Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Post-hoc calibration of prompted large language models for automated essay scoring: rubric agreement, uncertainty and between-group equity

Xuan Cheng
Macao Polytechnic University
Yapeng Wang
Macao Polytechnic University
Dingcheng Yang
Nanchang University
Benjamin K. Ng
Macao Polytechnic University
Sio-Kei Im
Macao Polytechnic University
Xu Yang
xuyang@mpu.edu.mo
Macao Polytechnic University

Article

Keywords: automated essay scoring, large language models, writing assessment, educational equity, uncertainty quantification, conformal prediction

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10801101/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Post-hoc calibration of prompted large language models for automated essay scoring: rubric agreement, uncertainty and between-group equity

Xuan Cheng[1], Yapeng Wang[1], Dingcheng Yang[2], Benjamin K. Ng[1], Sio-Kei Im[1], Xu Yang[1*]

[1*]Faculty of Applied Sciences, Macao Polytechnic University, Macao SAR, China.
[2]Information Engineering School, Nanchang University, Nanchang, China.

*Corresponding author(s). E-mail(s): xuyang@mpu.edu.mo;

Abstract

Large language models are increasingly used to score student writing, but a score becomes usable in a classroom or an examination only once it sits on the intended rubric scale, carries confidence that can be trusted, and does not err more heavily on some students than on others. This article addresses those requirements without retraining the model. We formulate calibration of prompted LLM essay scores as a problem defined by the ordinal, rubric-bounded structure of the scale, and instantiate it as a five-stage layer: anchor-referenced scoring against retrieved same-scale exemplars, repeated querying, monotone score calibration, temperature scaling, and conformal prediction sets that route uncertain essays to human readers. Across the ELLIPSE, PERSUADE and ASAP 2.0 corpora and three commercial models, calibration raises quadratic weighted kappa in all nine settings, from a mean of 0.240 to 0.416, and cuts mean absolute bias from 0.886 to 0.029 score points; paired bootstrap intervals exclude zero in all 45 comparisons. A simulated pass/fail decision improves from 0.471 to 0.225 misclassification before any referral. Between-group error gaps narrow in 24 of 27 demographic comparisons, though individual reductions are mostly unresolved at these subgroup sizes. The correction transfers across corpora and to unseen prompts.

Keywords: automated essay scoring, large language models, writing assessment, educational equity, uncertainty quantification, conformal prediction

1 Introduction

Automated essay scoring (AES) has moved quickly from feature-based and neural systems trained on task-specific data [1–4] to systems that simply prompt a general-purpose large language model [5]. The appeal is clear. A prompted model needs no labelled training corpus for the target rubric, adapts to a new prompt by editing text rather than by retraining, and can be deployed against a commercial endpoint in a matter of hours. For institutions facing a new writing task, a new cohort, or a rubric that has never been scored at scale, this is often the only feasible route.

The difficulty is that a prompted score is not the same kind of object as a score from a trained model. Work on the LLM-as-a-Judge paradigm has documented length preference, prompt sensitivity, position effects and self-preference in model-generated evaluations [6–8], and repeated calls to the same model with the same input frequently return different scores. In a research benchmark these properties merely inflate variance. In an assessment setting they carry consequences, because an ordinal score may determine whether a student passes, whether an essay is routed to a human reader, and what feedback is returned. What matters operationally is therefore not only whether the model ranks essays sensibly, but whether its scores land on the intended scale, whether the confidence attached to them can be trusted, and whether its errors fall evenly across the student population [9].

These are calibration questions, and calibration has a mature toolkit. Isotonic regression, Platt-style logistic scaling, beta calibration and temperature scaling all align model outputs with empirical frequencies after training [10–14], and conformal prediction converts a point estimate into a set with distribution-free coverage guarantees [15–17]. The toolkit has been applied to pre-trained transformers [18] and, more recently, to uncertainty in LLM outputs [19–22] including the LLM-as-a-Judge setting [23, 24]. What has not been established is how these tools behave when the object being calibrated is a prompted score on a rubric-defined ordinal scale, how much of the benefit survives when the calibrator is moved to a different corpus, what the sampling cost amounts to, and whether calibration helps or harms students in demographic groups that the model already scores less accurately.

We take up those questions here. We formulate calibration for prompted LLM essay scores as a problem defined by the ordinal, rubric-bounded structure of the target scale, and instantiate it as a five-stage layer whose components answer to that structure: anchor-referenced scoring against retrieved same-scale exemplars, repeated querying, monotone score calibration, temperature scaling of the resulting distribution, and conformal prediction sets that route uncertain essays to human readers. We evaluate the layer across three public corpora, three commercial LLMs, five calibrator families, five sampling budgets, five conformal risk levels, eighteen cross-corpus transfer pairs and twenty-seven demographic attribute-by-setting comparisons, all under a single cached-prediction protocol so that the numbers are directly comparable across experiments.

A single diagnosis organises the results. The dominant failure of a prompted judge on a rubric scale is a monotone displacement rather than a loss of discrimination: the model orders essays adequately but places that ordering in the wrong region of the scale. This predicts which corrections should work, and the comparison of calibrator

families bears it out. Three monotone families fall within a few thousandths of one another while a non-monotone binning baseline trails them all, which locates the active ingredient in the order-preserving constraint rather than in any particular estimator. With that settled, the remaining design choices can be assessed on their own terms. Repeated sampling contributes little agreement once calibration is in place and earns its cost only as the raw material for the distributional outputs on which review triage depends. A calibrator fitted on one corpus and applied unchanged to another recovers a median of roughly three quarters of the in-domain error reduction without using target calibration labels, so the correction is not tied to the split on which it was estimated. Finally we ask whom the correction helps, tracking how calibration redistributes error across demographic subgroups and attaching essay-level bootstrap intervals to those disparities rather than reporting counts alone, an audit that, to our knowledge, has not previously been reported for LLM-based scoring.

2 Related Work

2.1 Automated Essay Scoring

AES research has long balanced predictive accuracy against validity and interpretability. Feature-based systems such as e-rater were built around linguistically interpretable indicators so that scores could be defended to stakeholders [1], and the field has maintained an explicit concern with the evidentiary requirements for operational use [9]. Neural approaches improved agreement with human raters [2, 3, 25], and surveys of the area document the shift towards representation learning while noting persistent issues with ordinal label treatment, robustness and transparency [4]. More recent work has begun to evaluate prompted LLMs as scorers [5], generally reporting reasonable correlation with human scores alongside noticeable instability.

Operational scoring practice also supplies a mechanism that we adopt directly. Human raters are trained and kept aligned using anchor essays and rubric exemplars, that is, previously scored texts that fix what each level on the scale means. Retrieval of scored reference essays as in-context exemplars is the computational analogue [26, 27], and has been used to improve the interpretability of neural AES [28].

2.2 LLM-as-a-Judge and Evaluation Bias

Using an LLM to evaluate open-ended text is now standard practice, and so is the observation that such evaluations carry systematic distortions. Length-controlled analyses show that judges reward longer responses independently of quality [6], comparisons across judge designs show sensitivity to prompt formulation [7], and large-scale studies of judge agreement document position effects and self-preference [8]. These findings motivate treating a prompted score as a raw signal that must be constrained, corrected and audited before use, which is the stance adopted here.

2.3 Calibration and Uncertainty Quantification

Post-hoc calibration adjusts model outputs after training so that stated confidence matches observed accuracy. Classical methods include Platt scaling [10], isotonic

regression [11, 12], beta calibration [14] and temperature scaling for modern neural networks [13], with expected calibration error the standard summary [29]. Conformal prediction provides a complementary, distribution-free route to valid prediction sets [15–17] and has been extended to ordinal outcomes [23]. For LLMs specifically, work has examined whether models know what they know [19], how to elicit calibrated verbal confidence [20, 22], and how to define uncertainty over semantically equivalent generations [21]. Closest to our setting, Sheng et al.[24] analyse uncertainty in LLM-as-a-Judge using conformal intervals for general natural language generation evaluation. We differ in target and in construction: our labels are rubric-defined ordinal levels rather than a generic quality scale, so the calibrator must preserve order, and our layer combines monotone point-score calibration with temperature-scaled predictive distributions and conformal sets rather than intervals alone.

2.4 Research Gaps

Three limitations recur in existing evaluations of calibrated LLM scoring. Such evaluations typically compare a calibrated pipeline against its uncalibrated self, which cannot reveal whether the specific calibrator matters. They also assess robustness on the same corpus the calibrator was fitted on, leaving the portability of a fitted calibrator untested. Most consequential for the application, we are not aware of published evidence on whether post-hoc calibration of LLM essay scores changes the distribution of error across demographic subgroups. A calibration layer that improved aggregate metrics while widening group disparities would be unusable in assessment, and this needs to be checked rather than assumed.

3 Methods

3.1 Corpora and Evaluation Condition

We use three publicly documented English essay corpora, chosen so that the evaluation spans analytic and holistic scoring, fine and coarse score granularity, and different writing genres and populations. ELLIPSE [30] contains essays by English language learners, each scored on six analytic traits (cohesion, syntax, vocabulary, phraseology, grammar and conventions) on a 1.0 to 5.0 scale in half-point steps; of these we use 442 essays for anchor construction, 95 for calibration and 95 for testing, giving 570 essay-trait pairs in each of the calibration and test splits. PERSUADE [31] contains argumentative essays carrying a single holistic score on a 2.0 to 6.0 scale; after discourse-to-essay reconstruction we use 2,449 essays for anchor construction, 525 for calibration and 525 for testing. ASAP 2.0 [32] contains 5,457 source-based essays with a holistic 1 to 6 score, split by stratified sampling at a 70/15/15 ratio into 3,819 anchor-bank essays, 818 calibration essays and 820 test essays. Across the nine model-corpus settings the evaluation therefore covers 5,745 essay-trait pairs in total.

