

Seven Ways to Cut a Score: Mapping Essay-Scoring Predictions to Ordinal Ratings

Ahmed H. Bediwy Martha Bellows Sue Lottridge

Pearson Education

{ahmed.bediwy, martha.bellows, sue.lottridge}@pearson.com

Abstract

One approach to automated essay scoring relies on neural regression models that output *continuous* predicted scores, yet operational scores must be reported on a discrete, ordinal rubric scale to match human raters' scores. The mapping from continuous predictions to integer scores is governed by a small set of *cutpoints*, and the choice of cutpoint estimator materially changes both accuracy and the score distribution that examinees experience. We introduce, formalize, and compare seven cutpoint methods—simple rounding, exact-agreement optimization, balance optimization, quadratic-weighted-kappa (QWK) maximization, maximum-likelihood Gaussian boundaries, a Bayesian ordered-prior estimator, and an ordinal cumulative link model—under a single notation. We evaluate all seven on seven prompts from the ASAP 2.0 dataset (24,728 essays), using a shared Longformer regression backbone, and report agreement (quadratic weighted kappa), association (the Pearson correlation between integer machine and human scores), and distributional fidelity (the standardized mean difference and the maximum score-point distribution gap). Our results expose a consistent accuracy–calibration trade-off: QWK maximization and balance optimization tie for the highest agreement, balance optimization achieves the smallest score-point distribution gap, the ordinal link model achieves the smallest mean bias, the generative Gaussian and Bayesian estimators recover score *means* but distort the distribution, and the ordinal link model is a strong all-rounder.

Keywords: automated essay scoring · constructed-response items · cutpoint estimation · transformer models · fine-tuning · human-AI agreement.

1 Introduction

Automated essay scoring (AES) systems assign scores to written responses at a scale that makes

human-only scoring impractical (Ke and Ng, 2019; Yan et al., 2020). Modern AES systems are typically built on pretrained transformer encoders (Devlin et al., 2019; Vaswani et al., 2017) that are fine-tuned to predict a single number per essay. When the target rubric is ordinal—for example, a holistic score in $\{1, \dots, 6\}$ —it is common and effective to train a *regression* head that outputs a continuous value, because regression exploits the ordering of the scale and is robust to class imbalance (Johan Berggren et al., 2019). The continuous prediction, however, is not a reportable score. To return an integer rubric score, the system must apply a set of *cutpoints* (also called thresholds or cut scores) that partition the real line into ordered intervals, one for each score point.

The cutpoint step is easy to overlook and easy to get wrong. A naive choice, such as rounding to the nearest integer, implicitly assumes that the regression model is unbiased at every half-integer boundary, which is rarely true for skewed score distributions. Different estimators optimize different objectives: some maximize raw agreement with human raters, some balance the kinds of errors made at each boundary, and others fit an explicit probabilistic model of how continuous predictions are generated within each score class. These objectives do not coincide, and the resulting cutpoints can differ enough to change an examinee's reported score and visibly reshape the score distribution.

Despite its operational importance, the cutpoint problem is rarely treated as a first-class modeling decision in the AES literature. Evaluation metrics report quadratic weighted kappa (Cohen, 1968) against human raters but often do not isolate how much of the measured agreement is attributable to the cutpoint estimator rather than to the underlying model. Psychometric practice, meanwhile, emphasizes that an operational scoring engine must preserve the *score distribution*, not merely the mean or the rank order, because the distribution drives

pass rates, subscore reporting, and fairness audits (Williamson et al., 2012).

This paper makes the cutpoint estimator the object of study. Our contributions are:

- We unify seven cutpoint methods under a single notation and give an explicit mathematical formulation for each, spanning agreement optimizers, a balance criterion, generative likelihood models, a Bayesian estimator, and an ordinal cumulative link model (Section 5).
- We evaluate all seven approaches across the seven ASAP 2.0 prompts (Crossley et al., 2025), using a shared Longformer (Beltagy et al., 2020) regression backbone and reporting both agreement and distributional-fidelity metrics (Sections 6–8).
- We characterize a robust accuracy–calibration trade-off and give practical guidance on method selection (Section 9).

