A Teacher-Centered Evaluation of AI-Assisted Educational Assessment: Rubric Consistency, Feedback Quality, and Fairness Across Student Groups

Qianyi Fang

Johns Hopkins University, MD, USA

qfang10@jh.edu

Abstract. AI-assisted assessment is increasingly used to score student writing and to generate formative feedback, yet it is still validated mainly by aggregate agreement with consensus scores. For classroom adoption the decisive question is teacher-centered: does the AI apply the rubric the way a competent teacher does, diagnose the same weaknesses, and treat every student group alike? We propose a Teacher-Centered Evaluation (TCE) framework with four lenses—teacher-relative rubric consistency, rubric-guided counterfactual tests, feedback diagnostic validity, and score-conditional fairness—complemented by a teacher-in-the-loop routing policy with group-audited conformal intervals. TCE is instantiated on two public corpora released by the research team behind the ASAP 2.0 benchmark: PERSUADE 2.0 (5,531 argumentative essays with holistic scores, 55,068 expert-rated discourse elements, and student demographics) and ELLIPSE (6,462 essays double-scored by 27 trained raters on seven rubric traits). Five AI scoring engines are audited. The strongest engine reaches a quadratic weighted kappa (QWK) of 0.870 and, in a teacher-substitution test, agrees with a teacher almost exactly as well as a second teacher does (ΔQWK = −0.016). TCE nevertheless exposes failures that agreement statistics hide: scores are blind to sentence order, rise by 0.27 points when a paragraph is duplicated, and fall no more when evidence is deleted than when equally long non-evidence text is removed; analytic trait scores collapse into a single dimension (inter-trait r = 0.87 vs. 0.61 for teachers); feedback misses 70% of the weaknesses that experts flag in English-learner writing; and conditioning on the teacher's score uncovers subgroup gaps that raw statistics mask. Routing 20% of essays to teachers with a learned deferral model raises QWK to 0.907, reduces the worst subgroup bias by 31%, and nearly halves the English-learner agreement gap.

Keywords: Automated essay scoring, AI-assisted assessment, teacher-centered evaluation, rubric consistency, formative feedback, algorithmic fairness, learning to defer, conformal prediction.

1. Introduction

Scoring essays and writing feedback are among the most time-consuming duties of secondary-school teachers, and they are also among the most consequential: holistic scores inform placement and accountability, while formative feedback shapes how students revise [16]. Automated essay scoring (AES) has supported large-scale testing for two decades [9], and a national demonstration reported that commercial engines approximate human raters on most agreement measures [6]. Neural scorers [10, 11] and, more recently, large language models (LLMs) [12–15] have carried these capabilities into classroom tools that also draft feedback [17, 18]. The ASAP 2.0 benchmark, which pairs human scores with student demographics, signals that AI graders are now expected to be evaluated for equity as well as accuracy [1].

The dominant validation practice nevertheless remains model-centered. Engines are tuned and reported by their agreement with a consensus score, typically through QWK and the standardized mean difference (SMD) [5, 28]. Such numbers answer whether a model reproduces a label, not whether it does a teacher's job, and four gaps follow. First, teachers disagree with one another, so agreement with a consensus label is not comparable with the agreement a substitute teacher would achieve [20]. Second, high agreement can be reached through construct-irrelevant proxies such as length [7], and scoring models accept inputs that no teacher would reward [27]. Third, feedback quality is usually judged holistically [17] rather than verified against expert diagnoses of specific

argumentative elements. Fourth, fairness audits compare raw group means [8, 19, 21], which confounds scoring bias with genuine differences between the score distributions of student groups. In addition, the operational question of how much teacher review an AI grader needs, and for which essays, is rarely answered with evidence [15, 23].

This paper makes the teacher the reference point of evaluation. We ask whether an AI grader behaves like a competent colleague: whether it agrees with a teacher as much as another teacher does, changes its score when—and only when—the rubric says it should, flags the weaknesses an expert would flag, and treats students from different groups alike once the quality of their writing is held constant. The contributions are as follows.

A Teacher-Centered Evaluation (TCE) framework with four lenses and dedicated metrics: a teacher-substitution test with rater-population placement and a trait discrimination index (TDI); rubric-guided counterfactual tests with a length-controlled construct sensitivity index (CSI); budget-matched diagnostic feedback flags with a false-criticism rate; and score-conditional, prompt-stratified SMD.

A teacher-in-the-loop routing policy based on learned deferral, combined with split-conformal score intervals whose coverage is audited across student groups without using demographics at inference time.

A comprehensive audit of five AI engines on two public corpora with 11,993 essays, 27 trained raters, and 55,068 expert-rated discourse elements, revealing failure modes that remain invisible under standard agreement criteria.

Section II reviews related work, Section III presents the framework, Section IV describes the experimental setup, Section V reports the results, Section VI discusses implications, and Section VII concludes.

2. Related Work

2.1 Automated Essay Scoring and Its Validation

Feature-based engines such as e-rater combine measures of grammar, usage, mechanics, organization, and lexical sophistication [9]. The ASAP demonstration compared nine scoring systems with human raters and reported comparable performance on most measures [6], a conclusion contested on the grounds that much of the agreement can be explained by essay length [7]. Neural models learn representations directly from text [10], and surveys document steady gains on ASAP-style benchmarks [11]. The framework of Williamson et al. [5]—a minimum QWK of 0.70, a degradation of at most 0.10 relative to human–human agreement, and bounded SMD—remains the operational standard and is implemented in open tools such as RSMTool [28]. Loukina et al. [20] showed that rater noise distorts agreement-based comparisons, which motivates evaluating engines against individual raters rather than consensus labels. Robustness studies further show that content-scoring models accept shuffled or nonsensical answers [27], in line with the behavioral-testing paradigm of CheckList [26].

2.2 LLM-Based Scoring and Feedback

Mizumoto and Eguchi [12] scored the 12,100 TOEFL11 essays with GPT-3 and reported useful but imperfect accuracy. Pack et al. [13] evaluated the cross-sectional and longitudinal validity and reliability of four LLMs for scoring English-learner writing, and Tate et al. [14] examined whether generative AI can provide useful holistic essay scores. Human–AI collaborative scoring has been proposed to combine model efficiency with rater judgment [15]. Feedback research emphasizes that effective feedback clarifies the goal, the current progress, and the next step [16]. Steiss et al. [17] found that trained evaluators gave higher-quality feedback than ChatGPT on all criteria except criteria-based feedback, whereas Dai et al. [18] found ChatGPT feedback more detailed and fluent than instructor feedback. These studies rate feedback with feedback-quality rubrics; they do not test whether the feedback diagnoses the argumentative elements that experts judge ineffective.