All experiments reported here use a perturbed version of each corpus, which we refer to throughout as the controlled distribution shift. Essays at or above the 75th word-count percentile have their score labels shifted upward and essays at or below the 25th percentile shifted downward (± 0.5 for ELLIPSE, ± 1 for PERSUADE and

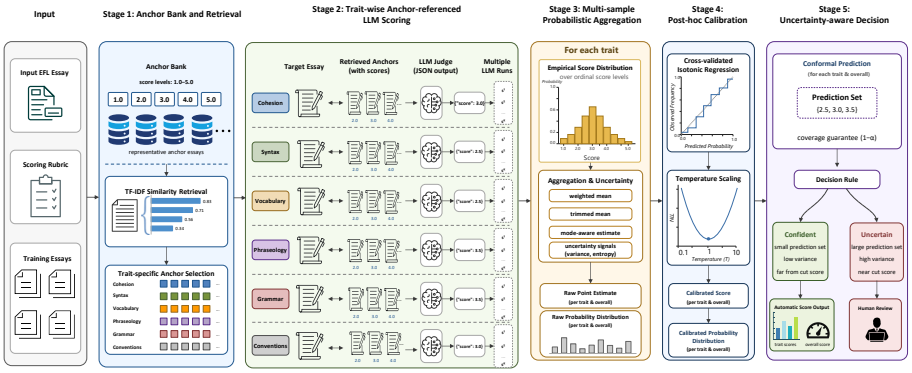

Fig. 1 Five-Stage Calibration Layer. The "coverage guarantee $(1-\alpha)$" label in stage 5 marks the nominal target only. As implemented the residual quantile comes from held-out folds of the calibration split, so the procedure is cross-conformal and the coverage reported in Section 4 is empirical.

ASAP 2.0), followed by clipping and rounding to legal levels; the test distribution is then reweighted by resampling with weights proportional to $\exp(1.2\bar{s} + 0.8\bar{w})$, where $\bar{s}$ and $\bar{w}$ are normalised score and word count, using a fixed seed so that the same perturbed corpora are presented to every model.

Two properties of this construction should be stated plainly, because they bound what the experiments can establish. The perturbation is applied to the reference labels rather than to the model outputs, so it is a synthetic target perturbation and not a reproduction of naturally occurring LLM length bias; what it creates is a scoring scale that is displaced in a length-dependent way relative to the model, which is the condition a calibration layer is designed to correct. The reweighting additionally produces a test cohort whose composition differs from the reference population. We adopt this as the primary evaluation condition so that every experiment shares one protocol and one set of cached model outputs, and we return in Section 5 to the limits of what a constructed perturbation can show.

3.2 The Calibration Layer

Figure 1 shows the pipeline. It has five stages, and every stage after the first treats the LLM as a black box whose parameters are never modified.

3.2.1 Anchor-Referenced Scoring

Let $\mathcal{T}$ denote the scoring dimensions and $\mathcal{Y}$ the legal score levels. For each dimension $t \in \mathcal{T}$ and level $y \in \mathcal{Y}$ we collect the training essays scored at that level into an anchor bank $\mathcal{A}_{t,y}$, so that every point on the rubric scale is represented by real texts rather than by a verbal description alone. For a target essay x the system retrieves the most similar anchor at each level by TF–IDF cosine similarity,

$$a^*_{t,y} = \arg\max_{a \in \mathcal{A}_{t,y}} \operatorname{sim}(x, a), \tag{1}$$

and inserts the retrieved set $A_t = \{a^*_{t,y} \mid y \in \mathcal{Y}\}$ into a dimension-specific prompt. The judge then scores one dimension at a time by comparing x against anchors whose scores are known, and returns structured JSON containing the dimension, the score and a short rationale. Writing the r-th call as

$$\hat{y}_t^{(r)} = f_{\text{LLM}}(x, A_t, r), \tag{2}$$

the stage as a whole replaces unconstrained score generation with bounded comparative judgment against an explicit reference set.

Retrieval is per level rather than global, which guarantees that the prompt spans the full scale instead of clustering near the modal score, and it is per dimension rather than per essay, so that an essay strong in vocabulary and weak in conventions is compared against different references for the two traits. TF–IDF was preferred to a dense retriever because it is transparent, cheap and reproducible without adding a further model dependency, and because the anchors serve to fix the scale rather than to supply semantic neighbours, for which lexical overlap is an adequate similarity signal.

3.2.2 Probabilistic Aggregation

Repeated scoring of the same essay-dimension pair yields a sample set $S_t(x) = \{\hat{y}_t^{(1)}, \ldots, \hat{y}_t^{(R)}\}$. Because the levels are ordered and adjacent levels are semantically close, the draws are not treated as unordered categorical outcomes. Instead a kernel of bandwidth h is placed on each draw and normalised over the legal levels,

$$p_t(y \mid x) \;\propto\; \sum_{r=1}^{R} K\left(\frac{y - \hat{y}_t^{(r)}}{h}\right) + \epsilon, \quad y \in \mathcal{Y}, \tag{3}$$

with ϵ a small additive constant that keeps every level strictly positive. The kernel spreads a little mass onto neighbouring levels, reflecting the fact that a draw of 3.0 is weak evidence against 3.5 and strong evidence against 1.5, whereas simple counting would treat both as equally excluded. The raw point estimate is the expectation $\tilde{y}_t = \sum_y y\, p_t(y \mid x)$, and the framework also supports median, trimmed-mean and mode-aware aggregation for deployments that prefer a more robust summary.

3.2.3 Calibration of Point Scores and Distributions

To correct distortion of the score scale we fit a dimension-specific calibration function $g_t : \mathbb{R} \to \mathbb{R}$ mapping the raw estimate to a calibrated score $\bar{y}_t = g_t(\tilde{y}_t)$ subject to

$$\tilde{y}_t^{(i)} \le \tilde{y}_t^{(j)} \;\Rightarrow\; g_t(\tilde{y}_t^{(i)}) \le g_t(\tilde{y}_t^{(j)}), \tag{4}$$

so that the correction may be arbitrarily non-linear but can never reorder essays. The constraint is what makes the operation safe to apply to an assessment system, since whatever the calibrator does to the scale, a student ranked above another before calibration is still ranked above them afterwards. In the main pipeline g_t is isotonic regression fitted on the calibration split, and Section 4.2 compares this choice against

four alternatives. Fitting is per dimension, because nothing guarantees that the model distorts the vocabulary scale in the same way as the conventions scale.

The predictive distribution is calibrated in the same layer. Each individual draw is first mapped through the calibrated score space and the kernel-smoothed distribution is re-formed from the mapped draws, after which temperature scaling is applied,

$$p_t^{\mathrm{cal}}(y \mid x) = \frac{\exp(\log \tilde{p}_t(y \mid x)/T_t)}{\sum_{y' \in \mathcal{Y}} \exp(\log \tilde{p}_t(y' \mid x)/T_t)}, \tag{5}$$

where $\tilde{p}_t$ denotes the recentred distribution and $T_t > 0$ is chosen on the calibration split by minimising negative log-likelihood over a grid. Applying the map to the draws rather than to the binned histogram matters, because a step-function calibrator sends many raw levels to the same calibrated level, so remapping binned mass collapses the distribution onto a few spikes and leaves temperature scaling with nothing sensible to do. Recentring the draws instead preserves the dispersion that repeated sampling paid for.

3.2.4 Conformal Sets and Review Triage

The final stage turns the calibrated output into a decision. Using absolute residuals of the calibrated score on held-out folds of the calibration split as the nonconformity measure, we take the $\lceil (n+1)(1-\alpha) \rceil / n$ empirical quantile q_α and form the prediction set

$$C_\alpha(x) = \{\, y \in \mathcal{Y} : |y - \bar{y}_t| \le q_\alpha \,\}. \tag{6}$$

Because the residual quantile is computed on the same split that supplies the isotonic map and the temperature, this is a cross-conformal rather than a strict split-conformal construction, and the coverage we report is empirical rather than a finite-sample guarantee. The point is reinforced by the evaluation condition itself, since a deliberately reweighted test cohort violates exchangeability with the calibration data. Three quantities feed the routing rule: the width of $C_\alpha(x)$, the variance of the draws in $S_t(x)$, and the distance from the calibrated score to the nearest decision boundary. An essay with a narrow set, low sampling variance and a score far from a cut point can be returned automatically, while an essay failing any of the three becomes a candidate for human review.

3.3 A Worked Example

The aggregate results reported below establish the size of the effect; a single case shows how it arises. Table 1 traces one essay-trait pair from the ELLIPSE test split, scored for vocabulary by GLM-4.1V, whose reference score is 3.5.

The judge is asked seven times and answers 2.5 on six occasions and 3.0 once, so the raw expectation is 2.57, which rounds to 2.5 and sits a full rubric point below the reference. Nothing in the model's own behaviour signals a problem, because the draws are almost unanimous, and a system that reported sampling agreement as confidence would describe this score as highly reliable. The error is systematic rather than random, and repeated sampling cannot detect it, since every draw is displaced in the same

Table 1 Worked Example for One Essay-Trait Pair. The pair is drawn from the ELLIPSE test split (vocabulary, GLM-4.1V, reference score 3.5). Seven near-unanimous draws place the raw estimate at 2.57; the calibrated estimate is 3.67.

Score level	1.5	2.0	2.5	3.0	3.5	4.0	4.5
Raw draws (count)	0	0	6	1	0	0	0
Raw distribution	0.10	0.20	0.25	0.22	0.12	0.05	0.02
Calibrated distribution	0.00	0.01	0.06	0.17	0.28	0.27	0.15
In conformal set	–	–	–	yes	yes	yes	yes

direction. The isotonic map fitted on the calibration split sends 2.57 to 3.67, which rounds to the correct level. The conformal set at $\alpha = 0.10$ is $\{3.0, 3.5, 4.0, 4.5\}$, which contains the reference score but spans four levels, so under the routing rule this essay would be referred for human review rather than scored automatically.

3.4 Calibrator Families Compared

A single calibrated pipeline compared against its uncalibrated self cannot show whether the choice of calibrator matters, so we compare five families spanning different degrees of flexibility. All are fitted on the same calibration split, applied to the same test split, and evaluated with the same metrics, and the uncalibrated scorer is retained as the reference.

Affine rescaling fits a least-squares line and can adjust only a global offset and slope; it is the natural null hypothesis, because if a simple shift and stretch recovers the gain then nothing about the ordinal structure is doing work. Ordinal logistic calibration fits a cumulative-link model with a shared slope and $K - 1$ learned thresholds and returns the expected level, which respects the ordinal scale but constrains the shape of the correction. Beta calibration [14] fits a three-parameter monotone family on the normalised score range and is more flexible than an affine map while remaining smooth and low-variance. Isotonic regression fits the best non-decreasing step function and is the most flexible monotone option, at the cost of needing more calibration data. Histogram binning partitions the raw estimate into ten equal-frequency bins and replaces each with its bin mean; it is included precisely because it is not constrained to be monotone, and therefore isolates the contribution of the ordering constraint rather than of flexibility.