2 Related Work

Automated essay scoring. Feature-based approaches like e-rater (Attali and Burstein, 2006) brought AES into operational testing, while shared benchmark corpora such as Cambridge ESOL (Yanakoudakis et al., 2011) and ASAP (The Hewlett Foundation, 2012) drove subsequent methodological advances. Neural approaches (Taghipour and Ng, 2016) and, later, transformer encoders (Devlin et al., 2019; Beltagy et al., 2020) improved agreement with human raters; Ke and Ng (2019) survey the state of the art. Most prior work treats scoring as either classification or regression and reports agreement metrics, leaving the continuous-to-ordinal mapping implicit.

Ordinal modeling. Rubric scores are ordinal, and the cumulative link, or proportional-odds, model (McCullagh, 1980; Agresti, 2010) is the classical tool for modeling ordinal responses. Bayesian treatments of ordinal data (Albert and Chib, 1993; Bürkner and Vuorre, 2019) place priors on ordered thresholds. We adapt both ideas to the cutpoint problem, where the predictor is a one-dimensional model score rather than a feature vector.

Evaluation and score reporting. Quadratic weighted kappa (QWK) (Cohen, 1968), a weighted extension of Cohen’s kappa (Cohen, 1960), is the primary AES agreement metric and was the

official measure of the ASAP competition (The Hewlett Foundation, 2012). Operational frameworks (Williamson et al., 2012; Yan et al., 2020) additionally require scoring engines to preserve the human score distribution. We therefore evaluate cutpoints on both axes, using QWK for agreement and standardized mean difference (SMD) (Cohen, 1988) and a score-point distribution gap for distributional fidelity.

3 Data and Continuous Scoring Model

Corpus. We use seven argumentative-writing prompts drawn from the ASAP 2.0 corpus (Crossley et al., 2025), totaling 24,728 single-human-scored essays. Two prompts received scores on a 1–5 scale and five received scores on a 1–6 scale. Each prompt is split 80/10/10 into training, validation, and test partitions with a fixed seed; cutpoints are fit and evaluated on the test data. Table 1 summarizes the corpora and the per-prompt quality of the continuous model.

Continuous model. For every prompt we fine-tune a Longformer (Beltagy et al., 2020) (allenai/longformer-base-4096) with a single-output regression head. For each prompt we tune over a grid of learning rates ($\{1, 2, 3, 4, 5\} \times 10^{-5}$), weight-decay values ($\{0.01, 0.1\}$), and learning-rate schedules ($\{\text{linear}, \text{cosine}\}$), with a fixed per-device batch size of 2, 8 gradient-accumulation steps (effective batch size 16), a 0.1 warmup ratio, and 4 training epochs, and select the configuration with the lowest validation-set root mean squared error (RMSE). All cutpoint methods operate on the *same* model’s continuous predictions output, so any difference in downstream metrics is attributable to the cutpoint estimator alone. The per-prompt RMSE (0.53–0.72) and Pearson correlation (0.80–0.88) in Table 1 confirm a strong but imperfect continuous signal—precisely the regime in which the cutpoint choice matters.

4 The Cutpoint Problem

Let the rubric scale be the ordered integers $s_1 < s_2 < \dots < s_K$ (e.g. 1, ..., 6 with $K = 6$). For essay i , let $r_i \in \mathbb{R}$ be the continuous model prediction and $y_i \in \{s_1, \dots, s_K\}$ the human score. A cutpoint vector is a strictly increasing sequence

$$t_1 < t_2 < \dots < t_{K-1}, \quad (1)$$

Prompt	Corpus	Test	Scale	Avg. words	RMSE	Pearson
A Cowboy Who Rode the Waves	2175	218	1–5	286	0.526	0.804
Car-free Cities	1959	197	1–5	452	0.648	0.841
Electoral College	2046	206	1–6	400	0.716	0.829
Driverless Cars	6170	617	1–6	388	0.554	0.851
Exploring Venus	4480	448	1–6	339	0.635	0.845
Facial Action Coding	4883	489	1–6	361	0.552	0.876
The Face on Mars	3015	302	1–6	322	0.614	0.836

Table 1: ASAP prompt corpora and per-prompt Longformer regression quality. Corpus is the full labeled set; Test is the held-out split on which cutpoints are fit and evaluated; Scale is the observed integer score range; RMSE and Pearson are on the continuous validation predictions.

with t_k separating score s_k from s_{k+1} . The induced rounding rule maps a prediction to the lowest score whose upper boundary it does not exceed,

$$\hat{y}(r; t) = \begin{cases} s_j, & j = \min\{m : r \leq t_m\}, \\ s_K, & r > t_{K-1}. \end{cases} \quad (2)$$

Every method below produces a vector t ; they differ only in the objective used to choose it. We write n for the number of test essays, n_k for the number with human score s_k , and $\bar{r}_k$ and $\hat{\sigma}_k^2$ for the sample mean and (unbiased) variance of $\{r_i : y_i = s_k\}$.

