

Long-Context Transformer and State-Space Architectures for Data-Efficient Automated Essay Scoring: A Comparative Study of ModernBERT, Longformer, and Mamba

Justin O. Barber and Michael P. Hemenway and Martha Bellows and Susan Lottridge

Pearson

{justin.barber,michael.hemenway,martha.bellows,sue.lottridge}@pearson.com

Abstract

Operational automated essay scoring (AES) programs must decide how much human-scored data to collect before deploying a model for a new prompt. Prior work showed that rigorous Bayesian hyperparameter optimization (HPO) enables transformer scorers to reach deployable quality with as few as 64–128 training essays. We ask whether a newer generation of long-context architectures shifts that data-efficiency frontier, and whether their relative ranking changes with training size. On ASAP-2 (7 argumentative prompts) we compare 4 architecture variants (ModernBERT, Longformer, and two Mamba state-space models, causal and bidirectional) across 4 training sizes (128, 256, 384, and 512). We hold validation ($n=300$) and test ($n=500$) fixed and draw 5 stratified replicates per cell, for 560 trained models in total. Each cell receives an independent Optuna HPO sweep followed by lightweight post-hoc adjustments (a deployment-criteria-aligned fine-tune and an affine calibration). We report agreement (QWK, exact/adjacent), distributional fidelity (KL, JSD, Earth Mover’s Distance), and intersectional fairness (gender $\times$ race/ethnicity) on the four largest prompts. No single architecture dominates. ModernBERT (0.741), bidirectional Mamba (0.740), and Longformer (0.737) are statistically indistinguishable, and causal Mamba trails (0.718). The decisive state-space finding is *directionality*. A bidirectional scan adds +0.022 QWK over the causal variant ($p = .007$) and fully closes the gap to the transformers in the data-scarce regime (0.713 vs. 0.668 at 128 training essays). Bidirectional Mamba also yields the most human-like predicted score distributions on the divergence measures (lowest KL and JSD, and essentially tied with ModernBERT on EMD) at roughly half the training cost of Longformer. We conclude that for data-efficient AES the choice of attention-vs-SSM backbone matters far less than two other factors. The first is the collection of enough labels to cross the deploy-

ment bar, where deployment attainment (our green/yellow operational performance-criteria tiers) rises from 60% at 128 to 84% at 512. The second, for SSMs, is a bidirectional scan.

1 Introduction

Assessment programs that deploy AES for a new prompt face a concrete question. *How much scored data must we collect?* The answer carries real cost, because every training essay must be scored by trained human raters under documented quality control, and that labor is the dominant expense in standing up a new automated scoring model. A program that can reach deployable quality with a few hundred training essays rather than several thousand saves both money and calendar time, and it reaches operational readiness sooner. Hemenway et al. (2026) and Barber et al. (2025) established that systematic hyperparameter optimization (HPO) sharply reduces the training sizes at which transformer scorers meet operational quality bars ($QWK \geq 0.70$ plus the green/yellow tiers of an operational performance-criteria scheme adapted from Williamson et al., 2012, defined in Section 4). That work fixed its architecture choices, DeBERTa for short responses and Longformer for essays. It left open whether newer long-context families improve data efficiency on the essay-length responses that dominate high-stakes writing assessment.

The question matters because essays are long. A holistic argumentative response routinely runs from several hundred words to roughly 800–1,000 words, particularly in higher grades, and the rubric rewards structure that spans the whole text, such as a thesis stated early and revisited in the conclusion. A scorer that cannot attend across the full response must truncate, compress, or partition the input, and each choice fundamentally alters the response representation. The backbone (the underlying model architecture) that encodes the essay therefore interacts with the data-efficiency question, because a

backbone that represents the entire essay may need fewer labeled examples to learn the rubric than one that only partially represents the essay.