The comparison is designed so that flexibility and monotonicity vary somewhat independently. Affine and ordinal logistic calibration are monotone and rigid, isotonic regression is monotone and flexible, and histogram binning is flexible but not monotone. If flexibility alone mattered, histogram binning and isotonic regression would behave alike; if monotonicity alone mattered, the three monotone families would cluster together regardless of their flexibility.

3.5 Metrics

Let y_i and $\hat{y}_i$ be the reference and predicted scores over N test items. We report tolerance accuracy $\mathrm{Acc}_{\pm 0.5}$, mean absolute error, signed bias $\frac{1}{N}\sum_i(\hat{y}_i - y_i)$, Spearman's ρ, and quadratic weighted kappa [33],

$$\kappa_w = 1 - \frac{\sum_{j,k} w_{jk} O_{jk}}{\sum_{j,k} w_{jk} E_{jk}}, \qquad w_{jk} = \frac{(j-k)^2}{(K-1)^2}, \tag{7}$$

where O and E are the observed and chance-expected confusion matrices over K levels.

The choice of metrics is not incidental to the argument, because they divide into two groups that respond to different kinds of error. Signed bias, mean absolute error, tolerance accuracy and quadratic weighted kappa are all sensitive to where a prediction sits on the scale, so a systematic offset degrades every one of them. Spearman's ρ is invariant under strictly increasing transformations of the scale. Isotonic regression is non-decreasing rather than strictly increasing and may introduce ties, and clipping to the legal range and rounding to discrete levels may introduce further ties, so small changes in ρ are expected even though the transformation cannot reverse a strict pairwise ordering before discretisation. Reporting both groups is what licenses the central claim of this paper: if the scale-sensitive metrics improve while ρ stays essentially flat, the layer is correcting the placement of scores rather than improving the model's ability to tell essays apart. All correlations are computed on the continuous calibrated scores before rounding to legal levels.

For distributional quality we report the multiclass Brier score [34] and negative log-likelihood, which are proper scoring rules and so cannot be improved by simply flattening a distribution, together with expected calibration error [29] computed over eight equal-frequency confidence bins rather than equal-width ones, because top-label confidence for a nine-level ordinal task occupies only part of the unit interval. For conformal outputs we report empirical coverage $\frac{1}{N}\sum_i \mathbf{1}(y_i \in C_i)$ against its nominal level and the mean set size in score levels, since coverage alone can always be met by returning wider sets. For ELLIPSE all metrics are computed per trait and then averaged, so that the six traits contribute equally rather than being pooled into one confusion matrix.

3.6 Experimental Settings

We query GLM-4.1V (glm-4.1v-thinking-flashx), DeepSeek-V3-Chat and Qwen3-Max through an OpenAI-compatible interface at temperature 0.3. Each essay-dimension pair is scored up to $R = 9$ times, with early stopping once at least seven calls have returned and the modal proportion reaches 85%; across the 5,745 evaluated test pairs this yields an average of 7.15 calls, a saving of 21% against a fixed nine-call schedule. Empirical distributions use bandwidth $h = 0.75$ and Laplace smoothing $\epsilon = 0.005$. Calibrators are fitted with five-fold cross-validation on the calibration split, the temperature is selected by grid search, and conformal thresholds use out-of-fold residuals at $\alpha = 0.10$ unless stated otherwise. As a supervised reference point we fine-tune DeBERTa-base [35] with a regression head on the ELLIPSE anchor split, predicting

Table 2 Uncalibrated (u) Versus Calibrated (c) Performance in All Nine Model-Corpus Settings. Arrows indicate the preferred direction. Bias is signed; values closer to zero are better.

		QWK ↑		MAE ↓		Bias → 0		Acc±0.5 ↑		Spearman		Brier ↓		NLL ↓	
Corpus	Model	u	c	u	c	u	c	u	c	u	c	u	c	u	c
ELLIPSE	GLM-4.1V	0.183	**0.368**	0.866	**0.548**	-0.797	**-0.022**	0.435	**0.537**	0.450	0.447	0.892	**0.797**	2.23	**1.71**
ELLIPSE	DeepSeek-V3	0.202	**0.436**	0.933	**0.537**	-0.886	**-0.036**	0.400	**0.542**	0.487	0.495	0.915	**0.793**	2.33	**1.70**
ELLIPSE	Qwen3-Max	0.128	**0.321**	0.902	**0.557**	-0.839	**-0.053**	0.440	**0.540**	0.411	0.409	0.903	**0.799**	2.29	**1.75**
PERSUADE	GLM-4.1V	0.290	**0.390**	1.057	**0.906**	-0.654	**+0.014**	0.244	**0.240**	0.523	0.524	0.892	**0.751**	1.94	**1.47**
PERSUADE	DeepSeek-V3	0.236	**0.382**	1.503	**0.927**	-1.393	**-0.035**	0.179	**0.234**	0.511	0.505	1.044	**0.751**	2.85	**1.46**
PERSUADE	Qwen3-Max	0.203	**0.315**	1.418	**0.932**	-1.241	**-0.023**	0.166	**0.267**	0.450	0.456	1.012	**0.754**	2.64	**1.51**
ASAP2.0	GLM-4.1V	0.342	**0.591**	1.084	**0.993**	-0.453	**+0.034**	0.289	**0.338**	0.645	0.645	0.879	**0.768**	2.08	**1.57**
ASAP2.0	DeepSeek-V3	0.341	**0.504**	1.244	**1.019**	-0.902	**+0.009**	0.222	**0.307**	0.615	0.615	0.938	**0.779**	2.31	**1.59**
ASAP2.0	Qwen3-Max	0.235	**0.438**	1.281	**1.151**	-0.810	**+0.035**	0.228	**0.208**	0.483	0.476	0.942	**0.807**	2.38	**1.69**
Mean		0.240	**0.416**	1.143	**0.841**	-0.886	**-0.009**	0.289	**0.357**	0.508	0.508	0.935	**0.778**	2.34	**1.61**

all six traits jointly and selecting checkpoints by mean quadratic weighted kappa on the calibration split.

Because every model response is cached, all analyses in this paper are recomputed from the same stored predictions without additional API calls. This makes the comparisons internally consistent and the analyses reproducible from the cache.

4 Results

4.1 Agreement, Error and Bias Across the Nine Settings

Table 2 reports the nine model-corpus settings before and after calibration. Quadratic weighted kappa improves in all nine, from a mean of 0.240 to 0.416. Mean absolute error falls from 1.143 to 0.841 score points. The clearest effect is on systematic bias: mean signed bias moves from −0.886, that is, all three models score the perturbed essays substantially too low, to −0.009, and mean absolute bias falls from 0.886 to 0.029, so the residual displacement is both small and no longer in a consistent direction. Tolerance accuracy improves in seven of the nine settings, from a mean of 0.289 to 0.357, and declines slightly in two: by 0.004 for GLM-4.1V on PERSUADE and by 0.020 for Qwen3-Max on ASAP 2.0. Both are threshold effects rather than reversals, since $\mathrm{Acc}_{\pm 0.5}$ scores a prediction only on one side of a fixed cut, and mean absolute error falls in both of those settings as well. The Brier score and negative log-likelihood of the predictive distributions fall throughout.

Spearman's ρ is the informative exception, remaining at a mean of 0.508 before and after calibration. This follows from the monotonicity constraint and indicates what the layer is actually doing. The models were already ordering essays reasonably well; what they were getting wrong was where that ordering sat on the rubric scale. Figure 2 shows the pattern setting by setting.

4.2 Comparison of Calibrator Families

Were the improvement above specific to isotonic regression, it would be an artefact of estimator choice rather than a property of the correction. Table 3 and Figure 3 show that it is not. Every monotone family recovers most of the gain and reduces mean absolute bias from 0.886 to roughly 0.03. Isotonic regression attains the best mean

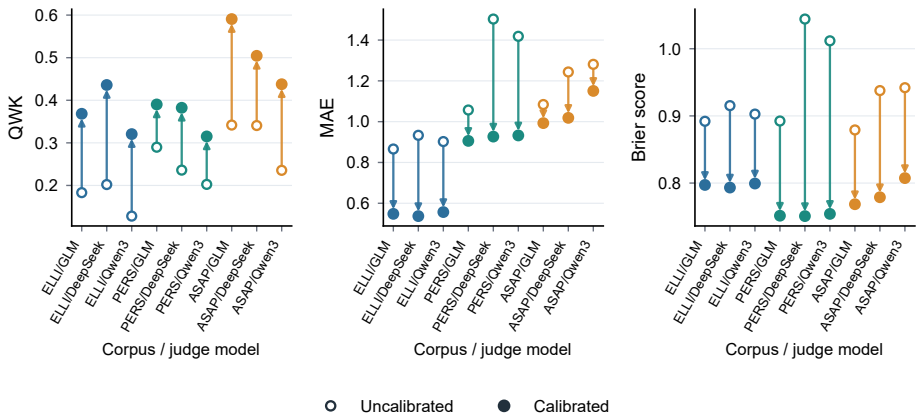

Fig. 2 Effect of Calibration in Each Model-Corpus Setting. Open circles are uncalibrated values, filled circles calibrated, and arrows show the direction of change. Colour identifies the corpus.

Table 3 Calibrator Families Under the Same Protocol. Per-corpus columns are mean QWK across the three models; the last three columns are means over all nine settings.

	QWK by corpus ↑			Mean over nine		
Calibrator	ELL.	PER.	ASAP	QWK	MAE	|Bias|
Uncalibrated	0.171	0.243	0.306	0.240	1.143	0.886
Histogram	0.377	0.314	0.432	0.374	0.862	0.027
Affine	0.388	0.302	0.478	0.389	0.839	0.027
Ordinal logistic	0.386	0.314	0.475	0.392	0.840	0.028
Beta	0.388	0.356	0.502	0.415	0.842	0.025
Isotonic	0.375	0.363	0.511	0.416	0.841	0.029

kappa at 0.416, but beta calibration is within a thousandth of it at 0.415, and ordinal logistic and affine calibration follow at 0.392 and 0.389. Histogram binning, the one family in the comparison that is not constrained to be monotone, is clearly the weakest of the calibrated variants at 0.374.