5 Cutpoint Methods

We group the seven methods into three families: a fixed baseline (Section 5.1); *agreement/error optimizers* that search threshold space directly (Sections 5.2–5.4); and *model-based* estimators that fit a probabilistic model of the predictions and derive cutpoints analytically (Sections 5.5–5.7).

5.1 Simple Rounding (baseline)

The baseline places cutpoints at the half-integers,

$$t_k = s_k + \frac{1}{2}, \quad k = 1, \dots, K-1, \quad (3)$$

so $\hat{y}(r)$ is ordinary nearest-integer rounding. It requires no fitting and is unbiased only if the continuous model is itself unbiased at each boundary.

5.2 Exact-Agreement Optimization

This method chooses thresholds that maximize the exact-agreement rate with human scores,

$$t^* = \arg \max_t \frac{1}{n} \sum_{i=1}^n \mathbb{1}[\hat{y}(r_i; t) = y_i]. \quad (4)$$

The objective in (4) is piecewise constant in t , so we optimize it with the derivative-free Nelder–Mead simplex (Nelder and Mead, 1965) for multi-threshold scales (and a bounded scalar search when

a single boundary lies between two populated classes). To stabilize the non-smooth search we wrap it in a bootstrap: over $B = 50$ resamples (fixed seed) we re-solve (4) and report the per-boundary median threshold. Exact-agreement optimization directly targets the fraction of essays scored identically to a human, but is blind to the *direction* and *magnitude* of disagreements.

5.3 Balance Optimization

Balance optimization estimates each cutpoint independently by equalizing the two opposing error directions at a given adjacent score boundary. For the boundary between scores s_k and s_{k+1} , and for a candidate threshold t , define

$$L_k(t) = \#\{i : r_i \leq t, y_i = s_{k+1}\}, \quad (5)$$

$$H_k(t) = \#\{i : r_i > t, y_i = s_k\}, \quad (6)$$

where $L_k(t)$ counts *false-low* assignments (essays truly in the upper category but placed at or below the boundary) and $H_k(t)$ counts *false-high* assignments (essays truly in the lower category but placed above the boundary). The balance criterion chooses

$$t_k = \arg \min_{t \in [\ell_k, u_k]} |L_k(t) - H_k(t)|. \quad (7)$$

Operationally, the method restricts attention to essays whose human scores are either s_k or s_{k+1} , scans candidate thresholds on a grid with step 0.01, and selects the threshold with the smallest absolute imbalance. Unlike the exact-agreement and QWK methods, there is no joint optimization over all boundaries: each t_k is solved separately.

The search interval $[\ell_k, u_k]$ is not learned from the data. Instead, it is inherited from the production balance rule, whose boundary ranges reflect decades of operational automated-scoring experience. Each boundary is confined to a predefined,

score-dependent interval around its nominal location, and these ranges were established through years of operational modeling rather than tuned to any single dataset. They encode accumulated institutional knowledge about plausible threshold placements, limiting the search to empirically reasonable regions while preserving the behavior of the production system.

5.4 QWK Maximization

Because quadratic weighted kappa (QWK) is the headline AES metric, this method optimizes it directly. With K categories, the quadratic weights and kappa are

$$w_{ab} = \frac{(a - b)^2}{(K - 1)^2}, \quad \kappa_w = 1 - \frac{\sum_{a,b} w_{ab} O_{ab}}{\sum_{a,b} w_{ab} E_{ab}}, \quad (8)$$

where O is the observed confusion matrix between $\hat{y}(r; t)$ and y , and E its expectation under independence with the same margins (Cohen, 1968). We solve

$$t^* = \arg \max_t \kappa_w(\hat{y}(r; t), y) \quad (9)$$

with Nelder–Mead restarts and the same $B = 50$ bootstrap median aggregation as exact-agreement optimization. Unlike (4), the quadratic weights penalize large disagreements more than adjacent ones, so QWK maximization trades a few exact matches for fewer far-off errors.

