

Beyond Clean Agreement: A Stability-Bounded Evaluation of Compact AES Distillation

Yi Gui

Measurement Incorporated
ygui@measinc.com

Abstract

Compact distillation is often motivated as a resource-aware strategy for automated essay scoring (AES), but target-design choices are typically compared using clean held-out agreement alone. This study evaluates whether such comparisons remain informative once broader sources of variability are considered. Three compact student models were trained against compressed teacher-derived targets (PCA, autoencoder, and multi-target conditions) using the PERSUADE 2.0 corpus. Under clean evaluation, all models achieved strong scoring agreement, but the best target condition differed by student family and the margins were small. Perturbation, distribution-shift, and subgroup-bootstrap analyses then showed that these clean differences were bounded by larger stability constraints: perturbation effects dominated target-design differences, distribution-shift variability was large, and most subgroup-gap estimates were sampling-sensitive. Compact distillation therefore preserves clean scoring performance but does not resolve robustness fragility or subgroup uncertainty. Uncertainty-aware evaluation may be necessary when comparing resource-constrained educational prediction systems, because comparisons based solely on clean held-out performance can be misleading.

Keywords: automated essay scoring, knowledge distillation, representation compression, robustness, subgroup fairness, educational data mining

1 Introduction

Automated essay scoring (AES) is increasingly studied as a representation problem rather than only a prediction problem. Recent transformer-based work shows that essay embeddings retain substantial score-relevant information even after compression, shifting attention from whether large models work to what structure they preserve and whether smaller models can inherit it (Li & Ng, 2024; Sasaki & Masada, 2023; Sun et al., 2025). This matters beyond engineering efficiency: in educational measurement, a score is useful only when its interpretation remains

defensible across the conditions under which it is used.

A central gap in the current literature is that target design is often treated as an implementation detail rather than as a substantive modeling decision. Choices such as aligning a student to a PCA-derived target, an autoencoder-derived target, or a joint multi-target objective are usually framed as alternative recipes for improving clean performance. That framing is incomplete for AES, because different compressed targets may preserve different aspects of the teacher representation, and those differences may affect not only average agreement but also robustness, subgroup behavior, and sensitivity to distribution shift.

Current AES evaluation often treats clean held-out performance as sufficient for comparing model designs. Clean target-design differences should instead be interpreted alongside seed variability, perturbation sensitivity, and subgroup uncertainty, because any of these can rival or exceed the observed differences. This study therefore evaluates compact AES distillation across clean performance, target variability, robustness, out-of-distribution (OOD) sensitivity, and subgroup uncertainty.

The contribution is twofold. First, the study examines compact distillation as a resource-aware modeling strategy for educational scoring, asking how alternative compressed target designs affect performance, robustness, and stability. Second, it introduces a joint evaluation framework that combines clean performance, perturbation robustness, distribution-shift sensitivity, and subgroup uncertainty, illustrating how conclusions drawn from conventional held-out evaluation can change when broader sources of variability are considered. Although demonstrated in an AES distillation setting, the framework is intended as a general approach for evaluating resource-constrained educational prediction systems.

The analysis proceeds sequentially: it first asks whether clean target-design differences are consistent across student families, then tests

whether those differences remain interpretable when the learned representations are stressed.

2 Related Work

Recent AES research has become increasingly representation-centered as transformer encoders have replaced earlier feature-engineered pipelines with dense contextualized essay embeddings (Li & Ng, 2024; Sun et al., 2025). An important line of work asks whether these representations are substantially redundant and therefore compressible without eliminating score-relevant structure. This motivates the present focus on target-space design: if compact representations preserve score information, the way they are constructed becomes part of the scoring system rather than a neutral preprocessing choice.

A second literature concerns compact models and distillation. In general NLP, student models retain substantial task-relevant knowledge from larger teachers, and later work extended that logic from output imitation to the transfer of intermediate or low-dimensional representational structure (Brown et al., 2023; Jiao et al., 2020; Sun et al., 2019; Zhao et al., 2022). For AES, the practical question is not only whether a smaller model can approximate clean score agreement, but whether it preserves the behavioral stability needed for score interpretation. Recent educational data mining (EDM) work has addressed related efficiency–performance trade-offs: Whitehill and LoCasale-Crouch (2024) showed that simpler models can match LLM-level accuracy for classroom scoring at lower cost, and Tran et al. (2024) evaluated LLMs for educational assessment across multiple quality dimensions rather than a single accuracy metric.