Counted setting by setting rather than on the mean, no family dominates: isotonic wins three of the nine settings, beta and ordinal logistic two each, and affine and histogram one each. What matters is therefore that the correction be order-preserving on a distorted ordinal scale; the choice among reasonable monotone estimators is a second-order decision. Practitioners can select on grounds of convenience, with the caveat that isotonic regression is the most flexible and the most data-hungry, and that beta calibration is the safer choice on small calibration splits.

4.3 Sampling Budget and Inference Cost

Repeated sampling is the expensive part of the pipeline. Table 5 and Figure 4 reduce the budget from nine draws to one, refitting the calibrator each time. A single draw

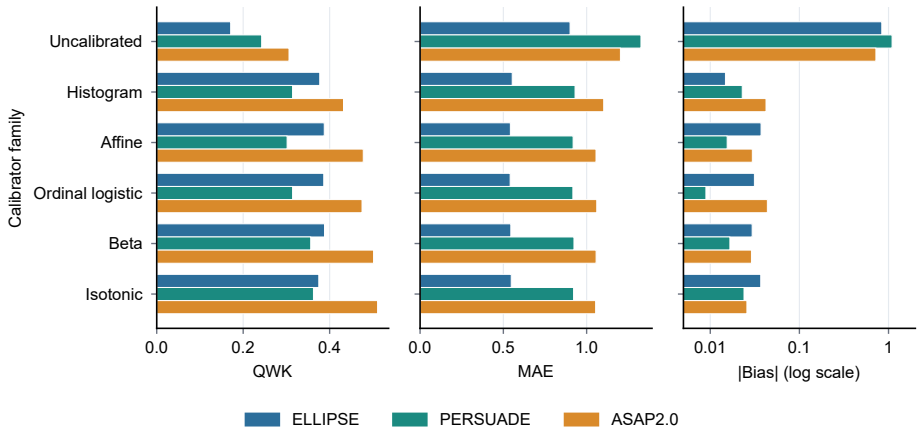

Fig. 3 Calibrator Families Compared on Agreement, Error and Absolute Bias. All monotone families cluster together and non-monotone histogram binning trails them. Absolute bias is drawn on a logarithmic axis because every calibrator reduces it by more than an order of magnitude relative to the uncalibrated scorer.

already retains 87% of the agreement obtainable at full budget, and three draws retain 94%, so the marginal kappa from the fourth draw onwards is small.

This is not, however, an argument for always using $R = 1$. The empirical distribution over levels is what temperature scaling calibrates and what the conformal sets and triage rule consume, and a single draw provides no such distribution at all. The budget question is therefore not how much agreement is lost but whether the deployment needs uncertainty outputs. A system that returns only a point score can run at $R = 1$ or $R = 3$ with an isotonic calibrator and pay a small penalty, whereas a system that routes low-confidence essays to human readers needs the samples, because that is what the routing signal is built from.

Table 4 makes the absolute cost concrete, since the multiplier differs sharply between analytic and holistic rubrics. On the two holistic corpora an essay costs about 7.1 calls. On ELLIPSE, where six traits are scored independently, the same policy costs about 43 calls per essay, so a trait-wise design is roughly six times more expensive than the holistic figure that is usually quoted for LLM scoring. Adaptive stopping saves about 20% against a fixed nine-call schedule in all three corpora. We do not convert these counts into monetary figures, because we did not log per-request token usage and any price estimate would depend on tariffs that change frequently; the call counts are the quantity a reader can re-price for their own endpoint. The calibration machinery itself is negligible beside this, since fitting the isotonic map, the temperature and the conformal quantile is a one-off cost on a few hundred cached scores and adds no per-essay inference.

4.4 Statistical Reliability of the Improvements

Table 6 and Figure 5 report a paired bootstrap over test essays with 500 resamples for each of the nine settings and five metrics. Resampling is at the essay level, so that

Table 4 Observed Inference Cost Under Adaptive Stopping. Calls per essay is the per-pair figure multiplied by the number of scored traits.

Corpus	Traits	Calls/pair	Calls/essay	Fixed R=9	Saving
ELLIPSE	6	7.17	43.0	54	20%
PERSUADE	1	7.16	7.2	9	20%
ASAP2.0	1	7.12	7.1	9	21%

Table 5 Sampling Budget Sweep With Isotonic Calibration. Per-corpus columns are means across the three models.

	QWK ↑				MAE ↓				
R	ELL.	PER.	ASAP	Mean	ELL.	PER.	ASAP	Mean	%
1	0.361	0.302	0.426	0.363	0.552	0.940	1.072	0.855	87%
3	0.368	0.320	0.481	0.390	0.551	0.930	1.061	0.847	94%
5	0.381	0.319	0.492	0.397	0.548	0.920	1.056	0.841	96%
7	0.373	0.362	0.511	0.415	0.548	0.923	1.055	0.842	100%
9	0.375	0.363	0.511	0.416	0.547	0.922	1.054	0.841	100%

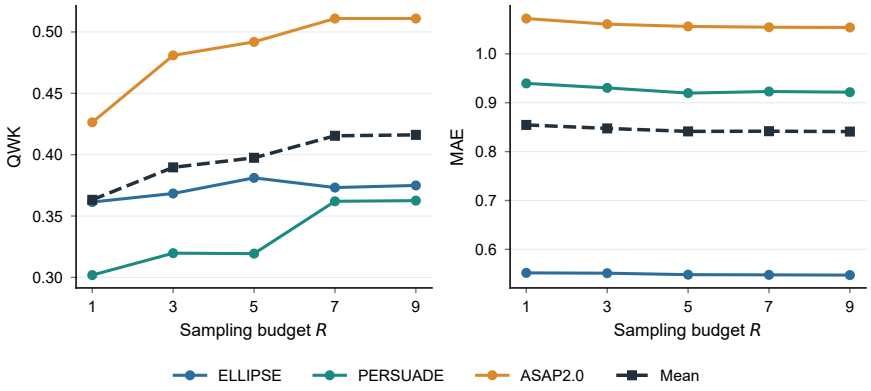

Fig. 4 Agreement and Error as a Function of the Sampling Budget R. Coloured lines are corpora; the dashed line is the mean over all nine settings.

the six ELLIPSE traits belonging to one essay move together and the intervals are not narrowed by treating traits as independent observations.

All 45 metric comparisons have 95% intervals that exclude zero, so every improvement in kappa, mean absolute error, absolute bias, Brier score and negative log-likelihood is separated from no effect in every setting. The intervals are widest, and the medians smallest, for GLM-4.1V on PERSUADE, which is also the setting with the least uncalibrated bias to remove; its kappa improvement of 0.100 has an interval of $[0.013, 0.177]$ that only just clears zero. The pattern is consistent with the

Table 6 Paired Bootstrap of the Calibrated Minus Uncalibrated Improvement. Intervals are 95% percentile bounds from 500 essay-level resamples. Positive values favour calibration for every metric shown.

Corpus	Model	ΔQWK	ΔMAE	Δ|Bias|	ΔBrier	ΔNLL
ELLIPSE	GLM-4.1V	0.182 [0.113, 0.242]	0.318 [0.226, 0.411]	0.766 [0.587, 0.798]	0.095 [0.063, 0.124]	0.51 [0.37, 0.66]
ELLIPSE	DeepSeek-V3	0.231 [0.177, 0.276]	0.396 [0.301, 0.487]	0.847 [0.725, 0.864]	0.122 [0.089, 0.153]	0.63 [0.49, 0.76]
ELLIPSE	Qwen3-Max	0.191 [0.130, 0.245]	0.350 [0.259, 0.439]	0.783 [0.662, 0.810]	0.105 [0.075, 0.134]	0.54 [0.42, 0.68]
PERSUADE	GLM-4.1V	0.100 [0.013, 0.177]	0.147 [0.043, 0.250]	0.631 [0.329, 0.694]	0.139 [0.093, 0.185]	0.46 [0.31, 0.63]
PERSUADE	DeepSeek-V3	0.139 [0.065, 0.210]	0.568 [0.418, 0.703]	1.346 [1.108, 1.386]	0.291 [0.223, 0.350]	1.38 [1.13, 1.61]
PERSUADE	Qwen3-Max	0.110 [0.052, 0.170]	0.483 [0.339, 0.619]	1.202 [0.927, 1.254]	0.256 [0.198, 0.312]	1.12 [0.90, 1.36]
ASAP2.0	GLM-4.1V	0.249 [0.219, 0.277]	0.092 [0.049, 0.136]	0.422 [0.250, 0.508]	0.111 [0.089, 0.132]	0.51 [0.43, 0.59]
ASAP2.0	DeepSeek-V3	0.162 [0.134, 0.189]	0.226 [0.160, 0.285]	0.887 [0.712, 0.934]	0.160 [0.130, 0.184]	0.72 [0.61, 0.82]
ASAP2.0	Qwen3-Max	0.202 [0.179, 0.224]	0.130 [0.070, 0.190]	0.782 [0.597, 0.859]	0.134 [0.110, 0.159]	0.69 [0.59, 0.79]

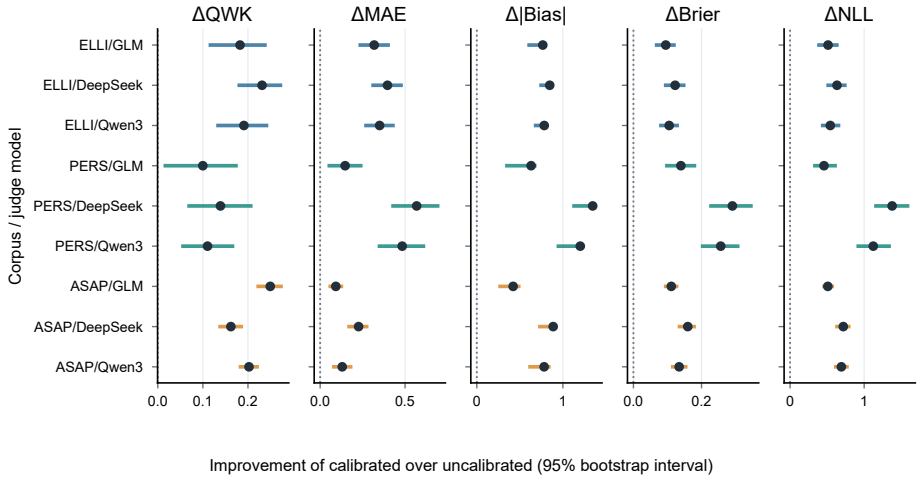

Fig. 5 Paired Bootstrap Intervals in All Nine Settings. Points are medians, bars are 95% intervals, and the dotted line marks no effect.

mechanism described in Section 5, in that the size of the gain tracks the size of the scale distortion being corrected.