5.5 Maximum-Likelihood Gaussian Boundaries

The first model-based approach method assumes that the predictions for each class follow a Gaussian distribution,

$$r \mid y = s_k \sim \mathcal{N}(\mu_k, \sigma_k^2), \quad (10)$$

estimated by maximum likelihood, $\hat{\mu}_k = \bar{r}_k$ and $\hat{\sigma}_k^2$ the unbiased class variance (heteroscedastic by default). Under equal class priors, the Bayes-optimal boundary between adjacent classes is the point where the two densities cross, $\mathcal{N}(t; \mu_k, \sigma_k^2) = \mathcal{N}(t; \mu_{k+1}, \sigma_{k+1}^2)$. Taking logarithms yields a quadratic $at^2 + bt + c = 0$ with

$$a = \frac{1}{2\sigma_{k+1}^2} - \frac{1}{2\sigma_k^2}, \quad (11)$$

$$b = \frac{\mu_k}{\sigma_k^2} - \frac{\mu_{k+1}}{\sigma_{k+1}^2}, \quad (12)$$

$$c = \frac{\mu_{k+1}^2}{2\sigma_{k+1}^2} - \frac{\mu_k^2}{2\sigma_k^2} + \ln \sigma_{k+1} - \ln \sigma_k. \quad (13)$$

We take the root lying in $[\mu_k, \mu_{k+1}]$ as t_k . When the variances are equal the quadratic degenerates and the boundary reduces exactly to the midpoint $t_k = (\mu_k + \mu_{k+1})/2$. Because ASAP score distributions are skewed and extreme classes can be sparse, the scale is fixed to its full nominal range, a class mean is estimated directly only when it has at least five essays, and sparse means are filled by monotone interpolation/extrapolation so that the μ_k remain strictly increasing.

5.6 Bayesian Ordered-Prior Estimator

The Bayesian estimator keeps the Gaussian likelihood of Section 5.5 but enforces ordering through the prior and quantifies uncertainty. We reparameterize the means through positive gaps $\delta_k = \mu_k - \mu_{k-1} > 0$ and place

$$\mu_1 \sim \text{Uniform on the observed range}, \quad (14)$$

$$\delta_k \sim \text{Exponential}(\lambda), \quad (15)$$

$$\log \sigma_k \sim \mathcal{N}(\log \sigma_0, \tau^2), \quad (16)$$

so the ordering $\mu_1 < \dots < \mu_K$ holds by construction. We fit the maximum-a-posteriori (MAP) parameters by bound-constrained quasi-Newton optimization (Byrd et al., 1995) and form a Laplace (Gaussian) approximation to the posterior from the inverse Hessian at the mode. Drawing 500 posterior samples (fixed seed) yields a median and 95% credible interval for each cutpoint, where each sampled parameter set induces cutpoints through the same density-crossing rule as Section 5.5. Sparse or empty extreme classes contribute no likelihood term; their incoming gap is then governed by the ordered prior—the Bayesian analogue of the interpolation used by the maximum-likelihood method.

5.7 Ordinal Cumulative Link Model

The final method models the ordinal score directly with a proportional-odds cumulative link model (CLM) (McCullagh, 1980; Agresti, 2010) in the one-dimensional model score r . With the logit link

function and ordered intercepts $\theta_1 < \dots < \theta_{K-1}$ and a single slope β ,

$$\Pr(y \leq s_k | r) = \frac{1}{1 + \exp(-(\theta_k - \beta r))}, \quad (17)$$

fit by maximum likelihood. The class probabilities follow by differencing (17), and the modal predicted category is a step function of r . We recover the raw-score cutpoint t_k as the value of r at which the modal prediction switches from s_k to s_{k+1} , located by a dense grid search over the score range. Boundaries adjacent to a sparse class, which the data cannot inform, are filled by continuing the observed spacing of the recovered cutpoints, with a monotonicity guard. The proportional-odds constraint shares a single slope across all boundaries, yielding smoothly ordered cutpoints.