2.3 Fairness in Automated Assessment

Bridgeman et al. [8] found that human and machine scores were similar for most subgroups but documented notable exceptions by gender, ethnicity, and country. Loukina et al. [19] showed that different fairness definitions for automated scoring can conflict and that total fairness may be unattainable. Baker and Hawn [21] reviewed evidence of algorithmic bias in education across race/ethnicity, gender, nationality, socioeconomic status, and disability. In uncertainty quantification, Romano et al. [22] proposed conformal intervals with equalized coverage across protected groups. AES fairness audits predominantly compare raw mean differences; we show that conditioning on the teacher's score, and on the prompt, changes which groups are flagged.

2.4 Human–AI Collaboration and Deferral

Learning to defer trains a model to pass difficult cases to a downstream expert and can improve both the accuracy and the fairness of the joint system [23]; consistent surrogate losses were derived by Mozannar and Sontag [24], and selective classification provides the reject-option foundation [25]. We adapt deferral to essay scoring, where the expert is the teacher whose judgment defines the rubric, and evaluate routing policies jointly on agreement and subgroup bias.

3. Teacher-Centered Evaluation Framework

Fig. 1 summarizes TCE. Let x denote an essay, $y \in \{1, \ldots, K\}$ a teacher's rubric score, f(x) the continuous output of an AI engine, and $\hat{y} = T(f(x))$ its discrete score obtained with ordered thresholds T. Every statistic is computed out of sample: engines are trained with cross-fitting, and the thresholds applied to a fold are learned from the other folds only. We refer to trained human raters as teachers because they perform the teacher's role of applying the rubric.

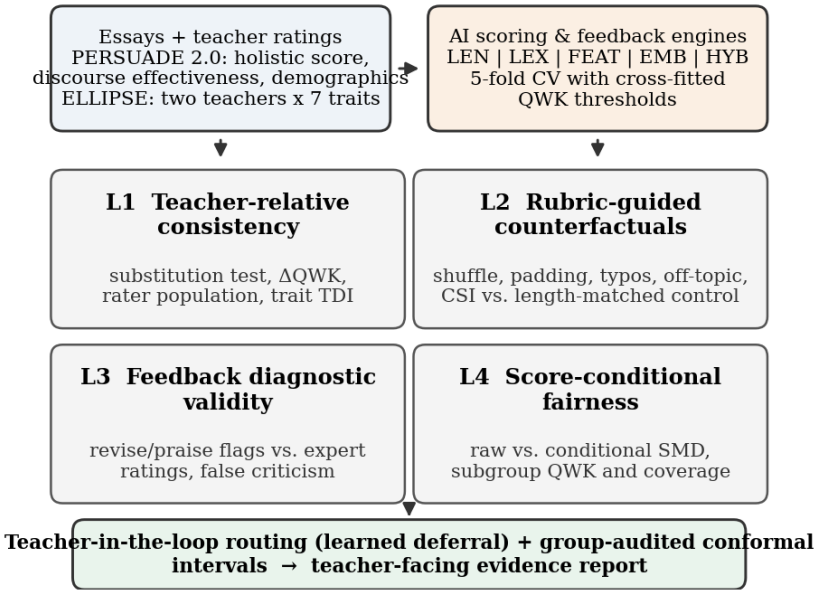


Figure 1. The Teacher-Centered Evaluation (TCE) framework. Four lenses compare an AI engine with teachers on the rubric, on the counterfactual behavior the rubric implies, on feedback diagnoses, and on group-level treatment; a routing layer decides which essays a teacher should review.

3.1 Lens 1: Teacher-Relative Rubric Consistency

When an essay is scored independently by teachers A and B, the relevant benchmark for an AI grader is the agreement $\kappa(y_A, y_B)$ between the two teachers, where κ denotes QWK. In the teacher-substitution test, engine f_A is trained only on A's ratings of other essays and mapped to A's scale; it

is then compared with teacher B, exactly the comparison that A faces. Averaging both directions gives

$$\Delta\kappa = \tfrac{1}{2}[\kappa(\hat{y}_A, y_B) + \kappa(\hat{y}_B, y_A)] - \kappa(y_A, y_B) \quad (1)$$

Where $\Delta\kappa \geq -0.10$ satisfies the degradation criterion of [5] and $\Delta\kappa \geq 0$ indicates that the AI is as interchangeable with a teacher as another teacher is. To place the AI inside the teacher population, each rater r is characterized by severity $s_r = E[y_r - y_c]$ relative to co-raters c and by QWK with co-raters; the AI receives the same statistics. For analytic rubrics with traits t = 1, …, M, a useful grader must also discriminate between traits. With $\rho(S_t, B_{t'})$ the correlation between trait t of source S and trait t′ of teacher B, the trait discrimination index is

$$TDI_t(S) = \rho(S_t, B_t) - (M-1)^{-1} \Sigma_{t' \neq t} \rho(S_t, B_{t'}) \quad (2)$$

A source whose TDI approaches zero carries no trait-specific information, so its analytic profile cannot support diagnostic feedback.

3.2 Lens 2: Rubric-Guided Counterfactual Tests

Agreement on naturally occurring essays cannot reveal whether an engine uses the construct defined by the rubric. Inspired by behavioral testing [26], we apply edits whose effect on a teacher's rubric judgment is unambiguous (Table IV): sentence shuffling with paragraph and sentence counts preserved (organization deteriorates), verbatim duplication of the longest paragraph (repetition earns no credit), adjacent-character transpositions in 10% of the words with at least four letters (conventions deteriorate), replacement of every Evidence span by a length-matched Evidence span written for another prompt (relevance deteriorates), and deletion of expert-annotated Evidence or Counterclaim/Rebuttal spans (support and counterargument are lost). Deletions confound content with length, so each targeted deletion is paired with a control that removes random non-target sentences of approximately the same length. With $\Delta = f(x') - f(x)$ and w the number of removed words, the construct sensitivity index is

$$CSI = 100 \cdot E[\Delta_{tgt}]/E[w_{tgt}] - 100 \cdot E[\Delta_{ctl}]/E[w_{ctl}] \quad (3)$$

A rubric-faithful engine yields CSI < 0: removing the evidence that supports an argument should cost more than removing the same amount of other text.