4.5 Component Ablation

Table 7 separates the contributions of the two expensive components. Configuration (A) makes a single anchored call with no calibration, (B) adds isotonic calibration to that single call, (C) restores the full sampling budget without calibration, and (D) is the complete layer.

The two components turn out not to be independent. Adding calibration to a single call, configuration (B), raises kappa by 0.125 and removes essentially all of the bias, from 0.892 to 0.013. Adding sampling without calibration, configuration (C), changes almost nothing on any measure: kappa moves by 0.002, the Brier score by 0.006 and the negative log-likelihood by 0.029. Repeated sampling on its own therefore buys almost nothing.

Table 7 Component Ablation Averaged Over the Nine Settings. Per-corpus columns are mean QWK.

	QWK by corpus			Mean over nine settings				
Configuration	ELL.	PER.	ASAP	QWK	MAE	|Bias|	Brier	NLL
(A) Single call	0.169	0.235	0.310	0.238	1.148	0.892	0.941	2.37
(B) A + isotonic	0.361	0.302	0.426	0.363	0.855	0.013	0.941	2.37
(C) A + sampling	0.171	0.243	0.306	0.240	1.143	0.886	0.935	2.34
(D) Full layer	0.375	0.363	0.511	0.416	0.841	0.009	0.778	1.61

Its value appears only in combination. Going from (C) to the full layer improves the Brier score by 0.157 and the negative log-likelihood by 0.732, far more than calibration delivers on a single call, where the distributional metrics are unchanged by construction because a point map cannot alter a one-sample distribution. Comparing (B) with (D) isolates what the extra draws contribute once calibration is in place: a further 0.053 in kappa and 0.163 in Brier score. Sampling is best understood not as a second source of accuracy but as the raw material that the distributional half of the calibration layer needs in order to operate. A deployment that returns only point scores can drop it and lose little; one that needs calibrated probabilities cannot.

4.6 Cross-Corpus Transfer

A calibrator fitted on the corpus it will be used on is the easy case. Table 8 and Figure 6 test the harder one: fit the isotonic map on one corpus, apply it unchanged to another after normalising both score ranges to a common interval, and compare against both the uncalibrated scorer and the in-domain calibrator on the target.

Transfer helps in almost every pair. Mean absolute error improves in 17 of 18 source-target-model combinations, absolute bias in 16, and kappa in 16. The median transferred calibrator recovers about three quarters of the error reduction that an in-domain calibrator achieves.

The scope of this result needs stating precisely. The transferred condition uses no target calibration labels, but the scorer still retrieves anchors from the target-domain scored training pool, so the setting is free of target calibration labels rather than free of target-domain labels. A fully label-free target condition, in which the model receives only the verbal rubric or anchors drawn from the source corpus, is not evaluated here. The transfer is also between corpora that share a broadly similar construct and a comparable ordinal scale, and it says nothing about transfer across languages, rubrics of different dimensionality, or model versions.

4.7 Displacement Injected into the Predictions

The main evaluation condition perturbs the reference labels, which makes it a synthetic target perturbation rather than a model failure. A reader may reasonably object that a calibrator asked to undo a displacement of the labels is solving an easier and differently shaped problem than one asked to undo a displacement of the model. We therefore repeated the analysis with the perturbation moved to the other side of the comparison:

Table 8 Cross-Corpus Transfer. The calibrator is fitted on the source corpus and applied unchanged to the target in normalised score space. Bold marks an improvement over the uncalibrated scorer. Recovery is the fraction of the in-domain MAE reduction achieved.

		MAE ↓			|Bias| ↓		QWK ↑		
Model	Transfer	uncal.	transf.	in-dom.	uncal.	transf.	uncal.	transf.	Recov.
GLM-4.1V	ELLIPSE → PERSUADE	1.057	**1.028**	0.906	0.654	0.803	0.290	0.262	19%
GLM-4.1V	ELLIPSE → ASAP2.0	1.084	1.132	0.993	0.453	0.603	0.342	**0.429**	-52%
GLM-4.1V	PERSUADE → ELLIPSE	0.866	**0.767**	0.548	0.797	**0.642**	0.183	**0.205**	31%
GLM-4.1V	PERSUADE → ASAP2.0	1.084	**1.056**	0.993	0.453	**0.048**	0.342	**0.432**	31%
GLM-4.1V	ASAP2.0 → ELLIPSE	0.866	**0.857**	0.548	0.797	**0.581**	0.183	**0.355**	3%
GLM-4.1V	ASAP2.0 → PERSUADE	1.057	**0.985**	0.906	0.654	**0.529**	0.290	**0.339**	47%
DeepSeek-V3	ELLIPSE → PERSUADE	1.503	**0.933**	0.927	1.393	**0.191**	0.236	**0.422**	99%
DeepSeek-V3	ELLIPSE → ASAP2.0	1.244	**1.031**	1.019	0.902	**0.225**	0.341	**0.437**	94%
DeepSeek-V3	PERSUADE → ELLIPSE	0.933	**0.671**	0.537	0.886	**0.402**	0.202	0.192	66%
DeepSeek-V3	PERSUADE → ASAP2.0	1.244	**1.143**	1.019	0.902	**0.499**	0.341	**0.490**	45%
DeepSeek-V3	ASAP2.0 → ELLIPSE	0.933	**0.633**	0.537	0.886	**0.054**	0.202	**0.385**	76%
DeepSeek-V3	ASAP2.0 → PERSUADE	1.503	**0.929**	0.927	1.393	**0.098**	0.236	**0.483**	100%
Qwen3-Max	ELLIPSE → PERSUADE	1.418	**0.922**	0.932	1.241	**0.198**	0.203	**0.394**	102%
Qwen3-Max	ELLIPSE → ASAP2.0	1.281	**1.170**	1.151	0.810	**0.331**	0.235	**0.312**	85%
Qwen3-Max	PERSUADE → ELLIPSE	0.902	**0.638**	0.557	0.839	**0.352**	0.128	**0.243**	76%
Qwen3-Max	PERSUADE → ASAP2.0	1.281	**1.177**	1.151	0.810	**0.206**	0.235	**0.427**	80%
Qwen3-Max	ASAP2.0 → ELLIPSE	0.902	**0.613**	0.557	0.839	**0.001**	0.128	**0.362**	84%
Qwen3-Max	ASAP2.0 → PERSUADE	1.418	**0.909**	0.932	1.241	**0.003**	0.203	**0.406**	105%

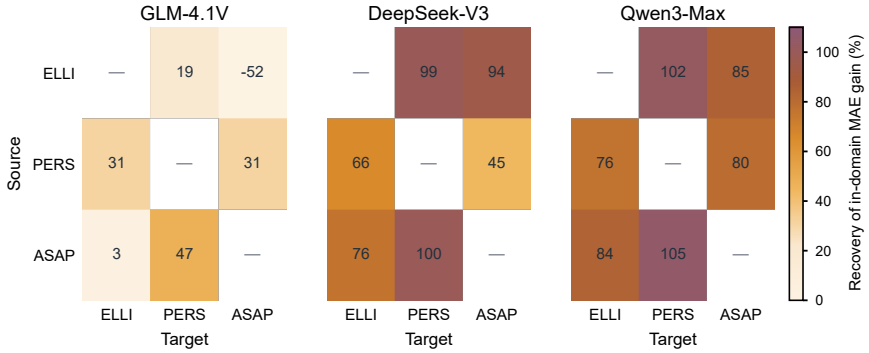

Fig. 6 Recovery of the In-Domain Mean Absolute Error Reduction. Rows are the source corpus supplying the calibrator and columns the target corpus it is applied to. Values above 100 indicate that the transferred calibrator matched or exceeded the in-domain one; the single negative cell is the one transfer pair in which mean absolute error increased.

the original labels are left untouched, and a length-dependent displacement of the same magnitude is added to the model's own point estimates before calibration, with the calibration split displaced in the same way so that no oracle information leaks.

Table 9 reports the outcome. Calibration recovers this displacement as effectively as the label-side one. Mean kappa rises from 0.520 to 0.704, mean absolute error falls from 0.973 to 0.626, and mean absolute bias falls from 0.888 to 0.072. The improvement holds on all three corpora, with the largest gain on ELLIPSE, where kappa nearly doubles from 0.354 to 0.632.

This is the more faithful analogue of the length preference documented for LLM judges, and it strengthens the central claim, because the correction is now being asked to remove a distortion of the scorer rather than of the scale it is measured against.

Table 9 Displacement Injected Into the Model Predictions Rather Than the Reference Labels. Values are averaged over the three judge models. Bias is absolute.

	QWK ↑		MAE ↓		|Bias| ↓	
Corpus	uncal.	cal.	uncal.	cal.	uncal.	cal.
ELLIPSE	0.354	0.632	0.867	0.491	0.840	0.033
PERSUADE	0.538	0.681	1.159	0.612	1.100	0.024
ASAP2.0	0.668	0.798	0.893	0.777	0.723	0.157
Mean	0.520	0.704	0.973	0.626	0.888	0.072

Two caveats remain. The injected displacement still has a known parametric form, so it is easier than an unanticipated shift, and the underlying reference labels here are the perturbed corpus labels used throughout, so this experiment isolates the direction of the injection rather than providing the fully unperturbed baseline discussed in Section 5.

4.8 Topic Shift within a Corpus

Cross-corpus transfer changes the corpus, the rubric and the population at once. A narrower and more realistic question is whether a calibrator estimated on one set of writing prompts still works on an unseen prompt from the same programme. ASAP 2.0 supports this directly, because its essays are written to three distinct source-based tasks with 336, 302 and 181 essays in the test split. We hold out one prompt at a time, fit the isotonic map only on calibration essays from the other two prompts, and evaluate on the held-out prompt. For reference we also fit a same-prompt calibrator on the calibration essays belonging to the held-out topic.