A third literature emphasizes validity, fairness, and robustness in AES and educational measurement. That work makes clear that agreement on a held-out split is not sufficient for a defensible score-use argument, and that subgroup behavior, distribution shift, and sensitivity to degraded inputs all matter (Ercikan & McCaffrey, 2022; Johnson et al., 2022; Johnson & McCaffrey, 2023; McCaffrey et al., 2022; Rupp, 2018). Within EDM, Cohausz et al. (2024) showed that demographic features interact with model predictions and fairness metrics in non-trivial ways, and Verger et al. (2024) proposed systematic approaches to evaluating and mitigating algorithmic unfairness. The present study links these concerns by asking

whether clean target-design differences remain interpretable once robustness, OOD sensitivity, and subgroup uncertainty are evaluated together.

3 Methods

The empirical design is organized as one sequential evaluation. The first part examines clean target-design performance across the full completed condition matrix. The second part uses the artifact-complete core matrix to evaluate whether the same target-design differences remain interpretable under perturbation, OOD evaluation, and subgroup bootstrap uncertainty. Figure 1 summarizes the integrated workflow: clean target-design evaluation is the starting point, and robustness, OOD, and subgroup-stability analyses are used to test the boundary conditions of those clean comparisons.

3.1 Data and Task

The study uses essay response data from the PERSUADE 2.0 corpus (Crossley et al., 2024), a large-scale publicly available dataset of student essays written by U.S. students in grades 9 through 12. Essays were written in response to seven writing prompts spanning argumentative and informational genres (e.g., car-free cities, distance learning, electoral college, driverless cars, exploring Venus); five were text-dependent tasks requiring students to draw on provided source material, and two were independent writing tasks. Each essay was assigned a holistic quality score on a 1–6 integer scale by trained human raters.

The dataset includes demographic metadata: gender (50.2% female, 49.8% male), English language learner (ELL) status (12.7% ELL), race/ethnicity (42.5% White, 27.4% Hispanic/Latino, 18.6% Black/African American, 6.9% Asian/Pacific Islander, 4.1% two or more races, 0.5% American Indian/Alaskan Native), and economic disadvantage status (available for about 71% of the sample and roughly balanced). These variables are used in subgroup analyses but are not model features.

The dataset was split into train ($N = 12,428$) and held-out test ($N = 1,387$) partitions using stratified sampling by prompt, task type, and grade level. Within the training portion, a stratified inner split created train_inner ($N = 10,563$) and validation ($N = 1,865$) sets using split seed 2026. The modeled outcome was the holistic essay score, and three text fields were

carried across the teacher-target and student-distillation pipeline: the full essay text, source reading material, and the writing assignment prompt. The same held-out test set provides the clean comparison base for all analyses reported here. The corpus is publicly available, and analyses used de-identified records with no individually identifiable student information.

3.2 Teacher Representation and Target Construction

Teacher representations were generated from a locally hosted DeBERTa-v3-Large model (He et al., 2023; hidden size 1,024). For each essay, the three text fields were encoded separately using a sliding-window strategy with maximum length 512 tokens and stride 256. Within each window, token-level hidden states were mean-pooled over the attention mask; when an input produced multiple windows, the window-level representations were averaged. The three resulting 1,024-dimensional field vectors were concatenated to form a 3,072-dimensional teacher representation per essay. PCA was fit on the training split only and applied to both training and test essays to produce the PCA:128 target.

Autoencoder (AE) targets were generated by training a symmetric bottleneck autoencoder on the same 3,072-dimensional teacher representations. The encoder path was $3,072 \rightarrow 1,024 \rightarrow 512 \rightarrow 128$ and the decoder was the reverse. Training used mean squared error reconstruction loss with learning rate 0.001, weight decay $1e-5$, batch size 128, and up to 80 epochs with early stopping (patience 10, minimum 5 epochs). Because autoencoder training is stochastic, three AE targets were produced using seeds 42, 1,234, and 2,025.

3.3 Student Models and Target Conditions

Three compact student model architectures were considered: MiniLM-L12-H384-uncased, DeBERTa-v3-xsmall, and electra-small-discriminator. These represent three lightweight transformer families with different pretraining strategies (distilled encoder, disentangled attention, and replaced-token detection). Students used the same three-field sliding-window encoding as the teacher. The three field representations were concatenated and passed through a shared representation head (linear projection $\rightarrow$ layer normalization $\rightarrow$ GELU $\rightarrow$ dropout), after which target-specific linear heads

predicted the compressed teacher spaces. Training optimized mean squared error between the student's target-head outputs and the corresponding compressed teacher representations.