3.3 Lens 3: Feedback Diagnostic Validity

Formative feedback helps only if it points to the right parts of the essay [16]. PERSUADE 2.0 provides, for each argumentative element (Lead, Position, Claim, Counterclaim, Rebuttal, Evidence, Concluding Statement), an expert effectiveness rating of Ineffective, Adequate, or Effective [2, 3]. An AI feedback model estimates the probabilities p_I, p_A, p_E of an element and issues a revise flag when p_I exceeds a threshold and a praise flag when p_E exceeds another. To compare models fairly, thresholds are budget-matched: on the training folds they are set so that the flag rate equals the expert base rate. We report revise precision and recall, the false-criticism rate FCR = P(revise | Effective), the misplaced-praise rate P(praise | Ineffective), and the accuracy of structural feedback that asks a student to add a missing counterclaim or rebuttal. Feedback–score coherence is the Spearman correlation between an essay's mean element effectiveness and its holistic score.

3.4 Lens 4: Score-Conditional Fairness

For a student group g with n_g essays, the raw standardized mean difference [5] is

$$SMDg = (m_{AI,g} - m_{T,g}) / \sqrt{[(s^2_{AI,g} + s^2_{T,g})/2]} \quad (4)$$

Where m and s denote the mean and standard deviation of AI and teacher (T) scores. Raw SMD mixes scoring bias with the interaction between a group's score distribution and the engine's calibration, such as regression toward the mean or the scale expansion induced by threshold tuning.

TCE therefore compares each essay with essays that received the same teacher score. With m(h) the mean AI score of all essays with y = h and σ_T the teacher-score standard deviation,

$$SMD^{cond}_g = (1/n_g)\, \Sigma_{i \in g}\, [\hat{y}_i - m(y_i)] / \sigma_T \quad (5)$$

A stricter variant computes m(·) within prompt × score cells, which also removes prompt-level calibration offsets. We flag a subgroup when |SMD| exceeds 0.10 [5].

3.5 Teacher-in-the-Loop Routing and Coverage Audit

Let u(x) be a routing score. For a teacher-review budget d, the fraction d of essays with the largest u is routed to the teacher and receives the teacher's score, while the remaining essays keep ŷ. We compare random routing, ensemble disagreement (standard deviation of the engines' predictions), threshold margin (distance of f(x) to the nearest score threshold), and learned deferral, a gradient-boosted regressor that predicts |y − f(x)| from linguistic features, engine outputs, disagreement, and margin, trained with cross-fitting in the spirit of learning to defer [23, 24]. An oracle that routes the largest actual errors provides an upper bound. Policies are evaluated by system QWK, the worst-case |SMD| over all audited groups, and the agreement gap between English-language learners (ELL) and other students. To communicate the remaining uncertainty, TCE wraps the engine in split-conformal intervals [22]:

$$C(x) = f(x) \pm q \cdot \hat{\sigma}(x), \quad q = Q_{1-\alpha}\{|y_j - f(x_j)| / \hat{\sigma}(x_j)\} \quad (6)$$

Where $Q_{1-\alpha}$ is the $\lceil (n+1)(1-\alpha) \rceil / n$ empirical quantile over calibration essays. The marginal variant uses $\hat{\sigma} \equiv 1$, a Mondrian variant computes q per predicted-score bin, and the normalized variant uses the learned-deferral difficulty estimate as $\hat{\sigma}(x)$. No router or interval uses demographic information; demographics serve only to audit coverage across groups.

4. Experimental Setup

4.1 Datasets

The ASAP 2.0 benchmark [1] is distributed through a competition platform, so TCE is instantiated on two openly licensed (CC BY 4.0) corpora released by the same research team (Table I). PERSUADE 2.0 [2] contains argumentative essays by U.S. students in grades 6–12, scored holistically on a 1–6 scale and annotated with argumentative elements and their effectiveness [3]; like ASAP 2.0 [1], it provides student demographic data, here gender, race/ethnicity, ELL status, economic disadvantage, and disability status. We use a publicly redistributed four-prompt subset with 5,531 essays from grades 8–12, covering two independent and two source-based prompts and 55,068 rated elements. ELLIPSE [4] contains essays by English-language learners in grades 8–12, scored for overall proficiency and six analytic traits (cohesion, syntax, vocabulary, phraseology, grammar, conventions) on a 1–5 scale. Its raw-rating release preserves both independent ratings and anonymized rater identities; after removing unscorable ratings, 6,462 essays rated by 27 raters remain, and 23 of these raters scored at least 100 essays.

Table I. Datasets Used for the Teacher-Centered Evaluation

Property	PERSUADE 2.0 (subset)	ELLIPSE
Essays	5,531	6,462
Prompts	4 (2 independent, 2 source-based)	44 (independent)
Grades	8–12	8–12
Rubric	Holistic, 1–6	Overall + 6 traits, 1–5
Teacher ratings	1 adjudicated score per essay	2 independent ratings, 27 raters
Expert feedback labels	55,068 rated elements	—
Score distribution	3.6 / 18.3 / 32.7 / 25.8 / 14.4 / 5.1%	—
Audited groups	Gender, race/ethnicity, ELL, econ. disadv., disability	Gender, race/ethnicity, SES, grade

Score distribution lists the shares of scores 1–6. The ELL group comprises 537 PERSUADE essays.

4.2 AI Engines

Five engines span the main AES paradigms. LEN is an isotonic regression on log word count, a deliberately construct-poor reference [7]. LEX is a ridge regression on word 1–2-gram and character 3–5-gram TF-IDF features. FEAT is a LightGBM regressor [29] on 48 interpretable linguistic features—length, sentence and paragraph structure, lexical diversity (MTLD, type–token ratios), lexical sophistication from word-frequency norms, spelling-error rate, punctuation, cohesive connectives, argumentation and source-use markers, pronoun use, readability, and prompt topicality—plus the prompt identity. EMB is a multilayer perceptron on frequency-weighted averages of 300-dimensional static word vectors, a semantic engine without explicit length features. HYB stacks the FEAT features, 100 latent semantic components, and out-of-fold LEX and EMB predictions in a second LightGBM model, mirroring hybrid commercial engines. On ELLIPSE, LEX (in dual ridge form) and HYB are trained for each trait.