Table 10 reports the nine folds. The topic-disjoint calibrator improves kappa in 8 of 9 folds, from a mean of 0.281 to 0.402, and reduces absolute bias in 6 of 9, from 0.673 to 0.373. A same-topic calibrator does better still, reaching kappa 0.448 and absolute bias 0.117, so topic identity carries real information that a disjoint calibrator cannot access. Mean absolute error is the weakest result, improving in only 5 of 9 folds, which is consistent with the pattern seen elsewhere in this paper that the layer corrects scale placement more reliably than it reduces per-essay error.

One boundary of the design deserves emphasis: only the calibrator is topic-disjoint. The anchor bank was constructed before caching and may contain essays written to the held-out prompt, so the experiment does not rule out retrieval leakage from same-topic anchors; a fully prompt-disjoint pipeline would require re-running the scorer with a restricted anchor bank.

4.9 Predictive Distributions and Conformal Sets

Calibration substantially improves the quality of the predictive distributions. Table 11 reports expected calibration error before and after, which falls from a mean of 0.234 to 0.073 and improves in all nine settings.

Table 10 ASAP 2.0 Leave-One-Prompt-Out. The topic-disjoint calibrator is fitted only on calibration essays from the other two prompts; the same-topic calibrator uses essays from the held-out prompt.

			QWK ↑			MAE ↓		
Model	Held-out prompt	n	uncal.	disj.	same	uncal.	disj.	same
GLM-4.1V	A Cowboy Who Rode the Waves	183	0.312	0.433	0.314	0.812	0.916	0.860
GLM-4.1V	Does the electoral college…	301	0.284	0.450	0.632	1.293	1.067	0.926
GLM-4.1V	The Face on Mars	336	0.338	0.499	0.512	1.044	1.039	1.032
DeepSeek-V3	A Cowboy Who Rode the Waves	183	0.302	0.309	0.509	0.920	0.980	0.910
DeepSeek-V3	Does the electoral college…	301	0.289	0.497	0.581	1.558	1.104	1.003
DeepSeek-V3	The Face on Mars	336	0.360	0.495	0.444	1.138	1.014	1.012
Qwen3-Max	A Cowboy Who Rode the Waves	183	0.248	0.237	0.391	0.962	0.966	0.947
Qwen3-Max	Does the electoral college…	301	0.190	0.442	0.335	1.570	1.218	1.098
Qwen3-Max	The Face on Mars	336	0.205	0.258	0.318	1.196	1.318	1.222
Mean			0.281	0.402	0.448	1.166	1.069	1.001

Table 11 Expected Calibration Error of the Predictive Distributions. Values are computed over eight equal-frequency confidence bins.

	GLM-4.1V		DeepSeek-V3		Qwen3-Max	
Corpus	uncal.	cal.	uncal.	cal.	uncal.	cal.
ELLIPSE	0.100	**0.063**	0.131	**0.057**	0.120	**0.040**
PERSUADE	0.261	**0.107**	0.338	**0.139**	0.347	**0.068**
ASAP2.0	0.220	**0.021**	0.301	**0.063**	0.291	**0.100**

The direction of the miscalibration explains why the correction helps. Pooled across settings, the uncalibrated distributions assign a mean confidence of 0.44 to their top-ranked level while that level is correct only 0.20 of the time, so they are over-confident by more than a factor of two. This is the expected consequence of forming a distribution from a handful of discrete draws that frequently agree with one another, because agreement among seven samples of a displaced scorer is not evidence of correctness. After the draws are recentred on the calibrated score and temperature-scaled, mean confidence falls to 0.32 while accuracy rises to 0.28, leaving a residual gap of 0.04 rather than 0.24. The fitted temperatures lie between 0.75 and 2.75 across the nine settings.

Table 12 and Figure 7 sweep the conformal risk level. Empirical coverage tracks the nominal level closely across all three corpora. The practical question is set width. At $\alpha = 0.10$ the mean set spans close to four score levels, which is too wide to be useful as an automatic score on a six-level scale, and only a minority of essays receive sets of three levels or fewer. Relaxing to $\alpha = 0.20$ or 0.30 narrows sets considerably at a coverage cost that may be acceptable when the set is used to triage rather than to certify. As noted in Section 3, these are empirical coverage figures under a deliberately shifted test distribution, not finite-sample guarantees.

4.10 Simulated Review Triage

To ask what the uncertainty output is worth operationally, we simulate a pass/fail decision at the midpoint of each scale and a human-review budget that resolves a fixed

Table 12 Conformal Prediction Sets as a Function of the Risk Level α. Coverage columns are means across the three models; set size is in score levels.

		Empirical coverage					
α	Nom.	ELL.	PER.	ASAP	Mean	Size	≤ 3
0.05	0.95	0.926	0.973	0.974	0.958	4.63	3%
0.10	0.90	0.863	0.938	0.900	0.901	3.88	16%
0.15	0.85	0.825	0.885	0.876	0.862	3.53	45%
0.20	0.80	0.771	0.871	0.841	0.828	3.32	64%
0.30	0.70	0.656	0.759	0.706	0.707	2.64	100%

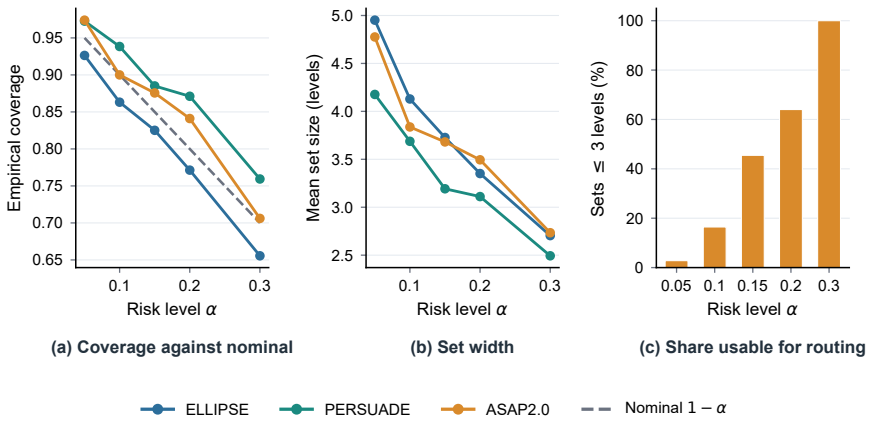

Fig. 7 Conformal Behaviour Across Risk Levels. Panel (a) plots empirical coverage against the nominal level $1 - \alpha$. Panel (b) shows the mean set width in score levels. Panel (c) shows the share of essays whose set spans at most three levels.

fraction of essays correctly. Essays are referred in decreasing order of the probability mass straddling the cut score. Table 13 and Figure 8 report the resulting error rates.

Most of the benefit comes from calibration itself: the pass/fail error rate falls from 0.471 to 0.225 before any essay is referred, because a systematically depressed score scale misclassifies large numbers of essays that sit near the boundary. Given calibrated scores, referring on calibrated uncertainty is then clearly better than referring at random. At a twenty per cent review budget the error rate falls to 0.144 with uncertainty-based referral against 0.180 for random referral, while uncertainty-based referral applied to uncalibrated scores stays far behind at 0.318. The largest operational gain is therefore scale correction, with uncertainty-based routing a real but secondary refinement.

4.11 Subgroup Disparities

An aggregate improvement is not sufficient evidence for use in assessment, because a correction that helps on average can still concentrate error in particular groups. The

Table 13 Simulated Pass/Fail Error Rate Under a Human-Review Budget. Rates are taken at the scale midpoint and averaged over the nine settings. Referral resolves the reviewed essays correctly.

Budget	Uncal. + unc.	Cal. + random	Cal. + unc.
0%	0.471	0.225	**0.225**
5%	0.429	0.214	**0.200**
10%	0.390	0.202	**0.180**
15%	0.353	0.191	**0.160**
20%	0.318	0.180	**0.144**
30%	0.254	0.157	**0.114**
50%	0.135	0.112	**0.062**

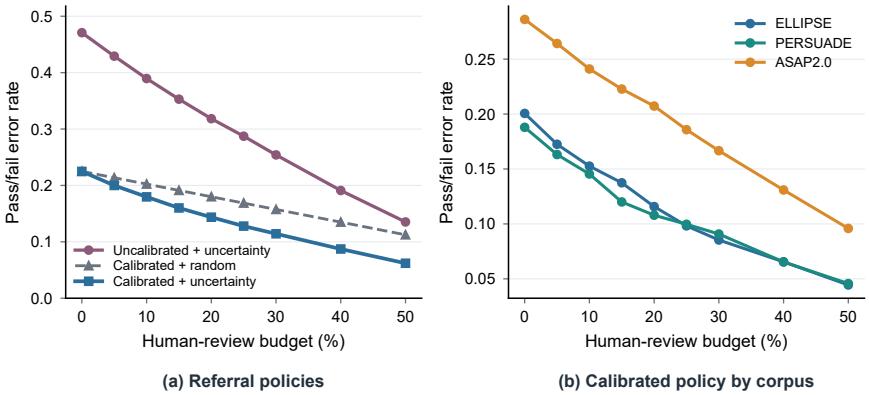

Fig. 8 Pass/Fail Error Rate Against Human-Review Budget. Panel (a) compares referral policies averaged over all nine settings. Panel (b) breaks the calibrated uncertainty policy down by corpus.

three corpora carry demographic metadata, which allows this to be checked directly. We consider gender, race and ethnicity, socioeconomic status, English language learner status and grade level, retain any group with at least 25 test essays, and compute for each attribute the gap between the best-served and worst-served group.

Table 14 and Figure 9 report the point estimates across 27 attribute-by-setting comparisons. The direction is consistent: the between-group gap in mean absolute error narrows in 24 of 27 comparisons, the gap in conformal coverage in 25 of 27, and the gap in signed bias in 20 of 27. Averaged over comparisons the error gap falls from 0.194 to 0.075 score points and the coverage gap from 0.111 to 0.038. Table 16 gives the per-group detail for one judge model, where uncalibrated bias ranges widely across groups and calibrated bias is compressed towards zero for all of them.

Counts of this kind say nothing about the reliability of the individual reductions, so we estimated intervals for them. Table 15 reports an essay-level paired bootstrap with 2,000 resamples for every attribute-by-setting comparison, in which all six ELLIPSE traits move with their essay. The intervals are much weaker than the counts suggest: although the mean reduction in the error gap is 0.113 score points, the 95% interval excludes zero in only 2 of the 27 comparisons, and for the coverage gap in only 4 of

Table 14 Between-Group Disparities Before and After Calibration. Values are averaged over attribute-by-setting comparisons. Each gap is the difference between the best-served and worst-served group within an attribute.

