. 4). Projecting each L2 group displacement onto the orthogonal complement of the AAE displacements leaves residual fractions ρ_b from 0.52 (agreement) to 0.94 (article errors). Agreement errors, the most entangled group, retain only 43% of their Conventions penalty under VS-CAIR and 51% with sensitivity anchoring, whereas anchoring restores 79–105% for the other groups. Proposition 1 bounds rather than predicts the retained sensitivity: a finite λ also shrinks the penalties of weakly entangled groups such as tense errors, so the rank correlation between ρ_b and retention is only 0.31. The group that the bound constrains most, however, is the one that no setting recovers. A missing third-person -s is at once a dialect feature and a learner error, and a linear head cannot treat the two readings differently.

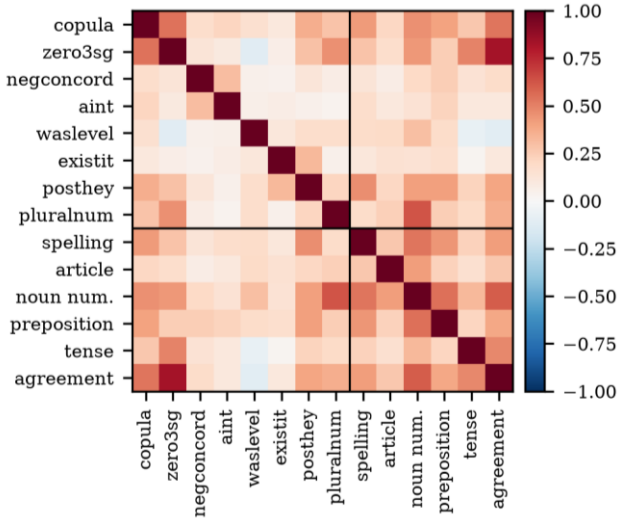


Figure 4. Cosine similarity between the mean RoBERTa displacements of the fourteen phenomena (prompt 7). Lines separate AAE features (top/left) from L2 groups.

6.6. Validity of the Synthetic Learner Direction

Synthetic counterfactuals are useful only if they move representations in the direction of authentic variation. We injected L2 errors into the first JFLEG correction of each learner sentence and compared the resulting displacement with the displacement between the authentic learner sentence and the same correction. In RoBERTa space the two mean displacements have a cosine of 0.52, and the ten leading directions of the synthetic displacements capture 29.0% of the energy of the authentic learner–correction differences, against 0.6% for a random ten-dimensional subspace.

7. Discussion

Three findings matter for the use of LLM representations in assessment. First, group-level and individual-level evidence diverge for every family studied: mean shifts far below operational screening thresholds coexist with flip rates of several per cent, so counterfactual flip rates belong in the evidence reported for automated scores next to agreement and subgroup statistics [7, 9]. Second, the phenomenon-level audit is comparative in a useful sense. The families differ in which phenomena they penalise, and the layer-wise profile shows that contextual layers amplify rather than remove sensitivity to dialect features, and the contextual model penalises almost every deviation from Standard English. Choosing a representation therefore also chooses a pattern of linguistic sensitivity, which can now be measured before deployment without demographic data. Third, construct maps are policy decisions. Whether a Conventions score should penalise dialect features is a question for assessment specialists; the variety-selective map makes the answer explicit, sensitivity anchoring implements it without silently erasing learner-error feedback, and Proposition 1 states what no linear scorer can deliver.

The study has limitations. The counterfactuals are rule-based approximations of AAE and learner English; they reproduce published generator statistics and align with authentic learner corrections, but they do not capture the full variability of either variety, and AAE rules applied to essays written by other students are not samples of AAE writing. All scorers are linear heads on frozen representations; fine-tuned encoders and prompted generative LLMs were outside the computational budget of this study (a single CPU core) and may behave differently. The data come from five ASAP prompts of grades 7–10, and the construct map for prompt 7 reflects our reading of its rubric.

8. Conclusion

This paper compared linguistic bias and its mitigation across five families of essay representations, from n-grams to a contextual LLM, using demographic-free counterfactual essays grounded in AAE grammar and learner-corpus statistics. Every family changes individual scores for reasons the rubric excludes, even when group-level differences look negligible; phenomenon-level potencies and layer profiles show where these changes come from. Construct-aligned invariance reduces them consistently across families, and the proposed variety-selective map with sensitivity anchoring separates dialect invariance from learner-error feedback up to a geometric limit that we characterise. Future work will extend the audit to fine-tuned and generative scorers, to further varieties of English and to multi-dimensional argumentative rubrics.

References

[1] Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. Assessing Writing, 20, 53–76. https://doi.org/10.1016/j.asw.2013.04.001

[2] Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1882–1891). https://doi.org/10.18653/v1/D16-1193

[3] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT) (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

[4] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692.