The clean target-design matrix comprised six conditions per student: Single PCA, AE:42, AE:1234, AE:2025, Multi 1:1 (joint alignment to PCA:128 and AE:128:1234 with equal loss weights), and Multi 1:2 screened (joint alignment with raw weights 1 and 2 plus an inner-screening rule). Because the Stage-2 stress analyses required inference-capable saved artifacts, the robustness, OOD, and subgroup-bootstrap analyses used the artifact-complete core matrix: Single PCA, Single AE:2025, and Multi 1:1 across the three student families (Table 1). This reflects the conditions for which full model checkpoints were preserved, not a post hoc selection of favorable conditions: Single PCA was the top clean condition for two of the three student families, while AE:2025 and Multi 1:1 represent the best-performing AE seed and the primary multi-target condition.

3.4 Stage-1 Clean Target-Design Evaluation

The clean target-design evaluation used 18 completed runs (three student families $\times$ six target conditions). Model selection followed the fixed split design: the inner split selected the best epoch, after which the student was refit on `train_full` and evaluated on the held-out test split. Downstream score evaluation used the same ordinal logistic regression framework (LogisticAT) as the companion compression pipeline, treating the learned student representation as the feature space, which keeps the comparison centered on target-space transfer rather than on downstream scoring architecture.

3.5 Stage-2 Robustness, OOD, and Subgroup-Stability Evaluation

The Stage-2 evaluation asks what happens when the learned representations are stressed. Robustness analyses used the saved artifact-complete reruns and the available perturbation conditions: OCR-like noise, spelling noise, 50% truncation with a minimum word floor, 128-word short crop, and long-repeat-tail perturbation. These perturbations are interpreted as controlled information-loss or surface-degradation conditions rather than as a comprehensive taxonomy of essay corruption.

OOD analyses used leave-grade-out (4 folds, one per grade level: 9, 10, 11, 12) and leave-prompt-out (7 folds, one per writing prompt) designs. In each fold, the held-out group was excluded from the training data, a new downstream ordinal logistic regression scorer was fit on the remaining training essays using the student's learned representations, and the held-out group served as the test set. All seven prompts and all four grade levels produced folds with sufficient sample sizes (minimum training $n = 100$, minimum test $n = 30$); no folds were skipped.

Subgroup-stability analyses used saved predictions and held-out metadata to estimate subgroup gaps with essay-level bootstrap intervals, asking whether observed gaps are stable enough to support model-level interpretation.

3.6 Metrics and Uncertainty Analysis

The primary score-agreement metric was quadratic weighted kappa (QWK). Additional summaries included mean absolute error (MAE), exact agreement, adjacent agreement, standardized mean difference (SMD), and signed-error shift, the last used to identify directional scoring movement, especially under perturbation.

For planned multi-target contrasts, paired bootstrap confidence intervals were computed on QWK differences by resampling the 1,387 held-out essays with replacement, each summary based on 20,000 resamples. For subgroup analyses, the Stage-2 bootstrap summaries used 1,000 essay-level resamples per run and subgroup variable. These intervals are interpreted descriptively, because the goal is to evaluate stability rather than to assign model-level ordering from single summaries. Bootstrap intervals were flagged using pre-specified CI-width thresholds. For QWK-gap intervals, width (upper – lower) $\geq .20$ was flagged as very wide and width $\geq .10$ as wide; intervals below .10 were labeled relatively stable. For other metrics, the flag was based on the ratio of CI width to the absolute value of the observed gap: a ratio ≥ 2.0 was flagged as very wide, ≥ 1.0 as wide, and below 1.0 as relatively stable.

3.7 Artifact and Scope Constraints

The analysis is constrained by the artifacts that the completed experiments saved. The clean target-design comparison uses the full completed

Stage-1 condition set; the stress and stability analyses use the artifact-complete Stage-2 core matrix rather than the full six-condition design, and missing robustness, OOD, or subgroup-bootstrap evidence was not simulated or reconstructed. This artifact-aware scope matters for interpretation: the study supports a stability-oriented account of compact AES distillation under the available evidence, but does not claim to exhaust all target designs, perturbation types, or repeated-seed student outcomes. These limitations are boundaries on the conclusions rather than reasons to discard the observed stability pattern.

4 Results

Results are organized as a funnel from clean target-design evidence to stress and stability evidence. The clean analysis uses the 18 completed Stage-1 runs; the robustness, OOD, and subgroup-bootstrap analyses use the artifact-complete core matrix of nine runs (three student families $\times$ Single PCA, Single AE:2025, and Multi 1:1).

4.1 Clean Target-Design Performance

Figure 2 and Table 2 summarize clean held-out QWK across the completed target-design runs. The highest observed condition differed by student family. For electra-small-discriminator, Multi 1:1 had the highest clean QWK (.835), with AE:2025 trailing by .002. For DeBERTa-v3-xsmall, Single PCA had the highest clean QWK (.839), with Multi 1:1 trailing by .005. For MiniLM-L12-H384-uncased, Single PCA had the highest clean QWK (.841), with AE:2025 trailing by .004.