4.3 Protocol

PERSUADE engines are trained with five-fold cross-validation stratified by prompt and score, which yields an out-of-fold prediction for every essay; cross-prompt transfer is assessed with leave-one-prompt-out (LOPO) training. Continuous outputs are discretized with QWK-optimized thresholds fitted by coordinate search on the other folds' out-of-fold predictions. ELLIPSE uses the same five-fold design with fold-specific teacher targets. Counterfactual tests use the 1,107 essays of one held-out fold, scored by models trained on the remaining folds; 649 essays qualify for evidence deletion (Evidence covers at most 60% of the words and a length-matched control is feasible) and 461 for counterclaim/rebuttal deletion. The feedback model is a two-stage element classifier—multinomial logistic regression on word and character n-grams, element type, and position, followed by LightGBM on element descriptors, essay context, and the stage-one probabilities—trained with essay-grouped folds identical to those of the scorers. Confidence intervals are percentile bootstraps with 500–1,000 resamples, and the conformal miscoverage level is $\alpha = 0.10$.

5. Results

5.1 Agreement and Cross-Prompt Consistency

Table II shows that every engine except EMB exceeds the 0.70 QWK criterion [5] on the pooled corpus, and HYB reaches 0.870 [0.863, 0.875] with 99.1% adjacent agreement and a negligible overall SMD of 0.032. The length-only LEN engine already attains 0.755, and length features account for 44.5% of the mean absolute attribution of the FEAT engine, whose two most important features are essay length in characters and in words. EMB, which has no explicit access to length, is the weakest engine (0.676). Pooled QWK also benefits from between-prompt variance: within prompts, HYB ranges from 0.770 to 0.849 (Fig. 2b). When a prompt is unseen during training, QWK drops by 0.12 on average and the calibration shifts by a mean |SMD| of 0.35, from −0.63 for Cell phones at school to +0.42 for Electoral college, far beyond the 0.15 tolerance. The confusion matrix in Fig. 2a shows that HYB separates adjacent levels well in the middle of the scale but places 42% of the essays that teachers scored 1 on the next level.

Table II. Agreement of AI Engines with Teacher Scores on PERSUADE 2.0

Engine	QWK [95% CI]	Exact	SMD	SDR	R@1	R@6	In-p.	LOPO
LEN	0.755 [.740, .768]	0.496	0.002	1.08	0.14	0.78	—	—
LEX	0.834 [.825, .842]	0.570	0.012	1.04	0.44	0.79	0.748	0.684
FEAT	0.851 [.842, .858]	0.622	0.003	1.04	0.47	0.85	0.791	0.695
EMB	0.676 [.661, .690]	0.414	−0.010	1.07	0.38	0.67	—	—
HYB	0.870 [.863, .875]	0.641	0.032	1.04	0.50	0.83	0.810	0.689

SDR: ratio of AI to teacher score SD; R@k: recall of score k; In-p./LOPO: mean QWK within prompt and on an unseen prompt.

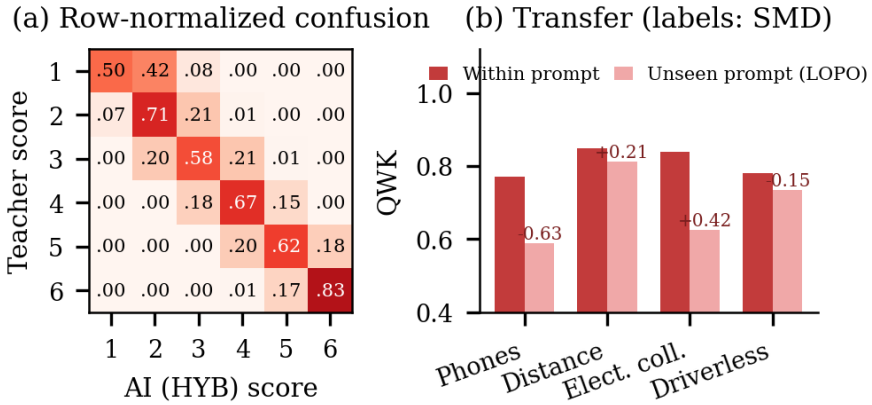


Figure 2. Rubric consistency on PERSUADE 2.0. (a) Row-normalized confusion matrix of HYB versus teacher scores. (b) Within-prompt versus unseen-prompt (LOPO) QWK; labels give the unseen-prompt SMD.

5.2 Teacher-Substitution Test

On ELLIPSE, two teachers agree with a QWK between 0.483 (vocabulary) and 0.591 (overall), as Table III shows, a reminder that rubric scoring is itself noisy. The substitute HYB engine, trained on one teacher's ratings, agrees with the other teacher at a practically identical level: $\Delta\kappa = -0.016$ [−0.030, −0.002] for the overall score and between −0.029 (grammar) and +0.032 (conventions) for the analytic traits, well inside the −0.10 tolerance [5]. Its correlation with the second teacher exceeds the teacher–teacher correlation for every trait (e.g., 0.622 vs. 0.592 overall), whereas its exact agreement is lower (0.554 vs. 0.592): the AI tracks the rank order of quality but places fewer essays on the teacher's exact level. A consensus-trained engine adds 0.011–0.020 QWK, which shows that claims of exceeding human agreement depend on the training target. Fig. 3b places the AI among the 23 teachers who rated at least 100 essays: its agreement (0.575) exceeds that of 7 teachers, and its severity (−0.003) sits near the center of a teacher distribution spanning −0.34 to +0.30. By the standard of rater interchangeability, the AI behaves like a typical member of the teacher pool.