		Bias gap ↓		MAE gap ↓		Cov. gap ↓	
Attribute	n	uncal.	cal.	uncal.	cal.	uncal.	cal.
ELL status	3	0.502	**0.297**	0.406	**0.129**	0.233	**0.029**
Gender	9	0.252	**0.230**	0.140	**0.043**	0.083	**0.020**
Grade	3	0.285	**0.314**	0.261	**0.083**	0.169	**0.070**
Race/ethnicity	6	0.563	**0.486**	0.268	**0.113**	0.134	**0.072**
SES	6	0.189	**0.140**	0.062	**0.054**	0.040	**0.018**
All	27	0.339	**0.284**	0.194	**0.075**	0.111	**0.038**

Table 15 Essay-Level Paired Bootstrap of the Between-Group Error Gap Reduction. Intervals come from 2,000 resamples. Positive values favour calibration. The last column counts comparisons whose 95% interval excludes zero.

		MAE gap			
Attribute	n	uncal.	cal.	Reduction [95% CI]	CI > 0
ELL status	3	0.406	0.129	0.254 [-0.013, 0.486]	1/3
Gender	9	0.140	0.043	0.080 [-0.095, 0.275]	0/9
Grade	3	0.261	0.083	0.163 [-0.040, 0.385]	0/3
Race/ethnicity	6	0.268	0.113	0.167 [-0.116, 0.495]	1/6
SES	6	0.062	0.054	0.012 [-0.118, 0.188]	0/6
All	27	0.194	0.075	0.113 [-0.089, 0.340]	2/27

27. Three comparisons show a widened error gap in the point estimate. The subgroups are simply too small for individual reductions to be resolved at this sample size, and no multiplicity correction has been applied.

The evidence therefore supports a narrower claim than the counts alone would suggest. Under this controlled perturbation and for the attributes recorded in these corpora, calibration did not achieve its aggregate gains at the expense of any measured subgroup, and the descriptive disparities moved consistently in the favourable direction. That consistency of direction across 27 largely independent comparisons is itself informative. But the individual reductions are mostly not statistically distinguishable from zero, the analysis establishes no fairness guarantee, and it may not generalise to naturally occurring deployment shifts, unmeasured attributes, intersectional groups, or future model versions. Groups below the 25-essay threshold are excluded, which removes precisely the cases where disparities are most likely to be severe.

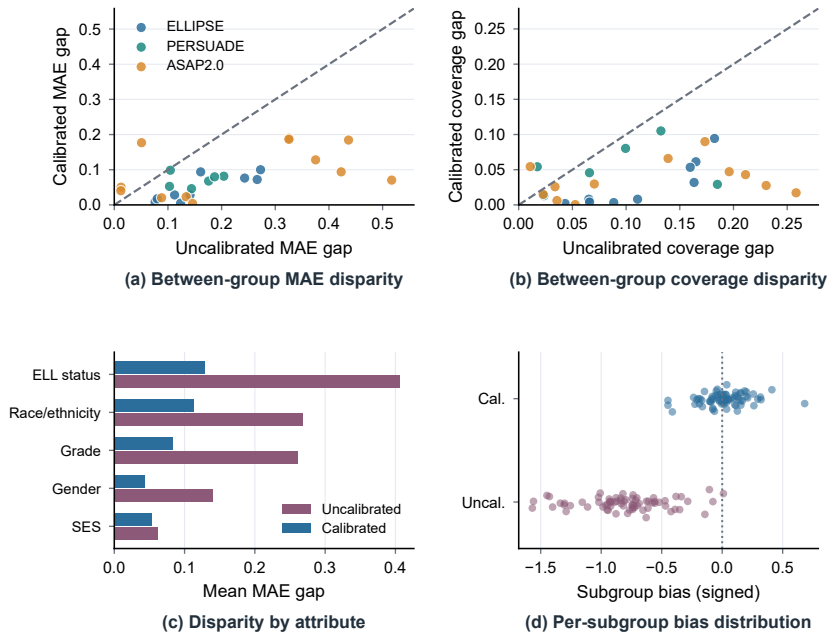

Fig. 9 Subgroup Analysis. Panels (a) and (b) plot calibrated against uncalibrated between-group gaps, where points below the diagonal indicate reduced disparity. Panel (c) averages the error gap by attribute. Panel (d) shows the distribution of signed bias over all subgroups.

Table 16 Per-Subgroup Detail for the GLM-4.1V Judge. Coverage is at $\alpha = 0.10$.

				Bias		MAE		Coverage	
Corpus	Attribute	Group	n	uncal.	cal.	uncal.	cal.	uncal.	cal.
ELLIPSE	Gender	Female	45	-0.888	-0.098	0.931	0.551	0.459	0.852
ELLIPSE	Gender	Male	50	-0.715	+0.047	0.807	0.546	0.570	0.860
ELLIPSE	SES	Economically disadvantaged	62	-0.835	-0.049	0.892	0.545	0.495	0.858
ELLIPSE	SES	Not economically disadvantaged	33	-0.726	+0.029	0.816	0.553	0.561	0.854
ELLIPSE	Race/ethnicity	Hispanic/Latino	68	-0.739	+0.029	0.807	0.545	0.559	0.865
ELLIPSE	Grade	8	28	-0.614	+0.166	0.716	0.503	0.619	0.905
ELLIPSE	Grade	11	29	-0.924	-0.169	0.989	0.603	0.437	0.810
ELLIPSE	Grade	12	33	-0.821	-0.040	0.875	0.565	0.510	0.843
PERSUADE	Gender	F	77	-0.811	-0.164	1.121	0.873	0.755	0.970
PERSUADE	Gender	M	143	-0.558	+0.123	1.018	0.926	0.778	0.957
PERSUADE	Race/ethnicity	Black/African American	125	-0.629	+0.037	1.068	0.939	0.756	0.959
PERSUADE	Race/ethnicity	Hispanic/Latino	41	-0.730	-0.064	1.035	0.879	0.764	0.991
PERSUADE	Race/ethnicity	White	33	-0.521	+0.155	0.963	0.841	0.822	0.945
ASAP2.0	Gender	F	469	-0.578	-0.057	1.122	1.002	0.657	0.900
ASAP2.0	Gender	M	351	-0.285	+0.154	1.034	0.981	0.692	0.906
ASAP2.0	SES	Economically disadvantaged	287	-0.077	+0.294	0.957	1.006	0.714	0.881
ASAP2.0	SES	Not economically disadvantaged	232	-0.351	+0.099	0.969	0.966	0.737	0.897
ASAP2.0	Race/ethnicity	Asian/Pacific Islander	30	-0.930	-0.448	1.130	1.031	0.633	0.867
ASAP2.0	Race/ethnicity	Black/African American	128	+0.010	+0.316	0.898	0.990	0.797	0.883
ASAP2.0	Race/ethnicity	Hispanic/Latino	224	-0.423	+0.015	1.089	0.986	0.696	0.906
ASAP2.0	Race/ethnicity	Two or more races/Other	46	-0.339	+0.155	0.835	0.844	0.804	0.913
ASAP2.0	Race/ethnicity	White	383	-0.609	-0.036	1.161	1.007	0.608	0.914
ASAP2.0	ELL status	No	788	-0.468	+0.033	1.098	1.002	0.662	0.901
ASAP2.0	ELL status	Yes	28	-0.106	+0.037	0.773	0.812	0.893	0.929

Table 17 ELLIPSE Analytic Traits. Uncalibrated (u) and calibrated (c) QWK are given for each judge model, with mean absolute error averaged over models.

	GLM-4.1V		DeepSeek		Qwen3-Max		MAE	
Trait	u	c	u	c	u	c	u	c
Cohesion	0.243	**0.456**	0.215	**0.514**	0.162	**0.293**	0.798	**0.521**
Syntax	0.191	**0.297**	0.190	**0.462**	0.104	**0.289**	0.806	**0.551**
Vocabulary	0.139	**0.375**	0.135	**0.366**	0.101	**0.363**	1.084	**0.507**
Phraseology	0.161	**0.340**	0.190	**0.395**	0.086	**0.301**	1.010	**0.572**
Grammar	0.136	**0.393**	0.188	**0.445**	0.131	**0.364**	0.930	**0.576**
Conventions	0.229	**0.349**	0.294	**0.435**	0.184	**0.312**	0.774	**0.556**
Mean	0.183	**0.368**	0.202	**0.436**	0.128	**0.321**	0.900	**0.547**

4.12 Behaviour across Analytic Traits

ELLIPSE scores six analytic traits separately, which allows a check of whether the layer works uniformly across the constructs a rubric measures or only on some of them. Table 17 reports each trait before and after calibration for the three judge models.

Averaged over models and traits, kappa rises from 0.171 to 0.375, and the improvement holds for every trait. This matters for a rubric-based system, because a correction that worked on some constructs and not others would change the relative weight the rubric places on them. What the layer does not do is equalise the traits: those that were harder to score before calibration remain comparatively harder afterwards, which is the expected behaviour of a correction applied to the score scale rather than to the model's representation of the construct.

4.13 Comparison with a Supervised Scorer

For context we fine-tuned DeBERTa-base on the ELLIPSE anchor split. It reaches a mean kappa of 0.712 and mean absolute error of 0.410 across the six traits, with per-trait kappa ranging from 0.653 on conventions to 0.769 on phraseology. The best calibrated prompted scorer on the same corpus reaches a kappa of 0.436, against 0.202 for the same model without calibration.

The comparison is deliberately asymmetric, and only a narrow conclusion follows from it. A supervised encoder trained on in-domain reference scores sets a higher ceiling for agreement than a frozen LLM with a lightweight calibrator, and where such training data exists it should be used. The setting this work addresses is the one where it does not: a new rubric, a new cohort, or a constraint that prevents in-domain training. In that setting the relevant comparison is not against DeBERTa but against the uncalibrated prompted scorer, which the calibration layer improves on every aggregate measure examined here.

5 Discussion

Post-hoc calibration is not usually this effective, so the size of the effect reported here needs accounting for. The explanation lies in the structure of the task. Rubric

scores are ordinal levels on a fixed scale, and the dominant failure of a prompted judge relative to such a scale is a monotone displacement: the model orders essays sensibly but places the whole ordering too low or too high, and compresses it near the ends of the range. Our data show exactly this signature, with large negative bias coexisting with essentially intact Spearman correlation.