Target-condition ordering was therefore conditional on student family rather than uniform: clean agreement was high across families, but the observed differences were small and changed direction across families.

4.2 AE Seed Variability and Multi-Target Contrasts

AE seed variability was not negligible. For electra-small-discriminator, AE QWK ranged from .820 to .833 (range = .013). For DeBERTa-v3-xsmall, the AE range was smaller (.823 to .826; range = .002). For MiniLM-L12-H384-uncased, AE QWK ranged from .826 to .836 (range = .010). In electra-small-discriminator and MiniLM, the AE seed range exceeded all three

reference target-method gaps used in the reporting summary (Figure 3 and Table 3).

For the planned multi-target contrasts, all nine 95% confidence intervals involving Multi 1:1 crossed zero (Table 4). The point estimates were small, and the intervals included both negative and positive differences relative to the corresponding single-target alternatives.

4.3 Robustness Under Perturbation

Robustness analyses showed a shared degradation trajectory across target conditions. As shown in Figure 4 and Table 5, performance remained close to the clean reference under OCR-like noise, declined more visibly under spelling noise, dropped substantially under 50% truncation, and then collapsed under the 128-word crop. Across the nine stress-evaluation runs, mean clean QWK was .835; under the 128-word crop, mean QWK fell to .195.

The scale of the perturbation effect far exceeded that of target condition: between-condition differences were approximately .01 to .02 QWK, whereas the crop produced an average QWK loss of approximately .64. Signed-error shifts were consistently negative under degradation, with larger negative shifts under truncation and short cropping, and under the most severe crop target-condition values converged near the bottom of the observed QWK range.

4.4 OOD Sensitivity

OOD summaries showed variability across folds and settings. In leave-grade-out evaluation, mean QWK across the nine runs was .663; in leave-prompt-out evaluation, mean QWK was .612. Leave-prompt-out folds were especially variable, with fold-level QWK ranging from .123 to .804 across the available summaries. Representation differences were visible in these summaries, but fold-level variability was large relative to the descriptive differences among target conditions.

4.5 Subgroup Behavior With Bootstrap Uncertainty

The subgroup-bootstrap analysis added essay-level interval estimates to the earlier read-only subgroup summaries. Across 324 bootstrap rows, 227 were flagged as wide or very wide, so the majority of subgroup-gap estimates had wide uncertainty intervals. For QWK gaps alone, 47 of 81 rows were wide or very wide.

A scale comparison showed that representation differences were often smaller than bootstrap uncertainty. Across model-by-subgroup QWK comparisons, the median range among target conditions was .024, whereas the median bootstrap CI width was .112. In all 27 model-by-subgroup QWK comparisons, the target-condition range was smaller than even the narrowest CI width for that comparison.

ELL-related rows showed visible differences in several runs, but the intervals were not stable enough to support target-condition or student-family ordering (Table 6). Under current uncertainty, subgroup analysis provides monitoring evidence rather than definitive fairness comparisons.

4.6 Integrated Summary

Across the integrated results, clean performance was strong, but clean differences were small, conditional on student family, and sometimes comparable to AE seed variability; the planned paired-bootstrap contrasts did not resolve a clear directional advantage for the competitive multi-target conditions. Stress and stability analyses added larger sources of variation around those clean differences: perturbation effects were much larger than target-condition differences, severe cropping produced collapse, OOD fold variability was large, and subgroup gaps were often sampling-sensitive. Across the artifact-complete matrix, compact distillation retained clean scoring performance while robustness and subgroup uncertainty patterns remained present, and differences observed under clean evaluation did not appear as large separations under stressed conditions.

5 Discussion

5.1 Target Design Effects Are Conditional

The clean results support a conditional view of target design. Different target choices produced small descriptive differences, but the highest observed condition was not the same across student families. This finding is consistent with the idea that target-space choice can shape what compact students preserve, but it does not support a universal best-target claim. The AE and multi-target findings strengthen this interpretation: AE seed variability shows that a target family can contain non-trivial internal variation, and the paired-bootstrap contrasts show that small clean differences should not be

overinterpreted when uncertainty intervals cross zero.

5.2 Stability Constraints on Target-Design Interpretation

Clean held-out performance remains important: it shows that compact students can preserve score-relevant structure from compressed teacher targets. However, clean performance should be interpreted alongside broader sources of variability. The perturbation results show that robustness limitations are tied to the amount of usable essay information rather than to target-design choice. OCR-like noise produced small declines; truncation and cropping produced much larger shared degradation; and the 128-word crop produced collapse across all conditions. Once input information falls below a critical threshold, differences among target conditions become negligible.