The analytic picture differs. Teachers' trait ratings are correlated (mean inter-trait r = 0.61), but the AI's trait predictions are nearly collinear (r = 0.87). The TDI in Table III shows that the AI carries less trait-specific information than a teacher for cohesion (0.021 vs. 0.042), syntax (0.013 vs. 0.028), phraseology (0.019 vs. 0.028), and grammar (0.022 vs. 0.060), and more only for vocabulary and conventions, the traits with direct surface correlates such as word frequency and spelling. Analytic AI scores thus largely restate the overall score; a teacher who reads them as a diagnostic profile is misled precisely on the traits that require reading for meaning. Across subgroups, $\Delta\kappa$ for the overall score stays within ±0.05 for every gender, SES, grade, and racial/ethnic group; the largest degradation (−0.047) and the largest SMD (+0.144) occur for Asian/Pacific Islander students, whose essays the AI scores more leniently than teachers do.

TABLE III. Teacher-Substitution Test and Trait Discrimination on ELLIPSE (HYB)

Trait	κ_{TT}	κ_{sub}	$\Delta\kappa$	κ_{cons}	r_{TT}	r_{sub}	TDI_T	TDI_{AI}
Overall	0.591	0.575	−0.016	0.595	0.592	0.622	0.064	0.079
Cohesion	0.497	0.500	+0.002	0.519	0.499	0.540	0.042	0.021
Syntax	0.511	0.517	+0.006	0.532	0.513	0.551	0.028	0.013
Vocabulary	0.483	0.501	+0.017	0.519	0.484	0.553	0.020	0.028
Phraseology	0.510	0.519	+0.008	0.530	0.512	0.554	0.028	0.019
Grammar	0.531	0.503	−0.029	0.519	0.534	0.537	0.060	0.022
Conventions	0.518	0.550	+0.032	0.565	0.519	0.593	0.049	0.083

TT: teacher–teacher; sub: AI trained on one teacher vs. the other teacher; cons: consensus-trained AI. Mean inter-trait r: teachers 0.61, AI 0.87.

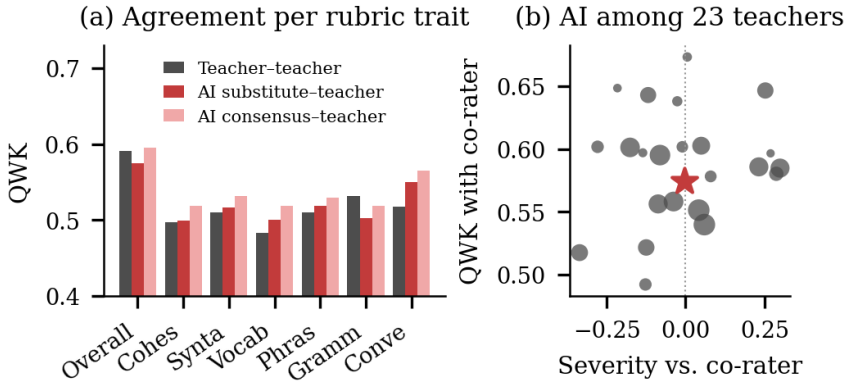


Figure 3. Teacher-substitution test on ELLIPSE (HYB). (a) QWK per rubric trait for teacher–teacher pairs, for the AI trained on one teacher and compared with the other, and for the consensus-trained AI. (b) The AI (star) among the 23 teachers who rated at least 100 essays; circle size is proportional to the number of essays rated.

5.3 Rubric-Guided Counterfactuals

Fig. 4 and Table IV show that the engines respond to the rubric only where a surface signal exists. Sentence shuffling, which destroys organization while keeping every word, leaves HYB unchanged (+0.003; 1.4% of essays move in either direction), and the same holds for every engine. Duplicating the longest paragraph adds 146 words of pure repetition and raises the HYB score by 0.27 points, so that 30.8% of essays gain at least one score level; FEAT rises by 0.35 and LEN by 0.66, whereas LEX and EMB are nearly unaffected. Typos lower all text-based engines (HYB −0.14), consistent with the conventions criterion. Replacing the evidence with off-topic evidence from another prompt lowers HYB by only 0.19 and raises the score of 8.4% of essays; LEX even increases on average (+0.075), and only EMB reacts strongly (−1.72). Deleting all evidence costs HYB 0.84 points, but deleting the same amount of other text costs 0.91: the length-normalized CSI is +0.029 for HYB and +0.040 for FEAT, i.e., these engines penalize the loss of evidence slightly less than the loss of arbitrary text. Counterclaims and rebuttals receive no specific credit either (HYB CSI = −0.007). Only LEX and EMB show negative CSIs for evidence (−0.046 and −0.112), confirming that sensitivity to argument content exists in the lexical and semantic representations but is overridden by length and mechanics in the stacked engine with the best agreement.

TABLE IV. Mean Score Change Under Rubric-Guided Counterfactuals

Edit	Exp.	LEN	LEX	FEAT	EMB	HYB	↑%	↓%
Shuffle	↓	0.000	+0.001	+0.016	0.000	+0.003	1.4	1.4
Pad (+146 w)	≤0	+0.655	+0.022	+0.346	−0.015	+0.270	30.8	0.7
Typo (10%)	↓	0.000	−0.147	−0.160	−0.401	−0.136	0.1	17.3
Off-topic evid.	↓	−0.003	+0.075	−0.090	−1.720	−0.192	8.4	30.0
Del. evidence	↓	−0.992	−0.692	−0.868	−0.307	−0.843	0.2	76.4
Ctrl. deletion	ref.	−1.006	−0.617	−0.954	−0.106	−0.907	0.2	78.6
CSI evidence	<0	0.000	−0.046	+0.040	−0.112	+0.029	—	—
CSI ctr./reb.	<0	+0.011	−0.047	+0.007	−0.058	−0.007	—	—

Exp.: change a teacher applying the rubric would expect. ↑%/↓%: share of essays whose discrete HYB score rises/falls. n = 1,107 (deletions: 649 and 461 eligible essays).

