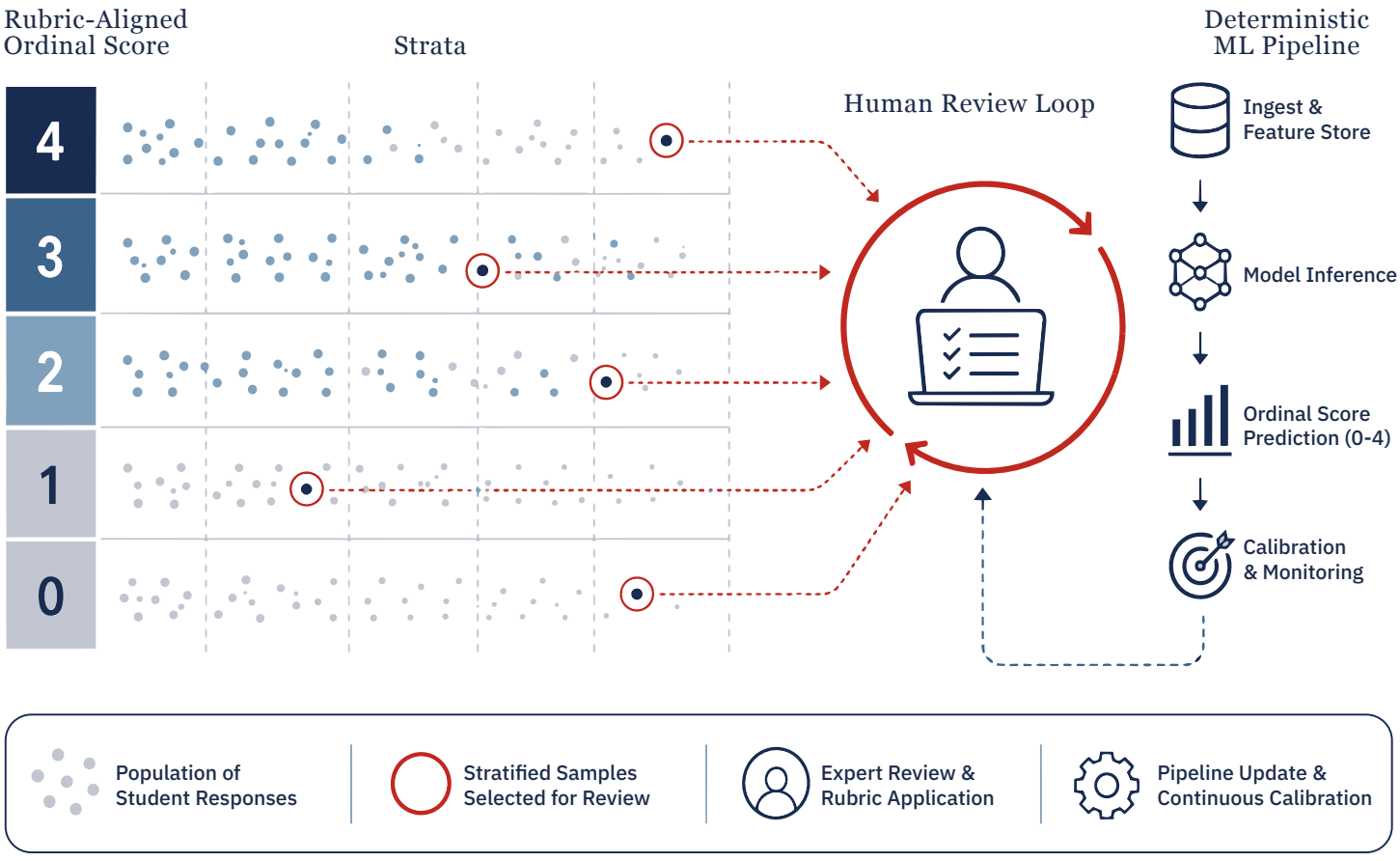
Impacter® Pathway
EDUCATION CORP.