Distillation therefore functions as a cost-efficient preservation strategy rather than a stability-improving one: the compact students retained strong clean performance, but robustness fragility and subgroup uncertainty remained largely unchanged. Target-design differences observed under clean evaluation should be interpreted as conditional evidence that may not extend to conditions where input information is limited or sampling variability is large.

5.3 Implications for Evaluation Practice

The subgroup results illustrate a broader evaluation concern. Representation differences were often smaller than bootstrap uncertainty, and many subgroup-gap estimates were sampling-sensitive, implying that observed subgroup differences may reflect estimation variability rather than stable differences in model behavior. Under current uncertainty, subgroup results identify monitoring priorities rather than supporting directional fairness rankings.

More generally, the joint evaluation framework used in this study suggests that evaluation design is part of the validity argument for educational prediction systems. Robustness evidence speaks to construct stability: whether the system continues to reflect the intended writing construct when surface conditions change. Subgroup evidence speaks to score interpretation across relevant student groups. When these sources of variability are considered alongside clean performance, conclusions about

model comparison can change. This pattern is not specific to distillation; it applies to any resource-constrained educational scoring system where clean held-out agreement is the primary basis for model selection or deployment. Such systems should prioritize uncertainty-aware model comparison—reporting bootstrap intervals, perturbation sensitivity, and subgroup stability alongside clean metrics—rather than relying on a single held-out comparison.

5.4 Limitations and Future Work

The conclusions are bounded by the available artifacts. The stress and stability analyses use the artifact-complete core matrix rather than the full clean target-design condition set. The study does not include full multi-seed student reruns, and the robustness analyses are limited to the perturbation types available in the saved outputs. The subgroup analyses should be treated as monitoring evidence rather than definitive fairness claims. Future work should use fuller artifact persistence, repeated-seed student designs, broader perturbation scenarios, and cross-dataset replication to strengthen the stability assessment.

6 Conclusion

This integrated evaluation shows that compact AES distillation can preserve clean scoring performance, but that clean target-design differences are conditional and bounded by stability constraints. Perturbation, OOD, and subgroup-bootstrap analyses show that robustness fragility and subgroup uncertainty can overwhelm the small differences observed under clean evaluation. The practical implication is not that distillation is unhelpful, but that its evidence should be interpreted as cost-efficient preservation within a bounded stability regime rather than as proof of improved robustness or reduced uncertainty.

References

- Brown, N., Williamson, A., Anderson, T., & Lawrence, L. (2023). Efficient transformer knowledge distillation: A performance review. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track* (pp. 54–65). Association for Computational Linguistics.
<https://doi.org/10.18653/v1/2023.emnlp-industry.6>
- Cohausz, L., Tschalzev, A., Bartelt, C., & Stuckenschmidt, H. (2024). Investigating