Three architectures now offer distinct ways to model long inputs. ModernBERT (Warner et al., 2025) applies efficient alternating local-global attention with RoPE and Flash-Attention-2, and it extends the usable context well beyond the original BERT limit. Longformer (Beltagy et al., 2020) combines sparse local and global attention so that cost grows linearly with sequence length rather than quadratically. Mamba (Gu and Dao, 2023) is a selective state-space model (SSM) with linear-time recurrence and no quadratic attention, and it represents a different design philosophy altogether. Whether any of these better serves data-efficient essay scoring, and whether their ordering shifts with training size, bears directly on model-infrastructure selection.

We hold validation and test fixed across the size conditions so that the manipulation cleanly isolates training size, and we treat Mamba as a primary comparison arm (our strongest state-space candidate) with a causal-versus-bidirectional contrast rather than as a drop-in third backbone. A short pilot gates the bidirectional variant so that we add it only when the directional contrast is worth the extra training cost.

This study makes three contributions. First, it benchmarks transformer and state-space backbones for essay scoring under matched and explicitly small label budgets, and it reports agreement, distributional fidelity, and fairness on the same footing. Second, it isolates scan direction within the state-space family through a weight-tied bidirectional Mamba variant, and it shows that direction rather than capacity explains the gap to the transformers. Third, it argues from the combined evidence that label quantity governs deployable quality far more than the choice of backbone, and it offers a practical selection rule for programs that must commit infrastructure before they collect data.

2 Prior Work

Automated essay scoring. Automated essay scoring has moved from feature-engineered regression toward end-to-end neural models. Early systems represented an essay through hand-built features such as length, lexical diversity, and syntactic complexity, and they fit a regression onto human scores. Taghipour and Ng (2016) introduced a re-

current scorer that learned essay representations directly from text and removed much of that manual feature design. Uto et al. (2020) later showed that handcrafted features and neural encoders complement one another, because each captures evidence the other tends to miss. The survey of Ke and Ng (2019) catalogs this progression and notes a persistent tension between predictive accuracy and the validity evidence that operational programs require. We share that operational framing and adopt a green/yellow/red performance-criteria scheme—our adaptation and extension of Williamson et al. (2012), defined by adherence to explicit performance specifications (Section 4)—as our deployment standard.

Transformer encoders for scoring. Pretrained transformer encoders now anchor most strong AES systems. BERT (Devlin et al., 2019) established the bidirectional encoder that later scorers fine-tune, and it showed that a single pretrained model transfers across many language tasks. DeBERTa (He et al., 2023) refined that encoder through disentangled attention and an improved position encoding, and it became a common backbone for short constructed responses. These encoders cap input length near 512 tokens, a limit that suits short responses but truncates essays, and truncation removes exactly the later-paragraph evidence that holistic rubrics reward. Our earlier studies built on these encoders and asked a different question, namely how little labeled data a well-tuned scorer needs (Hemenway et al., 2026; Barber et al., 2025).

Long-context architectures. Several architectures extend the context window that essay scoring needs. Longformer (Beltagy et al., 2020) replaces dense attention with a sparse local-plus-global pattern, so that it can read several thousand tokens at linear cost. ModernBERT (Warner et al., 2025) reaches long contexts through rotary embeddings and Flash-Attention-2, and it modernizes the encoder recipe with faster kernels and a longer pretraining context. Mamba (Gu and Dao, 2023) departs from attention altogether and models sequences with a selective state space in linear time, which lets it scale to long inputs without a quadratic memory cost. These models target language-modeling and retrieval benchmarks, and their value for data-efficient essay scoring remains untested. We close that gap and add a causal-versus-bidirectional contrast that isolates scan direction within the state-space family.

Data efficiency and evaluation. A key practical question is how to use human labels for modeling so that time and cost are managed appropriately, which makes data efficiency as important as peak accuracy. Hemenway et al. (2026) showed that systematic hyperparameter optimization lowers the training size at which transformer scorers clear deployment bars, and Barber et al. (2025) mapped when active learning does and does not help. Both studies treated the backbone as fixed and varied the data instead. We extend this line to long-context backbones and pair rank-agreement metrics with distributional and intersectional fairness measures, because operational reporting depends on the full score distribution rather than on rank agreement alone.