PREPARED FOR
EXTERNAL RESEARCH
COLLABORATORS

A Stratified, Human-in-the-Loop Sampling Layer for Continuous Calibration of a Production Rubric-Aligned Ordinal Scoring Pipeline

TECHNICAL NOTE — IMPACTER Pathway Assessment, Model Suite v6.0



Abstract

We describe a production-grade, automated stratified sampling layer that sits between high-throughput machine learning (ML) inference and the human review/retraining loop in a rubric-aligned ordinal scoring pipeline for K–12 human-skills assessment. The pipeline scores open-ended student voice and text responses against a combined rubric framework of eight Harvard Making Caring Common-derived character competencies and CASEL-aligned social and emotional learning (SEL) competencies. Scoring is performed with a fine-tuned DeBERTa-v3-small encoder and CORAL ordinal regression head. Promoted models are registered through MLflow Model Registry and frozen for deterministic production inference. The new sampling layer replaces ad hoc spot-checking with weekly, balanced draws across attribute, predicted score band, course/prompt, and item type, and routes selected responses into the reviewer-facing Manage Scoring interface. Re-scored items propagate automatically into validation sets and the next retraining corpus. The design operationalizes principles from active learning, importance-weighted evaluation, and concept-drift monitoring, and is purpose-built to remain robust as new prompts, populations, and partner content come online.

1. Problem and Motivation

Production natural language processing (NLP) scoring systems for educational measurement face a tension that is well documented in both the educational data mining and MLOps literatures: the empirical distribution of student response data in non-stationary, while the human-rated training corpus on which the scoring model was calibrated is fixed at the moment of model promotion (Bauer & Zapata-Rivera, 2020; McCaffrey et al., 2021). New cohorts, new prompts, new linguistic populations (e.g., English-learner subgroups; cross-state Skills for the Future deployments), and shifts in modality (typed vs. dialect-aware voice) all introduce covariate shifts relative to the training distribution. Without a structured process for sampling and re-rating responses over time, two problems can build: calibration drift that appears first in specific parts of the rubric, and weaker training data quality as new content is added without enough human-rated coverage.

The standard remedy, periodic random spot-checks performed by the scoring team, is operationally fragile. It tends to oversample the modal class while leaving the most ambiguous parts of the rubric underexamined. In this setting, the outer scores, 0 and 4, are often the most stable because they represent clearer low- and high-evidence responses. Most scoring ambiguity occurs in the middle of the rubric, especially among scores 1, 2, and 3, where neighboring levels can be harder to distinguish. Random spot-checking also leaves subgroup, prompt-type, and length-band coverage to chance. The result is a validation set that may look adequate in aggregate but provides too little information about the middle-score distinctions where calibration errors are most likely.

To address both problems, this note describes an automated sampling layer that connects daily inference with human review and monthly retraining, ensuring that production responses are reviewed in a structured way before they are used for validation or future model updates.

2. System Context

The scoring pipeline is discriminative and deterministic at inference. A student response — written or transcribed via dialect-aware ASR — is encoded by DeBERTa-v3-small (He et al., 2021; ~44M parameters, fine-tuned per competency) and passed through a CORAL ordinal regression head (Cao et al., 2019) that produces a 0–4 rubric-aligned score with rank structure preserved. Models are trained on double-scored human evaluations (Cohen's $\kappa \geq 0.65$ calibration threshold) following Harvard Making Caring Common and CASEL protocols, registered in MLflow, and promoted to Production only when held-out QWK improves. Production models are frozen for deterministic tokenization and scoring; identical inputs produce identical outputs across administrations. Current Model Suite v6.0 averages QWK = 0.918, exact-match accuracy = 77.3%, and adjacent (±1) accuracy = 98.0% across the eight character competencies (Curiosity, Perspective-Taking, Purpose, Self-Control, Grit, Growth Mindset, Gratitude, Compassion). Because inference is deterministic, identical responses receive identical scores under the same production model. When a disagreement appears later, it can usually be traced to a model-version change or a human re-rating. This makes the sampling layer easier to interpret because disagreements can be examined as evidence of calibration drift, rubric ambiguity, or borderline cases rather than random scoring variation. Generative LLM scoring approaches are harder to monitor in this way because temperature settings, prompt wording, and sampling behavior can introduce additional variance that complicates drift detection (Liu et al., 2025; Ober et al., in press; Peeperkorn et al., 2024).

3. The Stratified Sampling Layer

3.1 Stratification design

The sampling layer draws weekly batches from the rolling production inference log according to a four-factor stratification scheme:

- **Competency (attribute):** the thirteen rubric-aligned constructs are sampled independently so that no construct goes more than one cycle without coverage. This is critical because, as Liu et al. (2025) and Ober et al. (in press) document, agreement metrics behave very differently across constructs of varying operational clarity; aggregating across competencies masks construct-specific drift.