[5] Mayfield, E., & Black, A. W. (2020). Should you fine-tune BERT for automated essay scoring? In Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications (BEA) (pp. 151–162). https://doi.org/10.18653/v1/2020.bea-1.15

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

[7] Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. Educational Measurement: Issues and Practice, 31(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

[8] Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. Applied Measurement in Education, 25(1), 27–40. https://doi.org/10.1080/08957347.2012.635502

[9] Loukina, A., Madnani, N., & Zechner, K. (2019). The many dimensions of algorithmic fairness in educational applications. In Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA) (pp. 1–10). https://doi.org/10.18653/v1/W19-4401

[10] Yang, K., Raković, M., Li, Y., Guan, Q., Gašević, D., & Chen, G. (2024). Unveiling the tapestry of automated essay scoring: A comprehensive investigation of accuracy, fairness, and generalizability. Proceedings of the AAAI Conference on Artificial Intelligence, 38(20), 22466–22474. https://doi.org/10.1609/aaai.v38i20.30254

[11] Schaller, N.-J., Ding, Y., Horbach, A., Meyer, J., & Jansen, T. (2024). Fairness in automated essay scoring: A comparative analysis of algorithms on German learner essays from secondary education. In Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA) (pp. 210–221).

[12] Wang, J., Ge, Y., Liu, J., & Jiang, B. (2024). Early warning of liquidity stress in U.S. banks with seasonally aligned frailty hazards and delayed-feedback conformal alarms. Journal of Trends in Finance and Economics, 1(1), 85–100.

[13] Ziems, C., Chen, J., Harris, C., Anderson, J., & Yang, D. (2022). VALUE: Understanding dialect disparity in NLU. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 3701–3720). https://doi.org/10.18653/v1/2022.acl-long.258

[14] Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). AI generates covertly racist decisions about people based on their dialect. Nature, 633(8028), 147–154. https://doi.org/10.1038/s41586-024-07856-5

[15] Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. Patterns, 4(7), 100779. https://doi.org/10.1016/j.patter.2023.100779

[16] Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). 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) (pp. 5454–5476). https://doi.org/10.18653/v1/2020.acl-main.485

[17] Fang, Q. (2024). Bias-aware large language models for equitable educational feedback and assessment. Trends in Social Science and Humanities Research, 2(1), 41–50. https://doi.org/10.61784/tsshr4253

[18] Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 4593–4601). https://doi.org/10.18653/v1/P19-1452

[19] Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. E., & Smith, N. A. (2019). Linguistic knowledge and transferability of contextual representations. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT) (pp. 1073–1094). https://doi.org/10.18653/v1/N19-1112

[20] Napoles, C., Sakaguchi, K., & Tetreault, J. (2017). JFLEG: A fluency corpus and benchmark for grammatical error correction. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL) (Vol. 2, pp. 229–234).

[21] Flor, M., Fried, M., & Rozovskaya, A. (2019). 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 (BEA) (pp. 76–86). https://doi.org/10.18653/v1/W19-4407

[22] Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. Advances in Neural Information Processing Systems, 30, 4066–4076.

[23] Garg, S., Perot, V., Limtiaco, N., Taly, A., Chi, E. H., & Beutel, A. (2019). Counterfactual fairness in text classification through robustness. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES) (pp. 219–226). https://doi.org/10.1145/3306618.3317950

[24] Zmigrod, R., Mielke, S. J., Wallach, H., & Cotterell, R. (2019). Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 1651–1661). https://doi.org/10.18653/v1/P19-1161

[25] Teng, D. (2025). RSRBench: Reproducibility and performance benchmarking for regional seismic risk workflows. Academic Journal of Architecture and Civil Engineering, 3(1), 40–52.

[26] Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., & Goldberg, Y. (2020). Null it out: Guarding protected attributes by iterative nullspace projection. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 7237–7256). https://doi.org/10.18653/v1/2020.acl-main.647

[27] Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength natural language processing in Python. Zenodo. https://doi.org/10.5281/zenodo.1212303

[28] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics,

12(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

[29] Touvron, H., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288.

[30] Miller, D. L. (2024). WordLlama: Recycled token embeddings from large language models. https://github.com/dleemiller/wordllama

[31] Wang, Z., et al. (2026). FairPromote: Explainable and Fairness-Aware Talent Promotion Prediction via Adversarial Debiasing and SHAP-Based Interpretation. IEEE Access. https://doi.org/10.1109/ACCESS.2026.3692578

[32] Zou, J., Feng, E., Li, P., & Wang, G. (2026). Horizon-aware group-conditional conformal prediction: Post-hoc shrinkage for sparse multi-horizon coverage auditing. Machine Learning with Applications, 26, 101038. https://doi.org/10.1016/j.mlwa.2026.101038