6 Evaluation

We assess each cutpoint vector along three descriptive axes—agreement, association, and distributional fidelity. Because the ASAP dataset provides a single human score per essay, there is no human–human baseline against which operational tolerances are normally calibrated, so we present the metrics purely descriptively.

Agreement. We report the QWK between the machine and the human scores, computed with the full nominal label set $\{s_1, \dots, s_K\}$ so that empty categories do not inflate it.

Association. We report the Pearson correlation coefficient r between the machine human scores. Whereas κ_w rewards absolute agreement, r measures how strongly the two integer score vectors co-vary once cutpoints have been applied.

Distributional fidelity. Operationally, the reported score distribution should be within acceptable range difference from the human scores distribution. We summarize bias with the standardized mean difference (SMD) (Cohen, 1988),

$$\text{SMD} = \frac{\bar{\hat{y}} - \bar{y}}{s_p}, \quad (18)$$

where s_p is the pooled standard deviation of the machine and human scores; a positive SMD indicates systematic over-prediction. We also report the maximum score-point distribution gap,

$$\Delta_{\max} = \max_k |\hat{p}_k - p_k| \times 100\%, \quad (19)$$

Method	QWK	Pearson r	SMD	Max SP gap (pts)
Simple Rounding	0.787	0.808	+0.212	6.78
Exact Agreement	0.801	0.813	+0.112	7.23
Balance	0.807	0.809	+0.049	3.31
QWK Maximization	0.807	0.818	+0.127	6.45
MLE Gaussian	0.805	0.817	+0.013	7.25
Bayesian Ordered	0.805	0.815	+0.004	7.29
Ordinal CLM	0.799	0.804	-0.009	3.72

Table 2: Per-method agreement, association, and distributional fidelity, averaged across the seven ASAP prompts. QWK is the quadratic weighted kappa between integer machine and human scores; Pearson r is their linear correlation; SMD is the signed standardized mean difference (positive = over-prediction); Max SP gap is the maximum score-point distribution gap in percentage points. All values are means over the seven prompts.

where p_k and $\hat{p}_k$ are the human and machine proportions at score s_k . SMD captures mean bias while $\Delta_{\max}$ captures the largest shape discrepancy in the score distribution.

7 Experimental Setup

All seven methods are run in-process on the same test-set predictions for each of the seven prompts (49 method×prompt) combinations. Stochastic components—the bootstrap resampling in Sections 5.2 and 5.4 and the posterior sampling in Section 5.6—use fixed seeds, so every reported number is deterministic and reproducible from the predictions alone. Metrics are computed with a single shared implementation built on NumPy, SciPy, scikit-learn, and statsmodels (Harris et al., 2020; Virtanen et al., 2020; Pedregosa et al., 2011; Seabold and Perktold, 2010).

8 Results

Table 2 reports each metric averaged across the seven prompts. Figure 1 shows per-method agreement with prompt-level dispersion, Figure 2 plots the central accuracy–calibration trade-off, Figure 3 illustrates where the methods place their boundaries on one representative prompt, and Figure 4 gives the full prompt×method QWK matrix.

Agreement and association. QWK maximization and balance optimization tie for the highest mean QWK (0.807); QWK maximization leads on Pearson correlation ($r = 0.818$) ahead of balance ($r = 0.809$) and exact-agreement optimization (QWK = 0.801, $r = 0.813$). The model-

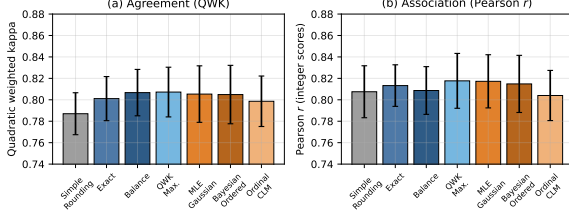


Figure 1: Per-method agreement (panel a, quadratic weighted kappa) and association (panel b, Pearson r between integer machine and human scores), averaged across the seven prompts.