- **Predicted score band (0–4):** Strata are defined over the model's ordinal output. Because the clearest low- and high-evidence responses often fall at scores 0 and 4, the sampling layer places greater emphasis on the middle score bands, especially 1, 2, and 3, where neighboring rubric levels are harder to distinguish. These middle-band responses are intentionally represented at higher rates than they would be under purely random sampling, so the review set does not simply confirm the model's strongest regions. This helps prevent agreement statistics from being inflated by easy extreme cases and ensures that the rubric levels most vulnerable to calibration error receive enough human review.

- **Course / prompt family:** samples are stratified across the active prompt corpus (Portrait of a Graduate, behavioral health screener items, workforce/CTE prompts, ETS-collaborative item families). This protects against prompt-level overfitting and ensures coverage of newly authored content within one sampling cycle of release.

- **Item / response type:** voice-elicited vs. text-elicited responses, and short- vs. long-form, are sampled separately. Automatic Speech Recognition (ASR) transcription introduces a known source of input-level noise that we want explicitly represented in the validation distribution rather than averaged away.

Within each stratum cell, draws are randomized but capped, producing a balanced sample whose size per cycle is set to the throughput of the human review team. When sampling rates are stored with each selected response, downstream analyses can reweight the reviewed sample back toward the production distribution, allowing balanced review for operational efficiency while still supporting population-level metric estimates.

3.2 Reviewer workflow

Selected responses are surfaced inside the Manage Scoring interface in the IMPACTER admin platform. Reviewers can filter directly into the active sample, view the response alongside the model's predicted score and rubric anchor language, and either confirm or override with their own rubric-aligned rating. The interface supports trait-, district-, school-, and timeframe-level filtering, which means review effort can be allocated to the cells where it has the highest marginal value (e.g., a newly onboarded district with limited representation in the training corpus, or a competency whose recent QWK trace shows non-trivial slope).

Re-scored items are written back to a labeled-data store with full provenance: original prompt, transcription (where applicable), model version that produced the original score, predicted score, reviewer-assigned score, reviewer ID, and timestamp. This provenance is what allows the layer to function simultaneously as a drift monitor and as a training-data generator.

3.3 Feedback into retraining and validation

Re-rated samples flow into two downstream workflows automatically:

- **Calibration / drift validation.** On a rolling window the layer recomputes QWK, exact-match accuracy, and adjacent accuracy against the most recent human re-ratings, partitioned by stratum. Movement beyond pre-registered thresholds (degradation in QWK > 0.10 from the production-promoted value) flags the relevant competency for either targeted retraining or rubric re-examination.

- **Retraining corpus augmentation.** Confirmed and corrected ratings are versioned into the training partition for the next scheduled fine-tune. Because draws are stratified rather than convenience-sampled, each new training cycle inherits a corpus whose marginal informativeness — in the active-learning sense (Settles, 2009) — is substantially higher per labeled instance than under random or ad hoc selection. New model versions are still gated through the existing MLflow promotion criteria: a candidate is only promoted if held-out QWK improves, regardless of how the new training data was sourced.

4. Why This Works (and Why It's Efficient)

The design draws on three converging lines of methodological work, applied to an operational pipeline rather than a one-off study.

Active learning under a fixed labeling budget. In any production scoring system, the rate-limiting resource is qualified human rater time, not compute. Random sampling is inefficient when the goal is to refine a model that is already strong as most random draws confirm what the model already does well. Stratified sampling concentrates labeling effort on regions of the input distribution where reviewer disagreement carries the most information such as low-frequency strata and newly introduced prompts. This is the same logic that motivates uncertainty- and disagreement-based active learning, instantiated here as a deployable, auditable layer rather than a query function inside a training loop.

Importance-weighted, prevalence-aware evaluation. Rubric-based agreement statistics are sensitive to score prevalence. When one rubric level dominates the response distribution, percent agreement and kappa can overstate reliability while masking weaker performance on lower-frequency or more consequential score bands. The sampling layer addresses this by oversampling tail and boundary bands, then applying sampling weights when reporting aggregate metrics. This allows the review process to focus on operationally important cases, such as borderline tier assignments or top-end recognitions, while preserving unbiased headline estimates.