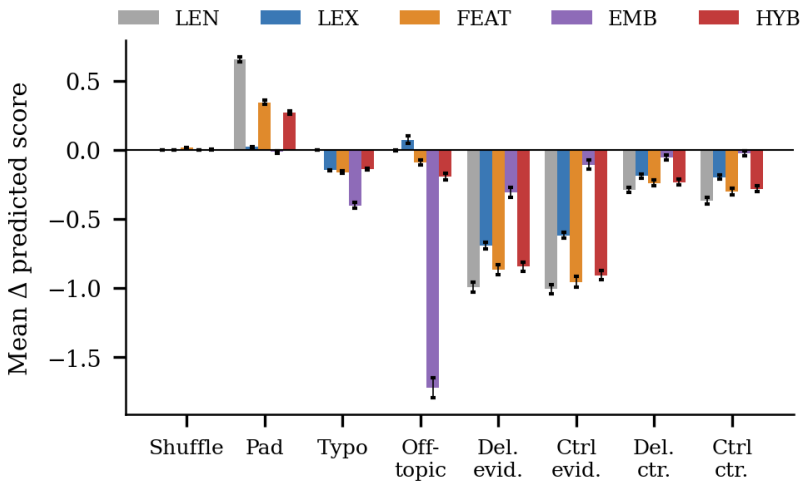


Figure 4. Mean change of the predicted holistic score under rubric-guided counterfactual edits (error bars: 95% CI). Ctrl denotes the length-matched deletion of non-target sentences.

5.4 Feedback Diagnostic Validity

Table V summarizes the feedback model. Element effectiveness is predicted well above the majority prior (macro-F1 0.549 vs. 0.289; AUC 0.871 for Ineffective and 0.862 for Effective elements). Under a budget equal to the expert base rate (4.8% of elements), 37.5% of the revise flags hit an element that experts rated Ineffective, and 37.6% of such elements are flagged. False criticism of Effective elements is rare (0.6%), whereas 3.8% of Adequate elements are flagged. Recall varies strongly by element type: weak Evidence is flagged in 51.3% of cases, but ineffective Claims, Positions, and Counterclaims in fewer than 19%. Structural feedback is more reliable: the advice to add a counterclaim is correct for 70.7% of the essays that receive it (rebuttal: 74.5%).

Two teacher-centered findings stand out. First, AI feedback is more strongly coupled to overall quality than expert diagnoses are: the AI effectiveness profile correlates with the holistic score at $\rho = 0.79$, compared with $\rho = 0.69$ for the expert profile, and only at $\rho = 0.70$ with the expert profile itself. The model largely predicts whether the essay is good and distributes that judgment over its elements, an element-level halo that mirrors the trait collinearity found in Lens 1. Second, the diagnostic burden falls unevenly. Experts rate more ELL elements as Ineffective (7.2% vs. 4.6%), yet the AI flags only 29.5% of them (39.3% for non-ELL students), doubles the false-criticism rate (1.2% vs. 0.6%), and its praise is far less precise (0.376 vs. 0.590). Structural advice is also less trustworthy for ELL students: 55.9% of ELL essays that contain a counterclaim are told to add one (46.7% for non-ELL students), and 84.2% vs. 64.9% for rebuttals. For Asian/Pacific Islander students, the AI flags only 19.8% of the weaknesses.

Table V. Feedback Diagnostic Validity on 55,068 Expert-Rated Elements

Model	Acc.	Macro-F1	QWK	AUC_I	AUC_E
(a) Element-level effectiveness prediction					
Majority prior	0.767	0.289	0.000	0.500	0.500
LR (lexical)	0.801	0.534	0.397	0.840	0.829
Hybrid (LR+GBM)	0.815	0.549	0.431	0.871	0.862
(b) Budget-matched flags of the hybrid model by student group					
Group	Expert I%	Rev. prec.	Rev. rec.	FCR%	Praise prec.
All students	4.8	0.375	0.376	0.6	0.582
Female / Male	4.4 / 5.2	0.381 / 0.370	0.356 / 0.395	0.5 / 0.7	0.587 / 0.575
Non-ELL	4.6	0.370	0.393	0.6	0.590
ELL	7.2	0.419	0.295	1.2	0.376
White	4.1	0.349	0.403	0.6	0.595
Hispanic	5.8	0.400	0.372	1.0	0.528
Black	6.6	0.392	0.365	0.8	0.527
Asian/PI	1.9	0.378	0.198	0.0	0.656
Other/Multi	3.6	0.397	0.370	0.0	0.597

Expert I%: share of elements rated Ineffective; Rev.: revise flags; FCR: false-criticism rate P(revise | Effective).

5.5 Score-Conditional Fairness

Fig. 5 and Table VI contrast raw and conditional SMDs of HYB on PERSUADE 2.0. Raw SMDs are positive for every group (0.010–0.112) and are driven by score distributions: across groups they correlate at r = 0.71 with the group's mean teacher score, a signature of the threshold-induced scale expansion (SD ratio 1.04). Conditioning on the teacher's score changes the picture in both directions. Artifactual gaps vanish, for ELL (0.053 → −0.011) and Hispanic students (0.045 → −0.001), while gaps hidden by the distributions emerge: essays by Black students receive AI scores 0.042 SD below those of essays with the same teacher score (95% CI [−0.076, −0.009]), essays by male students 0.029 SD below, and essays by students who are not economically disadvantaged 0.064 SD above. The lenient treatment of Asian/Pacific Islander students persists and grows (0.112 → 0.130, CI [0.076, 0.180]), exceeding the 0.10 flag and replicating the ELLIPSE result. The stricter prompt × score comparison reveals a further masked gap: ELL essays receive AI scores 0.18 points (≈0.15 SD) below those of non-ELL essays with the same teacher score on the same prompt (95% CI [−0.23, −0.13]). Agreement also differs by group: QWK is 0.793 for ELL students and 0.875 for other students, a gap present in every engine (LEX 0.753 vs. 0.839; FEAT 0.756 vs. 0.858). HYB has the smallest QWK range across groups (0.096 vs. 0.113–0.242 for the other engines), yet its worst conditional SMD (0.130) still exceeds the flag.