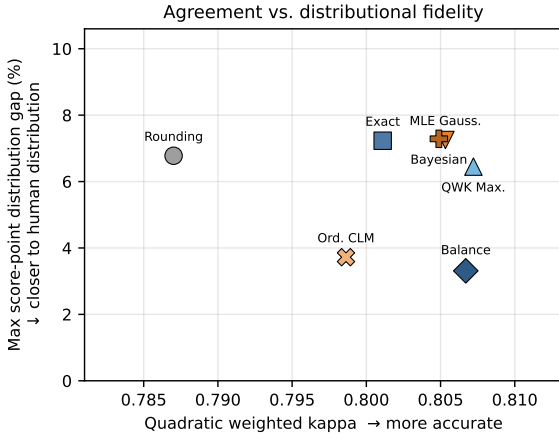


Figure 2: Agreement vs. distributional fidelity. Each marker is one method, positioned by mean quadratic weighted kappa (right is more accurate) and mean maximum score-point distribution gap (down is closer to the human distribution).

based generative estimators are competitive on association—maximum-likelihood Gaussian and Bayesian boundaries reach $r = 0.817$ and $r = 0.815$. Simple rounding is the weakest on QWK (0.787) and, as the baseline, leaves a clear margin for every fitted method.

Distributional fidelity. The ranking changes when it comes to score point distributions. On distributional fidelity the picture is nuanced. Balance optimization achieves the smallest maximum score-point gap (3.31 points) while the ordinal CLM achieves the smallest mean bias ($|\text{SMD}| = 0.015$, gap 3.72 points), with a slight under-prediction tendency (its SMD is negative). Together these two methods dominate the calibration axis. The agreement optimizers show moderate score-point gaps (6–7 points) and a positive over-prediction bias. Most strikingly, the generative estimators, despite small mean bias ($|\text{SMD}| \leq 0.03$), produce large score-point gaps (≈ 7 points): they match

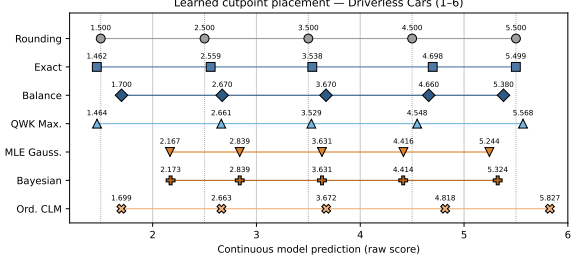


Figure 3: Learned cutpoints for the *Driverless Cars* prompt (1–6 scale). Agreement optimizers (blues) sit near the half-integer grid (dotted vertical lines); the model-based methods (oranges) pull the lower boundaries rightward to track the class-conditional densities.

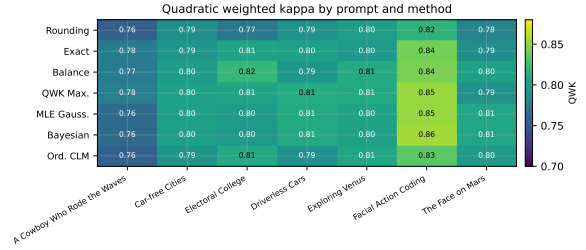


Figure 4: Quadratic weighted kappa for every prompt $\times$ method cell. Columns are ordered by corpus; rows follow the method families of Section 5.

the score *mean* yet distort the score *distribution*. Simple rounding shows the largest systematic over-prediction ($\text{SMD} = +0.212$).

Where the boundaries land. Figure 3 explains the trade-off mechanically. On *Driverless Cars*, the agreement optimizers keep boundaries close to the half-integer grid, whereas the Gaussian and Bayesian estimators shift the lower boundaries substantially to the right (e.g. the first cut moves from 1.50 to ≈ 1.95) to align with where the class densities actually cross. That rightward shift is what reshapes the low end of the score distribution and drives the generative methods’ large score-point gaps.

Consistency across prompts. Figure 4 shows that the agreement ordering is stable: QWK is highest on the larger 1–6 prompts (e.g. Facial Action Coding, Exploring Venus) and lowest on the smaller 1–5 prompts, mirroring the continuous model quality in Table 1. No method outperforms the others in QWK by more than a few hundredths on any prompt, suggesting that cutpoint selection functions primarily as a distributional fidelity mechanism rather than a driver of predictive accuracy.