Continuous concept-drift monitoring with attributable signal. Because inference is deterministic and the layer fixes a stratified denominator, any movement in agreement statistics across cycles is attributable rather than ambient. We can decompose drift into (i) input-distribution shift (new prompts, new populations, ASR changes), (ii) construct-level rubric ambiguity, and (iii) model-version effects after a promotion. This is the property that lets the layer support both our own model lifecycle and the more demanding monitoring requirements of partner programs, where evidence of ongoing fairness and stability is part of the contract.

The efficiency gains are concrete. Compared to the prior ad hoc spot-check process, the new layer (i) bounds the time-to-detect for construct-level drift to one sampling cycle; (ii) raises the average information yield per re-rated response, both for validation and for retraining; (iii) eliminates the silent under-sampling of low-prevalence and boundary cells that previously biased our internal QWK estimates upward; and (iv) generates a continuously growing, provenance-tagged human-rated corpus that compounds in value as it accumulates.

5. Connections to Recent Research

The design is consistent with, and partly motivated by, recent results on the limits of generative-LLM coding for educationally meaningful constructs. Ober et al. (in press) and Liu et al. (2025) report that human-LLM agreement is strongly modulated by the operational clarity of the construct: well-specified constructs (e.g., Self-efficacy) admit reliable LLM coding under appropriate temperature settings, while constructs with lower clarity (Persistence, Prior KSAs) yield high LLM-LLM internal consistency but poor alignment with human raters. The implication for production scoring is that construct definition does the heavy lifting, and that no amount of prompt engineering on a generative model substitutes for (a) a rubric that specifies observable developmental levels, (b) a discriminative architecture that learns ordinal structure from human-anchored labels, and (c) a continuous re-rating mechanism that keeps the human anchor live.

The sampling layer described here is the third leg of that tripod. It is the mechanism by which human judgment remains in the loop at production scale, without reverting to either purely generative coding or to one-time calibration.

6. Current Status and Forward Work

The system is live as of the current operating cycle. The first stratified sample over the prior month of production responses has been drawn and distributed to the scoring team; re-rated outputs will populate the next validation pass and the next retraining corpus on the standard cadence. Near-term development priorities include: (i) per-stratum confidence-band reporting on agreement statistics; (ii) integration of subgroup invariance tests (multinomial logistic regression on score residuals against demographic indicators) directly into the cycle-level drift report; (iii) prompt-level item analysis hooks for newly authored ETS-collaborative content so that new items can be flagged for additional-sampling weight during their first deployment cycle; and (iv) an auditable export of cycle-level metrics suitable for inclusion in partner-facing model cards.

We expect this layer to be especially load-bearing as the pipeline scales into new deployments, each of which introduces new prompts, new populations, and new evidentiary expectations. The sampling layer is what allows a single ML pipeline to serve all of those programs while remaining defensibly calibrated against the human-anchored rubric for each.

References

Bauer, M. I., & Zapata-Rivera, D. (2020). Cognitive foundations of automated scoring. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), Handbook of Automated Scoring: Theory into Practice (pp. 13–28). CRC Press.

Cao, W., Mirjalili, V., & Raschka, S. (2019). Rank-consistent ordinal regression for neural networks. Pattern Recognition Letters.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. Journal of Clinical Epidemiology, 43(6), 543–549.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical and Statistical Psychology, 61(2), 297–308.

He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with disentangled attention. ICLR.

Liu, X., Zambrano, A. F., Baker, R. S., et al. (2025). Qualitative coding with GPT-4: Where it works better. Journal of Learning Analytics, 12(1), 169–185.

McCaffrey, D. F., Casabianca, J. M., Ricker-Pedley, K. L., Lawless, R. R., & Wendler, C. (2021). Best practices for constructed-response scoring. ETS Research Report Series.

Ober, T. M., Zhang, S., Zapata-Rivera, D., Schroeder, N. L., & Botelho, A. F. (in press). Using LLMs to identify indicators of persistence from students' dialogues with a pedagogical agent. Journal of Educational Data Mining.

Peeperkorn, M., Kouwenhoven, T., Brown, D., & Jordanous, A. (2024). Is temperature the creativity parameter of large language models? arXiv:2405.00492.

Settles, B. (2009). Active learning literature survey. University of Wisconsin–Madison.