3 Research Questions

1. **RQ1. Architecture under data constraints.** Which variant achieves the highest scoring quality at each training size, and does the ordering change as data increases?
2. **RQ2. Minimum viable training size.** For each architecture, what is the smallest training size that reliably meets deployment standards (val/test fixed)?
3. **RQ3. Directionality in SSMs.** Does bidirectional scanning recover the gap between Mamba and bidirectional transformers, or is any shortfall architectural rather than directional?
4. **RQ4. Distributional fidelity and fairness.** Beyond rank agreement, how well do predicted score distributions match human distributions, and do reduced training sizes introduce intersectional fairness risk?

4 Setup

Data. ASAP-2¹ provides 24,728 essays on 7 argumentative prompts scored on 1–5 or 1–6 holistic rubrics, with demographic metadata (gender, race/ethnicity, ELL, disability, economic disadvantage). The corpus is well suited to our question, because it pairs essay-length argumentative writing with the demographic fields that a fairness analysis requires. Over 95% of responses fit within 1,024 tokens, which we adopt as the operational sequence

cap. This cap keeps the long tail of unusually long essays from inflating compute cost, and it still admits the full text of nearly every response.

Splits. Validation ($n=300$) and test ($n=500$) are drawn once per prompt (stratified on score) and *held fixed* across all conditions. We hold these partitions fixed so that any change in measured quality reflects the training manipulation rather than a lucky or unlucky evaluation draw. Because the fixed validation and test partitions (800 human-scored essays per prompt) are themselves a standing labeling cost, the total labeled requirement is roughly 800 essays plus the training increment. Our data-efficiency claims therefore concern the *training* portion of that budget rather than the total. Only the training partition varies, across four sizes (128, 256, 384, and 512). For each prompt $\times$ size $\times$ architecture we draw 5 independent stratified training samples, for 560 trained models in all (7 prompts $\times$ 4 sizes $\times$ 4 architectures $\times$ 5 replicates). The replicates give us a sampling distribution over training draws, and they let us attach confidence intervals to every reported mean.

Architectures. We compare 4 variants. ModernBERT-base and Longformer-base-4096 use a [CLS] regression head. Two Mamba variants come from the size-matched mamba-130m checkpoint, a causal SSM with an attention-pooling head and a weight-tied bidirectional variant. The bidirectional variant runs a forward and a backward selective scan, then concatenates and mean-pools the two passes, so that every position can draw on both left and right context. We tie the weights of the two scans so that the directional contrast does not also double the parameter count, which keeps the comparison clean. Mamba runs through Hugging Face `transformers` with fast CUDA kernels (`mamba-ssm` and `causal-conv1d`).

Training. Each cell receives an independent Op-tuna (Akiba et al., 2019) Tree-structured Parzen Estimator (TPE) sweep, with 18 trials for transformers and 10 for the slower Mamba arm. A patient percentile pruner operates over a frequency-balanced MSE objective, so that the search spends its budget on promising configurations and stops weak trials early. Optimization uses AdamW with gradient clipping, mixed precision, and early stopping on validation QWK. The HPO winner then receives a *green surrogate* fine-tune: a short, head-only fine-tune against a differentiable surrogate of

¹<https://www.kaggle.com/competitions/learning-agency-lab-automated-essay-scoring>

the green deployment criteria, a composite of differentiable QWK, SMD, and MSE that nudges the model toward the deployment criteria rather than toward squared error alone. A post-hoc affine calibration (Kuleshov et al., 2018), fit without leakage on training predictions, then aligns the predicted score mean and spread with the human scale.

Agreement metrics. Per prompt we report quadratic weighted kappa (QWK; Cohen, 1968), Pearson r , the standardized mean difference (SMD) between model and human score means, and exact and adjacent agreement. QWK is our headline metric, because it credits near-misses on an ordinal scale and penalizes large disagreements more heavily. The remaining metrics guard against the failure modes that a single number can hide, such as a systematic bias in the mean or a collapse onto the central score points.