- demographic features and their connection to performance, predictions, and fairness in EDM models. *Journal of Educational Data Mining*, 16(2), 177–213.
- Crossley, S. A., Tian, Y., Baffour, P., Franklin, A., Kim, Y., Morris, W., Benner, J., & Boser, U. (2024). A large-scale corpus for assessing written argumentation: PERSUADE 2.0. *Assessing Writing*, 61, Article 100846. <https://doi.org/10.1016/j.asw.2024.100846>
- Ercikan, K., & McCaffrey, D. F. (2022). Optimizing implementation of artificial-intelligence-based automated scoring: An evidence centered design approach for designing assessments for AI-based scoring. *Journal of Educational Measurement*, 59(3), 272–287. <https://doi.org/10.1111/jedm.12332>
- He, P., Gao, J., & Chen, W. (2023). DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*. <https://openreview.net/forum?id=sE7-XhLxHA>
- Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., & Liu, Q. (2020). TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 4163–4174). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.findings-emnlp.372>
- Johnson, M. S., Liu, X., & McCaffrey, D. F. (2022). Psychometric methods to evaluate measurement and algorithmic bias in automated scoring. *Journal of Educational Measurement*, 59(3), 338–361. <https://doi.org/10.1111/jedm.12335>
- Johnson, M. S., & McCaffrey, D. F. (2023). Evaluating fairness of automated scoring in educational measurement. In V. Yaneva & M. von Davier (Eds.), *Advancing natural language processing in educational assessment* (pp. 142–164). Routledge. <https://doi.org/10.4324/9781003278658-12>
- Li, S., & Ng, V. (2024). Conundrums in cross-prompt automated essay scoring: Making sense of the state of the art. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 7661–7681). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.acl-long.414>
- McCaffrey, D. F., Casabianca, J. M., Ricker-Pedley, K. L., Lawless, R. R., & Wendler, C. (2022). Best practices for constructed-response scoring. *ETS Research Report Series*, 2022(1), 1–58. <https://doi.org/10.1002/ets2.12358>
- Rupp, A. A. (2018). Designing, evaluating, and deploying automated scoring systems with validity in mind: Methodological design decisions. *Applied Measurement in Education*, 31(3), 191–214. <https://doi.org/10.1080/08957347.2018.1464448>
- Sasaki, T., & Masada, T. (2023). Sentence-BERT distinguishes good and bad essays in cross-prompt automated essay scoring. In K. S. Candan, T. N. Dinh, M. T. Thai, & T. Washio (Eds.), *IEEE International Conference on Data Mining Workshop, ICDMW 2022* (pp. 274–281). Institute of Electrical and Electronics Engineers. <https://doi.org/10.1109/ICDMW58026.2022.00045>
- Sun, J., Song, T., Peng, W., & Song, J. (2025). A survey of automated essay scoring: Challenges, advances, and future. *Neurocomputing*, 650, Article 130916. <https://doi.org/10.1016/j.neucom.2025.130916>
- Sun, S., Cheng, Y., Gan, Z., & Liu, J. (2019). Patient knowledge distillation for BERT model compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 4323–4332). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1441>
- Tran, N., Pierce, B., Litman, D., Correnti, R., & Matsumura, L. C. (2024). Multi-dimensional performance analysis of large language models for classroom discussion assessment. *Journal of Educational Data Mining*, 16(2), 304–335.
- Verger, M., Fan, C., Lallé, S., Bouchet, F., & Luengo, V. (2024). A comprehensive study on evaluating and mitigating algorithmic unfairness with the MADD metric. *Journal of Educational Data Mining*, 16(1), 365–409.
- Whitehill, J., & LoCasale-Crouch, J. (2024). Automated evaluation of classroom instructional support with LLMs and BoWs: Connecting global predictions to specific feedback. *Journal of Educational Data Mining*, 16(1), 34–60.
- Zhao, X., Yu, Z., Wu, M., & Li, L. (2022). Compressing sentence representation for semantic retrieval via homomorphic projective distillation. In *Findings of the Association for Computational Linguistics: ACL 2022* (pp. 774–781). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.findings-acl.64>

Tables

Table 1: Student Models and Target Conditions.

Component	Included Conditions	Purpose
Student families	MiniLM-L12-H384-uncased; DeBERTa-v3-xsmall; electra-small-discriminator	Compact transformer families for efficient AES distillation
Clean target conditions	Single PCA; AE:42; AE:1234; AE:2025; Multi 1:1; Multi 1:2 screened	Stage-1 clean target-design comparison (18 runs)
Stress/stability core	Single PCA; Single AE:2025; Multi 1:1	Artifact-complete Stage-2 matrix (9 runs)

Note. The full clean matrix and the artifact-complete stress matrix are related but not identical; see Section 3.7 for scope constraints.

Table 2: Highest Clean Condition Within Each Student Family.

Student Family	Best Condition	QWK	MAE	Exact	Adjacent	SMD	Gap (QWK)
electra-small	Multi 1:1	.835	.437	.591	.972	-.007	.002
DeBERTa-v3-xsmall	Single PCA	.839	.436	.587	.978	-.002	.005
MiniLM-L12-H384	Single PCA	.841	.428	.596	.977	.002	.004

Note. Gaps to the runner-up condition range from .002 to .005 QWK.

Table 3: AE-Seed Variability Summary.

Student Family	AE Mean	AE Min	AE Max	AE Range	Best Seed	Interpretation
electra-small	.829	.820	.833	.013	AE:2025	Exceeded all three reference gaps
DeBERTa-v3-xsmall	.824	.823	.826	.002	AE:2025	Did not exceed reference gaps
MiniLM-L12-H384	.832	.826	.836	.010	AE:2025	Exceeded all three reference gaps

Note. AE Mean, Min, Max, and Range refer to held-out QWK across the three autoencoder seeds. AE seed variability can rival between-method differences, so AE should not be interpreted as a single deterministic alternative to PCA.

Table 4: Multi-Target Paired-Bootstrap Contrasts (QWK).