9 Discussion

An accuracy–calibration trade-off. The central finding is that no single method wins on both axes. Maximizing agreement (QWK or exact) buys a point or two of kappa at the cost of a score distribution that drifts from the human one, while balance and the ordinal link model give up little agreement for markedly better calibration. The generative estimators occupy an instructive failure mode: by construction they match class-conditional densities, which controls the *mean* but not the *shape* of the reported distribution, so they exhibit the largest score-point distribution gaps even though their SMD is small. This is a concrete reminder that mean-level unbiasedness and distributional fidelity are different requirements.

Practical guidance. For operational deployment where the score distribution is contractually or psychometrically constrained, balance optimization is the strongest default: it is simple, deterministic, and achieves the smallest mean bias and score-point gaps on all seven prompts with near-top QWK. When the reporting requirement is a smoothly ordered set of boundaries with quantified uncertainty—for example to support standard setting—the ordinal CLM and the Bayesian estimator are attractive, the latter supplying credible intervals at the price of distributional distortion. When the sole published target is QWK, QWK maximization is the literal optimizer of that objective, but reviewers should be aware that it does not protect the distribution. Simple rounding is a reasonable sanity baseline but is dominated by every fitted method here.

Limitations

Our study fixes a single continuous backbone (Longformer regression) and a single corpus family (seven ASAP prompts); the magnitude of the accuracy–calibration trade-off may differ for other model classes, languages, or numbers of score points. Cutpoints are fit and evaluated on the same held-out test partition, so the reported metrics are in-sample with respect to the cutpoint estimator; an additional examinee-disjoint partition would better estimate operational stability. Because the ASAP labels provide a single human score per essay, we cannot compute human–human agreement baselines, so our distributional fidelity is assessed with machine–human metrics alone rather

than the full set of operational checks. Finally, the generative and Bayesian methods assume Gaussian class-conditional predictions; heavier-tailed or multimodal prediction distributions would warrant richer likelihoods.

Conclusion

We framed the continuous-to-ordinal cutpoint step in automated essay scoring as a first-class modeling decision, formalized seven estimators under one notation, and compared them on seven ASAP prompts with a shared regression backbone. The methods separate cleanly into agreement optimizers, which maximize kappa and the correlation with human scores, and calibration-oriented methods, which preserve the score distribution; balance optimization and the ordinal cumulative link model offer the best overall compromise, while generative Gaussian and Bayesian boundaries recover score means but distort the distribution. We hope that treating cutpoints explicitly—and reporting distributional fidelity alongside agreement—becomes standard practice in AES evaluation.

References

- Alan Agresti. 2010. *Analysis of Ordinal Categorical Data*, 2nd edition. John Wiley & Sons, Hoboken, NJ.
- James H. Albert and Siddhartha Chib. 1993. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679.
- Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v.2. *Journal of Technology, Learning, and Assessment*, 4(3).
- Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.
- Paul-Christian Bürkner and Matti Vuorre. 2019. Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, 2(1):77–101.
- Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyu Zhu. 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.
- Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

- Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Scott A. Crossley, Perpetual Baffour, L. Burleigh, and Jules King. 2025. A large-scale corpus for assessing source-based writing quality: ASAP 2.0. *Assessing Writing*, 65:100954.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics.
- Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, and 7 others. 2020. Array programming with NumPy. *Nature*, 585:357–362.
- Stig Johan Berggren, Taraka Rama, and Lilja Øvrelid. 2019. Regression or classification? automated essay scoring for Norwegian. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 92–102, Florence, Italy. Association for Computational Linguistics.
- Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 6300–6308.
- Peter McCullagh. 1980. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–142.
- John A. Nelder and Roger Mead. 1965. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Skipper Seabold and Josef Perktold. 2010. Statsmodels: Econometric and statistical modeling with Python. In *Proceedings of the 9th Python in Science Conference*, pages 92–96.
- Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891. Association for Computational Linguistics.
- The Hewlett Foundation. 2012. The hewlett foundation: Automated essay scoring. Kaggle competition.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.
- Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, and 1 others. 2020. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272.
- David M. Williamson, Xiaoming Xi, and F. Jay Breyer. 2012. A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1):2–13.
- Duanli Yan, André A. Rupp, and Peter W. Foltz, editors. 2020. *Handbook of Automated Scoring: Theory into Practice*. Chapman and Hall/CRC, Boca Raton, FL.
- Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189. Association for Computational Linguistics.