Deployment status. Our operational practice, one adaptation and extension of Williamson et al. (2012), grades each model *green*, *yellow*, or *red* against the fixed thresholds in Table 1. A model earns *green* only when it clears both the *primary* agreement bars ($\text{QWK} \geq .70$, $r \geq .70$, $|\text{SMD}| \leq .15$) and the *secondary* distribution-shape bars (exact $\geq .65$,² adjacent $\geq .95$, and every per-score-point frequency within 4 percentage points of the human distribution). A model that clears all primary bars but misses a secondary bar earns *yellow*, and a model that misses any primary bar earns *red*. Green/yellow marks the operational deploy/monitor boundary and is the attainment target we report throughout.³

Why distribution-level metrics. Rank agreement (QWK, r) can stay high even when a model systematically compresses the score distribution and avoids the rubric extremes. Such compression distorts the aggregate score-band statistics that operational score reports depend on. Relatedly, agreement coefficients such as QWK are themselves sensitive to the marginal score distri-

²We adopt the conventional exact-agreement bar of .65 used for 4-point scales. Because our 5- and 6-point holistic rubrics spread credit over more categories, exact agreement rarely exceeds $\sim .71$ here, so this is a stringent bar for the green (full-deploy) tier. We report its effect below and find the deploy/monitor boundary insensitive to it.

³Our bars are absolute, which aligns with Williamson et al. (2012). We do not additionally require that model-human agreement fall within a margin of human-human agreement, a relative criterion also used operationally. We treat that as a complementary check (see Limitations).

Tier	Criterion	Threshold
Primary	QWK	≥ 0.70
	Pearson r	≥ 0.70
	$ \text{SMD} $	≤ 0.15
Secondary	Exact agreement	≥ 0.65
	Adjacent agreement	≥ 0.95
	Max score-point freq. diff.	≤ 4 pts

Table 1: Deployment-status criteria. **Green** = all primary + all secondary met; **yellow** = all primary met, ≥ 1 secondary fails; **red** = ≥ 1 primary fails.

butions of the two raters and can be inflated or deflated when those distributions diverge (Lewis and Casabianca, 2026), which is a further reason to measure distributional fidelity directly rather than to read it off a single agreement number. We therefore add three distance metrics between the predicted and human score *distributions*. Writing p for the human and q for the predicted score-frequency distribution over the ordinal points, KL divergence $\text{KL}(p \parallel q) = \sum_i p_i \log(p_i/q_i)$ (Kullback and Leibler, 1951) is asymmetric and penalizes missed mass; Jensen-Shannon divergence, the symmetric, bounded smoothing of KL (Lin, 1991), is well behaved when a score point is empty; and Earth Mover’s (Wasserstein-1) distance (Rubner et al., 2000) respects the *ordinal* scale. Earth Mover’s distance charges more to move probability mass from a 1 to a 6 than from a 5 to a 6, whereas KL and JSD do not. All three are non-negative and equal zero only when the predicted and human distributions coincide, so smaller values mean a closer match and 0 is ideal. These metrics are independent of the directionality question. They test distributional fidelity (RQ4) and in fact surface a ranking that the agreement metrics alone conceal (Section 5.4). This distributional emphasis follows a presentation by Ormerod (2026) on long-sequence scoring with Mamba.

Fairness. We compute intersectional fairness (gender $\times$ race/ethnicity, cells with ≥ 30 test essays) on the four largest prompts and report per-cell QWK, SMD, and mean residual.

5 Results

5.1 Architecture $\times$ training size (RQ1)

Table 2 reports mean test QWK by architecture and training size. The top of the leaderboard is a statistical tie among three very different backbones. ModernBERT (0.741 [0.731, 0.750]), bidi-

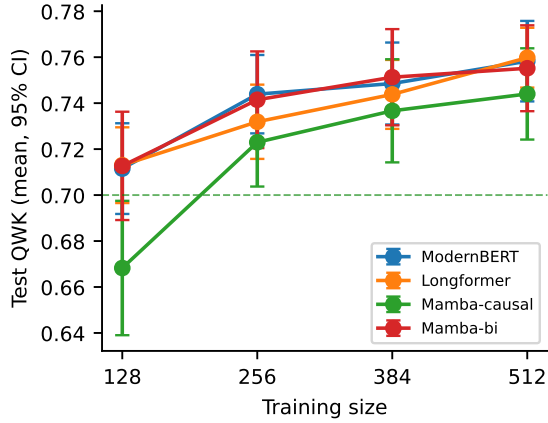


Figure 1: Test QWK vs training size per architecture (mean, 95% CI). The dashed line marks the QWK = 0.70 deployment threshold.