Student Family	Comparison	Point Diff	95% CI Low	95% CI High
electra-small	Multi 1:1 vs Multi 1:2 screened	+.005	-.005	+.014
electra-small	Multi 1:1 vs best single AE	+.002	-.009	+.013
electra-small	Multi 1:1 vs Single PCA	+.007	-.005	+.019
DeBERTa-v3-xsmall	Multi 1:1 vs Multi 1:2 screened	+.001	-.008	+.010
DeBERTa-v3-xsmall	Multi 1:1 vs best single AE	+.008	-.003	+.020
DeBERTa-v3-xsmall	Multi 1:1 vs Single PCA	-.005	-.015	+.005
MiniLM-L12-H384	Multi 1:1 vs Multi 1:2 screened	+.001	-.009	+.011
MiniLM-L12-H384	Multi 1:1 vs best single AE	.000	-.011	+.011
MiniLM-L12-H384	Multi 1:1 vs Single PCA	-.004	-.016	+.007

Note. All nine 95% confidence intervals cross zero; the clean multi-target contrasts are unresolved under the available paired-bootstrap intervals (20,000 resamples each).

Table 5: Key Perturbation Effects (Means Across Nine Stress-Evaluation Runs).

Condition	Mean QWK	Δ QWK	Mean MAE	Δ MAE	Signed-Error Shift
Clean reference	.835	—	.441	—	—
OCR-like noise	.827	−.008	.452	+.010	−.072
Spelling noise (5%)	.768	−.066	.547	+.105	−.299
50% truncation	.665	−.170	.894	+.452	−.856
128-word crop	.195	−.640	1.594	+1.153	−1.587

Note. Perturbation effects are much larger than observed target-condition differences.

Table 6: Selected Subgroup Gaps With Essay-Level Bootstrap Intervals.

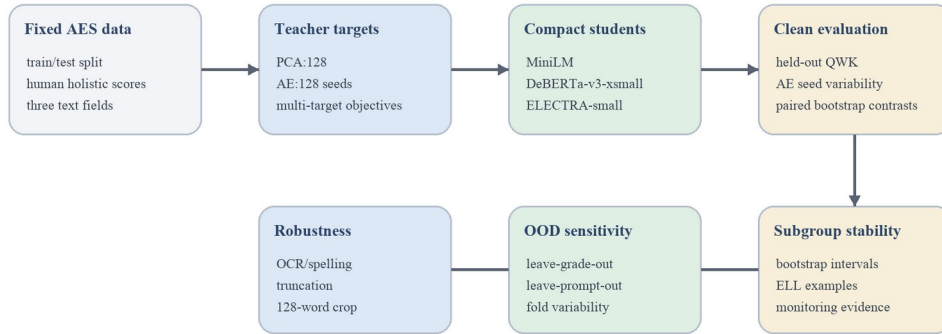
Student	Condition	Subgroup	Metric	Gap	95% Bootstrap CI	Min n	Flag
MiniLM	Single PCA	Grade level	QWK gap	.091	[.032, .300]	41	Very wide
MiniLM	Single AE:2025	Grade level	QWK gap	.116	[.054, .300]	41	Very wide
DeBERTa xsmall	Single PCA	Grade level	QWK gap	.123	[.068, .270]	41	Very wide
ELECTRA small	Single PCA	Score bin	QWK gap	.339	[.252, .428]	283	Wide
DeBERTa xsmall	Single PCA	Econ. disadv.	QWK gap	.048	[.016, .099]	401	Stable
MiniLM	Single PCA	Essay length	MAE gap	.140	[.095, .211]	461	Stable
DeBERTa xsmall	Single PCA	ELL status	QWK gap	.124	[.063, .203]	178	Wide
ELECTRA small	Multi 1:1	ELL status	QWK gap	.036	[.002, .094]	178	Stable
MiniLM	Single AE:2025	ELL status	QWK gap	.053	[.006, .109]	178	Wide

Note. Rows were selected to illustrate all three uncertainty categories (very wide, wide, and relatively stable) and to include ELL-related examples, which showed visible gaps in several runs. Subgroup analysis serves as monitoring evidence rather than a ranking. See Section 3.6 for width-flag definitions.

Figures

Integrated AES Distillation Evaluation

Clean target-design comparison followed by stress and stability evaluation



Main interpretation: target-design differences under clean evaluation are bounded by perturbation, OOD, and subgroup uncertainty.

Figure 1: Integrated workflow for clean target-design and stress/stability evaluation. The workflow frames the paper as one sequential evaluation: clean target-design comparison followed by robustness, OOD, and subgroup-stability analyses.