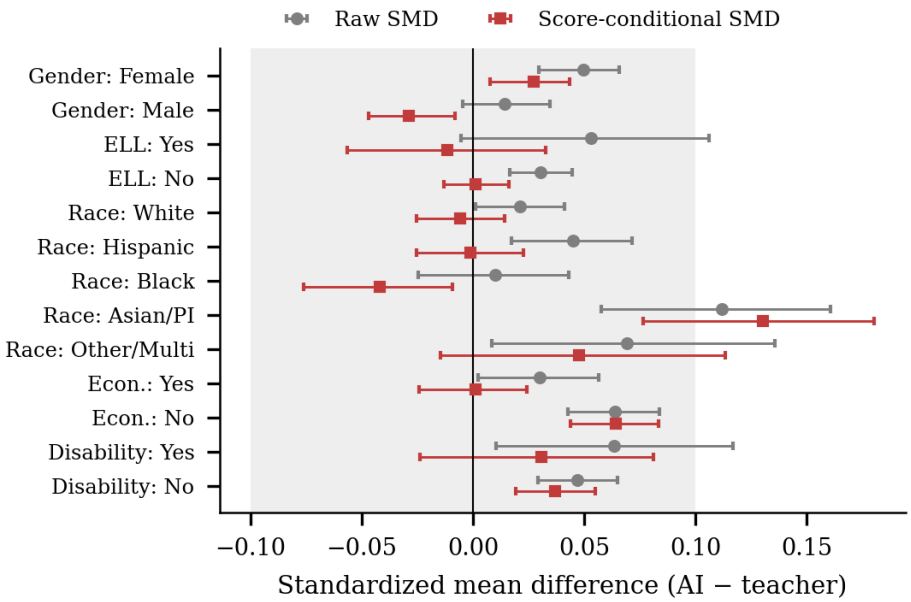


Figure 5. Raw versus score-conditional SMD of HYB for each student group on PERSUADE 2.0 (95% bootstrap CIs). The shaded band marks |SMD| ≤ 0.10.

Table VI. Raw and Score-Conditional Subgroup SMD of HYB on PERSUADE 2.0

Group	n	Mean T	SMD raw	SMD cond. [95% CI]	QWK
Female	2,843	3.57	0.049	0.027 [.008, .043]	0.868
Male	2,688	3.31	0.014	−0.029 [−.047, −.008]	0.868
ELL	537	3.10	0.053	−0.011 [−.057, .032]	0.793
Non-ELL	4,928	3.49	0.030	0.001 [−.013, .016]	0.875
White	2,456	3.52	0.021	−0.006 [−.025, .014]	0.870
Hispanic	1,472	3.26	0.045	−0.001 [−.025, .023]	0.850
Black	1,000	3.21	0.010	−0.042 [−.076, −.009]	0.856
Asian/PI	385	4.22	0.112	0.130 [.076, .180]	0.870
Other/Multi	194	3.55	0.069	0.048 [−.015, .113]	0.889
Econ. disadv.	1,627	3.41	0.030	0.001 [−.024, .024]	0.849
Not disadv.	2,070	3.86	0.064	0.064 [.044, .083]	0.876
Disability	350	3.41	0.063	0.031 [−.024, .081]	0.862
No disability	3,363	3.68	0.047	0.037 [.019, .055]	0.870

Mean T: mean teacher score. Economic and disability status are available for 3,697 and 3,713 essays.

5.6 Teacher-in-the-Loop Routing

Fig. 6 and Table VII show how teacher effort is best spent. Every informed router beats random routing, and learned deferral is the most effective: routing 20% of the essays raises QWK from 0.870 to 0.907 (random: 0.895) and reduces the worst-group |SMD| from 0.112 to 0.077 (random: 0.091). Random routing needs 29.1% of the essays to reach the agreement that learned deferral attains with 20%, a 31% saving in teacher effort; at a 10% budget, the equivalent random rate is 15.9%. Without seeing demographics, learned deferral also targets the students the AI serves worst: at a 20% budget it routes 26.6% of ELL essays to teachers and nearly halves the ELL agreement gap (0.083 → 0.043; random: 0.069). The oracle (QWK 0.945 at 20%) marks the headroom for better error predictors.

Conformal intervals make the residual uncertainty explicit. Marginal 90% intervals (half-width 0.88 points) achieve nominal coverage overall but cover only 87.0% of ELL essays and 65.0% of the essays that teachers scored 1. Mondrian calibration by predicted score does not repair this (ELL:

87.2%), because the undercoverage stems from the essays' difficulty rather than from their predicted level. Normalizing by the learned difficulty estimate raises ELL coverage to 89.2% and the coverage of the lowest-scored essays to 72.5%, and it shrinks the coverage gap across all audited groups from 0.057 to 0.022 at a modest increase in mean interval width (1.76 → 1.89 points).

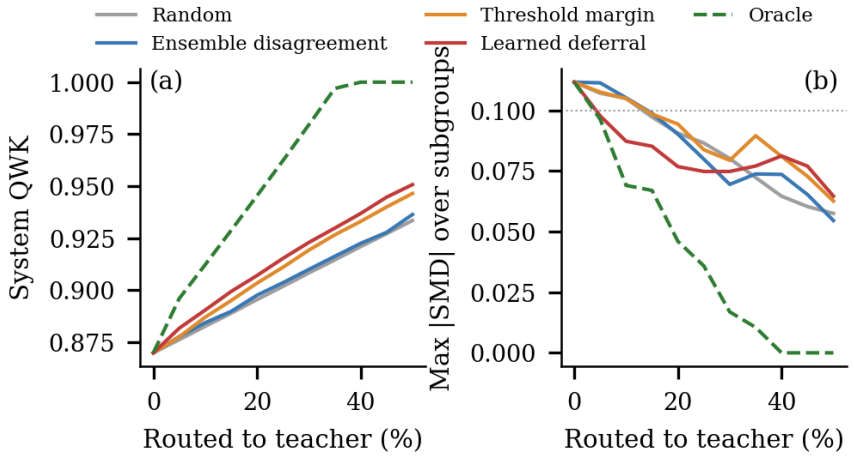


Figure 6. Teacher-in-the-loop routing on PERSUADE 2.0: (a) system QWK and (b) worst-group |SMD| as functions of the share of essays routed to the teacher.

Table VII. Teacher-in-the-Loop Routing and Conformal Coverage

Router	QWK 10%	QWK 20%	QWK 30%	Max|SMD|	ELL gap	ELL routed
(a) Routing (no routing: QWK 0.870, max|SMD| 0.112, ELL gap 0.083)						
Random	0.883	0.895	0.908	0.091	0.069	20.0%
Ens. disagreement	0.884	0.898	0.910	0.090	0.059	23.6%
Threshold margin	0.887	0.903	0.919	0.094	0.052	19.7%
Learned deferral	0.890	0.907	0.923	0.077	0.043	26.6%
Oracle (bound)	0.912	0.945	0.979	0.046	0.022	21.8%
(b) Split-conformal intervals, α = 0.10						
Interval	Coverage	Width	ELL cov.	Score-1 cov.	Group gap	Min group
Marginal	0.900	1.755	0.870	0.650	0.057	0.870
Mondrian (pred.)	0.900	1.755	0.872	0.640	0.050	0.870
Normalized	0.901	1.892	0.892	0.725	0.022	0.885

Max|SMD|, ELL gap, and ELL routed are reported at a 20% routing budget. Group gap: max − min coverage over audited groups.