A monotone calibrator is the matched correction for a monotone displacement. It can undo an arbitrary order-preserving deformation of the scale, which a linear rescaling cannot, while being unable to disturb the ranking the model got right. This is why every monotone family in our comparison performs similarly and the non-monotone one does not, and it is why the effect is much larger than typical calibration gains on tasks where the model's errors are not primarily a scale distortion.

Anchoring addresses a different part of the same problem. Adjacent rubric levels are separated by fine and partly subjective distinctions, and an unanchored prompt gives the model no fixed reference for where those boundaries lie. Supplying retrieved same-scale exemplars converts free scoring into bounded comparative judgment, which reduces variance across repeated calls and gives the calibrator a more stable input to map.

The mechanism carries directly into how such a layer should be configured. For a team scoring essays with a prompted LLM, a calibration split is worth more than further prompt engineering. A few hundred scored essays are enough to remove a displacement that would otherwise misclassify a large fraction of borderline cases, and the transfer results suggest that even a calibration split from a different corpus is much better than none.

The sampling budget should then be chosen from the intended output rather than from a target accuracy. A point-score deployment can drop to one or three draws and lose little agreement, which on a six-trait rubric is the difference between roughly 43 and roughly 6 calls per essay; a deployment that routes uncertain essays to human readers needs the repeated draws, since that is where the routing signal comes from.

How the layer is described to stakeholders matters as much as how it is configured. It should be presented as scale correction and confidence alignment, not as improved scoring accuracy, since its effect on ranking is negligible by construction. What it changes is whether the number that reaches a student means what the rubric says it means, and whether the system knows which of its own scores to doubt.

Several features of the design bound how far that guidance carries. The evaluation conditions are constructed perturbations. The main condition displaces the reference labels, and Section 4.7 shows that the same correction works when the displacement is instead applied to the model predictions, which is the more faithful analogue of LLM length preference. Both are nonetheless distortions of a known parametric form, and a calibrator facing such a distortion has an easier task than one facing an unanticipated shift.

Two experiments that would sharpen the argument were not run, and we name them rather than let the present evidence appear to cover them. The first is a fully unperturbed baseline in which the anchor bank, the calibrator and the test evaluation are all built from the original labels. Because the cached model responses were collected against the perturbed splits, this baseline cannot be recomputed from the

cache and requires a fresh scoring run; every result reported here should therefore be read as conditional on a perturbed corpus. The second is a no-anchor baseline in which the model receives only the rubric and the legal score range, which would separate the contribution of retrieved anchors from that of the prompt itself. Section 4.8 takes a step towards realistic shift by holding out whole writing prompts, but only the calibrator is topic-disjoint there, since the anchor bank was fixed before caching. Naturally occurring deployment shifts, including new rubrics, changes in the student population, adversarial essays and silent model version updates, remain untested.

The conformal component is cross-conformal rather than strictly split-conformal, because the residual quantile is estimated on held-out folds of the same split that supplies the calibrator and the temperature. Its coverage figures are accordingly empirical, and the reweighted test cohort breaks exchangeability in any case. A design with a fully independent conformal-calibration split would restore the standard guarantee at the cost of a smaller calibration set.

The three judge models are commercial APIs. This matches the deployment scenario we target, since institutions without training data are also typically institutions using hosted models, but it limits reproducibility, because the endpoints can change without notice and our cached predictions are a record of a particular moment. Replicating the layer on open-weight models would strengthen the evidence.

The subgroup analysis is bounded by the metadata the corpora provide and by the minimum group size, which excludes the smallest groups precisely where disparities may be largest. It also treats attributes one at a time, does not examine intersections, and reports uncorrected descriptive counts rather than inferential tests. Finally, the review-triage results are a simulation in which referral is assumed to resolve an essay correctly; a study with human readers would be needed to establish operational value.

6 Conclusion

This paper studied post-hoc calibration of prompted LLM essay scoring across three corpora, three commercial models, five calibrator families, five sampling budgets and twenty-seven subgroup comparisons, under a single cached-prediction protocol and a controlled distribution shift. Calibration improved ordinal agreement in every setting, recentred a large systematic bias to near zero, improved the quality of the predictive distributions, and did so with statistical support in all 45 bootstrap comparisons. The benefit derives from the monotone structure of the correction rather than from a particular estimator, survives when the displacement is injected into the predictions instead of the labels, transfers across corpora at about three quarters strength, and holds for unseen writing prompts within a corpus. Measured disparities between demographic subgroups narrow consistently in direction, though individual reductions are mostly not resolved by essay-level bootstrap intervals at these subgroup sizes.

The effect on ranking is negligible, which is a property of the design rather than a shortcoming of it. A prompted scorer that already ranks essays acceptably does not need a better ranker; it needs its scale corrected, its confidence made meaningful, and its errors distributed evenly. Those are exactly the properties a post-hoc layer can

supply without retraining. Future work should test the layer on open-weight models, under naturally occurring rather than constructed shift, with bias injected into predictions rather than labels, and with human readers in the loop.

Data availability. The three corpora analysed in this study are publicly available from their original publications [30–32]. The code implementing the calibration mechanism, together with the evaluation scripts used to produce the results reported here, is available at https://github.com/EV4-CX/homework_evlaution_code.

Code availability. All analysis code is available at the repository given above. Every result in this article is recomputed from a single cached set of model responses, so the analyses can be reproduced without further API calls.

Declarations

Funding. The authors reported there is no funding associated with the work featured in this article.

Competing interests. The authors declare no competing interests.

Ethics approval. Not applicable. The study is a secondary analysis of three publicly released, de-identified essay corpora and involved no new data collection from human participants.

References

[1] Attali, Y. & Burstein, J. Automated essay scoring with e-rater V.2. *Journal of Technology, Learning, and Assessment* **4** (2006). URL https://ejournals.bc.edu/index.php/jtla/article/view/1650.

[2] Taghipour, K. & Ng, H. T. A neural approach to automated essay scoring. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* 1882–1891 (2016).

[3] Dong, F., Zhang, Y. & Yang, J. Attention-based recurrent convolutional neural network for automatic essay scoring. *Proceedings of the 21st Conference on Computational Natural Language Learning* (2017).

[4] Ke, Z. & Ng, V. Automated essay scoring: A survey of the state of the art. *Proceedings of the 28th International Joint Conference on Artificial Intelligence* (2019).

[5] Mizumoto, A. & Eguchi, M. Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics* **2** (2023).

[6] Dubois, Y., Galambosi, B., Liang, P. & Hashimoto, T. B. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475* (2024).

[7] Liu, Y. *et al.* G-Eval: NLG evaluation using GPT-4 with better human alignment. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* 2511–2522 (2023).

[8] Zheng, L. *et al.* Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems* **36**, 46595–46623 (2023).

[9] Williamson, D. M., Xi, X. & Breyer, F. J. A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice* **31**, 2–13 (2012).

[10] Platt, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers* (1999).

[11] Zadrozny, B. & Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 694–699 (2002).

[12] Niculescu-Mizil, A. & Caruana, R. Predicting good probabilities with supervised learning. *Proceedings of the 22nd International Conference on Machine Learning* 625–632 (2005).

[13] Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning* 1321–1330 (2017). URL https://proceedings.mlr.press/v70/guo17a.html.

[14] Kull, M., Silva Filho, T. M. & Flach, P. Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (2017). URL https://proceedings.mlr.press/v54/kull17a.html.

[15] Vovk, V., Gammerman, A. & Shafer, G. *Algorithmic Learning in a Random World* (Springer, 2005).

[16] Romano, Y., Patterson, E. & Candès, E. Conformalized quantile regression. *Advances in Neural Information Processing Systems* **32** (2019). URL https://papers.nips.cc/paper_files/paper/2019/hash/5103c3584b063c431bd1268e9b5e76fb-Abstract.html.

[17] Angelopoulos, A. N. & Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511* (2021).

[18] Desai, S. & Durrett, G. Calibration of pre-trained transformers. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (2020).

[19] Kadavath, S. *et al.* Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221* (2022).

[20] Tian, K. *et al.* Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (2023).

[21] Kuhn, L., Gal, Y. & Farquhar, S. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *International Conference on Learning Representations* (2023). URL https://openreview.net/forum?id=VD-AYtP0dve.

[22] Xiong, M. *et al.* Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. *International Conference on Learning Representations* (2024). URL https://openreview.net/forum?id=gjeQKFxFpZ.

[23] Dey, P., Merugu, S. & Kaveri, S. R. Conformal prediction sets for ordinal classification. *Advances in Neural Information Processing Systems* **36**, 879–899 (2023).

[24] Sheng, H., Liu, X., He, H., Zhao, J. & Kang, J. Analyzing uncertainty of LLM-as-a-Judge: Interval evaluations with conformal prediction. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* 11297–11339 (2025).

[25] Yannakoudakis, H., Briscoe, T. & Medlock, B. A new dataset and method for automatically grading ESOL texts. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (2011). URL https://aclanthology.org/P11-1019/.

[26] Dong, Q. *et al.* A survey on in-context learning. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* 1107–1128 (2024).

[27] Zebaze, A. R., Sagot, B. & Bawden, R. In-context example selection via similarity search improves low-resource machine translation. *Findings of the Association for Computational Linguistics: NAACL 2025* 1222–1252 (2025).

[28] Gabriel, J. S. & Kimura, M. Enhancing automated essay scoring model's interpretability using triplet loss with teacher-defined rubric as anchors. *Joint International Conference on AI, Big Data and Blockchain* 124–135 (2025).

[29] Naeini, M. P., Cooper, G. F. & Hauskrecht, M. Obtaining well calibrated probabilities using Bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence* (2015).

[30] Crossley, S. *et al.* The English Language Learner Insight, Proficiency and Skills Evaluation (ELLIPSE) corpus. *International Journal of Learner Corpus Research* **9**, 248–269 (2023).

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

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

[33] Cohen, J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin* **70**, 213–220 (1968).

[34] Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**, 1–3 (1950).

[35] He, P., Liu, X., Gao, J. & Chen, W. DeBERTa: Decoding-enhanced BERT with disentangled attention. *International Conference on Learning Representations* (2021). URL https://openreview.net/forum?id=XPZIaotutsD.