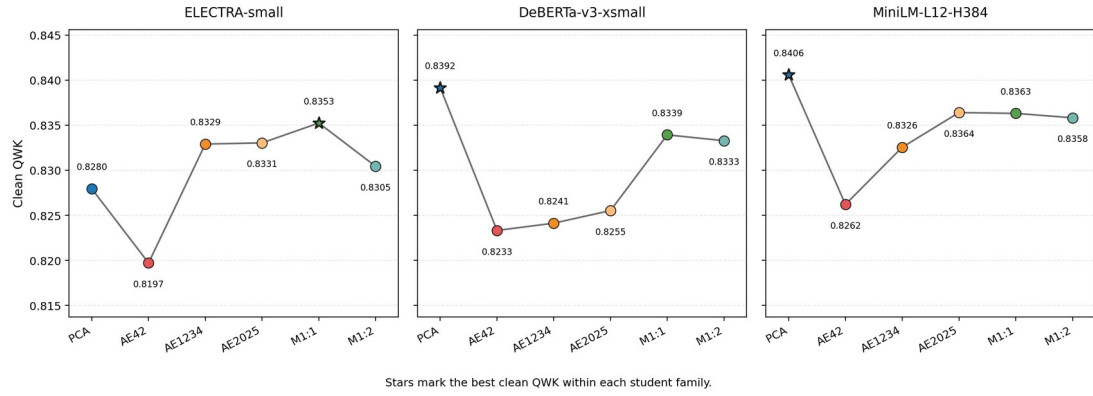


Figure 2: Clean QWK by student family and target condition. No target condition is uniformly highest across the three compact student families.

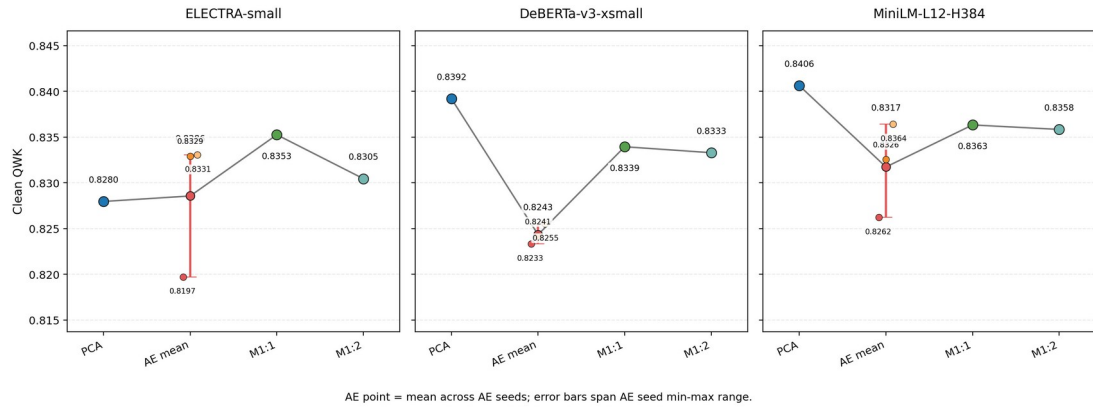


Figure 3: AE seed variability compared with target-condition differences. Within-method stochastic variation can rival the differences between target designs.

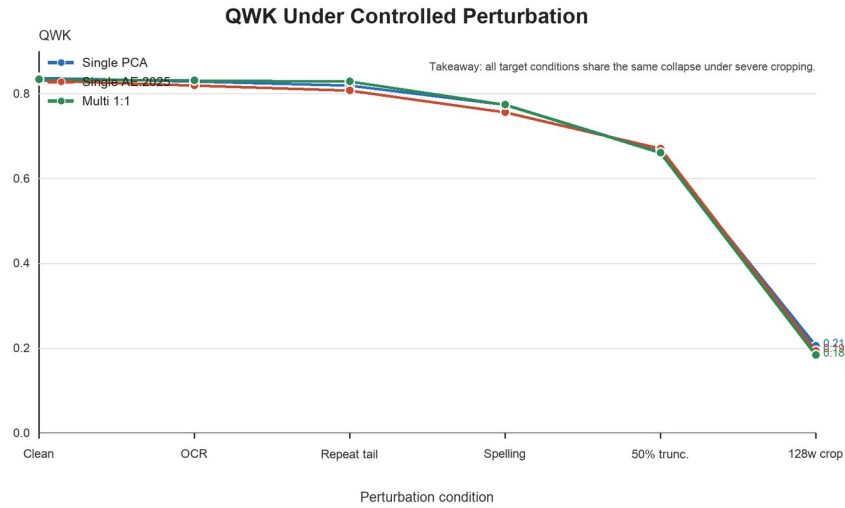


Figure 4: QWK under controlled perturbation by target condition. All target conditions follow the same degradation trajectory, with collapse under severe cropping.