6. Discussion

Implications for teachers. Agreement-based validation would certify HYB for operational use: it meets every aggregate criterion of [5] and behaves like a typical member of the teacher pool. TCE shows what that certification leaves out. The engine rewards length and repetition, ignores organization, barely distinguishes on-topic from off-topic evidence, and produces analytic and element-level feedback that mostly echoes its overall judgment. For teachers, this suggests a clear division of labor: the AI is a dependable first reader for overall placement, whereas its feedback on argument structure and its analytic profiles call for teacher review before they reach students, especially English learners, whose weaknesses it misses most often.

Implications for designers and evaluators. Three practices follow directly. First, report teacher-relative agreement ($\Delta\kappa$ for both consensus and single-teacher training targets) alongside agreement with consensus labels. Second, include counterfactual tests with length-matched controls in model acceptance; an engine that cannot tell evidence from filler should not be presented as evaluating argumentation. Third, audit fairness with score-conditional and prompt-stratified SMDs, because raw SMD both invented and concealed subgroup gaps in our data. The routing results add a constructive message: a modest, well-targeted teacher budget improves fairness as well as accuracy, and difficulty-normalized conformal intervals give teachers nearly group-uniform uncertainty estimates without using demographics.

Generality. TCE is model-agnostic: the same lenses apply unchanged to LLM graders and LLM-generated feedback [12–15, 17], for which the length, repetition, and halo effects identified here are natural first tests. Teacher re-scoring of counterfactual essays and an audit of the complete ASAP 2.0 release are direct extensions.

7. Conclusion

We introduced a teacher-centered evaluation of AI-assisted assessment that asks whether an AI grader can do a teacher's job rather than whether it reproduces a label. On two public corpora with 11,993 essays, the best engine matched the agreement of a second teacher and ranked within the teacher population, yet it was blind to organization, rewarded repetition, gave trait and element feedback dominated by overall quality, missed most weaknesses of English learners, and showed subgroup gaps that only score-conditional analysis revealed. A learned deferral policy that routes one essay in five to the teacher reduced the worst-group bias by 31% and nearly halved the ELL agreement gap. Evaluating AI graders through the teacher's eyes is therefore both a validation standard and a practical design tool for deciding where teachers' time matters most.

References

[1] Crossley, S. A., Baffour, P., Burleigh, L., & King, J. (2025). A large-scale corpus for assessing source-based writing quality: ASAP 2.0. Assessing Writing, 65, 100954.

[2] Crossley, S. A., et al. (2024). A large-scale corpus for assessing written argumentation: PERSUADE 2.0. Assessing Writing, 61, 100865.

[3] Crossley, S. A., Baffour, P., Tian, Y., Picou, A., Benner, M., & Boser, U. (2022). The persuasive essays for rating, selecting, and understanding argumentative and discourse elements (PERSUADE) corpus 1.0. Assessing Writing, 54, 100667.

[4] Crossley, S. A., et al. (2023). Measuring second language proficiency using the English Language Learner Insight, Proficiency and Skills Evaluation (ELLIPSE) corpus. International Journal of Learner Corpus Research, 9(2), 248–269.

[5] 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.

[6] 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.

[7] Perelman, L. (2014). When "the state of the art" is counting words. Assessing Writing, 21.

[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.

[9] Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. Journal of Technology, Learning, and Assessment, 4(3).

[10] 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).

[11] Ke, Z., & Ng, V. (2019). Automated essay scoring: A survey of the state of the art. In Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI) (pp. 6300–6308).

[12] 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.

[13] Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. Computers and Education: Artificial Intelligence, 6, 100234.

[14] Tate, T. P., et al. (2024). Can AI provide useful holistic essay scoring? Computers and Education: Artificial Intelligence, 7, 100255.

[15] Xiao, C., Ma, W., Xu, S. X., Zhang, K., Wang, Y., & Fu, Q. (2024). Human-AI collaborative essay scoring: A dual-process framework with LLMs. arXiv:2401.06431.

[16] Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81–112.

[17] Steiss, J., et al. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. Learning and Instruction, 91, 101894.

[18] Dai, W., et al. (2023). Can large language models provide feedback to students? A case study on ChatGPT. In Proceedings of the IEEE International Conference on Advanced Learning Technologies (ICALT) (pp. 323–325).

[19] 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).

[20] Loukina, A., et al. (2020). Using PRMSE to evaluate automated scoring systems in the presence of label noise. In Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications (BEA).

[21] Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. International Journal of Artificial Intelligence in Education, 32(4), 1052–1092.

[22] Romano, Y., Barber, R. F., Sabatti, C., & Candès, E. (2020). With malice toward none: Assessing uncertainty via equalized coverage. Harvard Data Science Review, 2(2).

[23] Madras, D., Pitassi, T., & Zemel, R. (2018). Predict responsibly: Improving fairness and accuracy by learning to defer. In Advances in Neural Information Processing Systems 31 (NeurIPS).

[24] Mozannar, H., & Sontag, D. (2020). Consistent estimators for learning to defer to an expert. In Proceedings of the 37th International Conference on Machine Learning (ICML) (pp. 7076–7087). PMLR 119.

[25] Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. In Advances in Neural Information Processing Systems 30 (NeurIPS).

[26] Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). Beyond accuracy: Behavioral testing of NLP models with CheckList. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 4902–4912).

[27] Ding, Y., Riordan, B., Horbach, A., Cahill, A., & Zesch, T. (2020). Don't take "nswvtnvakgxpm" for an answer—The surprising vulnerability of automatic content scoring systems to adversarial input. In Proceedings of the 28th International Conference on Computational Linguistics (COLING) (pp. 882–892).

[28] Madnani, N., & Loukina, A. (2016). RSMTool: Collection of tools building and evaluating automated scoring models. Journal of Open Source Software, 1(3), 33.

[29] Ke, G., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems 30 (NIPS) (pp. 3146–3154).