rectional Mamba (0.740 [0.730, 0.751]), and Longformer (0.737 [0.729, 0.745]) are indistinguishable, and causal Mamba trails by roughly .02 QWK (0.718 [0.706, 0.730]). The three leaders span dense attention, sparse attention, and a bidirectional state space, so the tie is notable. It suggests that the rubric signal these models must learn is accessible to several very different sequence-mixing mechanisms. The ordering stays stable across training sizes, yet every model improves markedly as the training set grows from 128 to 512 (Fig. 1). The steepest gain belongs to causal Mamba (0.668→0.744), which is also the only model that sits below the QWK=.70 deployment bar at 128. The size effect therefore dwarfs the architecture effect across the range we study, and the gap between the best and worst backbone at a fixed size is smaller than the gain any single backbone earns by moving from 128 to 512 essays.

5.2 Deployment-status attainment (RQ2)

Table 3 reports the fraction of configurations that reach green or yellow status (the combined deploy-or-monitor rate). Attainment climbs monotonically with training size, from 60% at 128 to 76% at 256, 81% at 384, and 84% at 512. Separating the two tiers, the stricter green-only rate (all primary and secondary bars met, i.e. deploy without monitoring) rises in parallel but stays well below the combined rate, from 1% at 128 to 1% at 256, 5% at 384, and 9% at 512; the bulk of the gain at each size is therefore in yellow (deployable with monitoring) rather than in green. Because exact agreement is a *secondary* criterion, the choice of its bar shifts

models only between green and yellow and leaves the combined deploy/monitor rate untouched: the green-only rate is sensitive to the .65 exact bar (few models exceed it on 5- and 6-point rubrics), whereas the green/yellow rate is invariant to it. The pattern holds across architectures, so no backbone earns deployable status at 128 that the others cannot also reach by 512.

5.3 Directionality in Mamba (RQ3)

The within-Mamba contrast (bidirectional – causal) is +0.022 QWK ($p = .007$, Welch’s t over matched architecture/size cells). The effect concentrates in the data-scarce regime. At 128 training essays a bidirectional scan lifts Mamba from 0.668 to 0.713 and erases the entire gap to the transformer baselines, whereas at 512 the two Mamba variants nearly converge (0.744 vs. 0.755). The pattern fits a simple reading. When labels are scarce, the model cannot afford to relearn from data the right-context dependencies that a backward scan supplies for free, so the architectural prior matters most. When labels are plentiful, the causal model can recover much of that signal from the data itself, so the two variants converge. Any “SSM shortfall” for AES is therefore largely a *directionality* artifact rather than an intrinsic limitation of the selective state-space backbone.

5.4 Distributional fidelity (RQ4)

Table 4 reports KL, JSD, and Earth Mover’s Distance between predicted and human score distributions. Here rank agreement and distributional fidelity come apart. Bidirectional Mamba produces the most human-like score distributions on the divergence measures, with the lowest KL and JSD, and it is essentially tied with ModernBERT on EMD. Longformer ranks competitively on QWK yet is the least faithful on the distribution, with the highest KL and EMD, so a strong rank score does not guarantee a human-like spread. A transformer can win on rank order while it quietly compresses the score distribution toward the center, and a state-space model can reproduce the human spread while it ranks essays slightly less sharply. Figure 2 compares EMD across architectures and sizes, and it shows that the bidirectional-Mamba advantage holds across training sizes rather than at one alone. We read these distance metrics conservatively. The robust, repeated pattern is that Longformer and causal Mamba are least faithful while bidirectional Mamba and ModernBERT are

Architecture	Train=128	Train=256	Train=384	Train=512	Overall
ModernBERT	0.712	0.744	0.749	0.758	0.741
Longformer	0.713	0.732	0.744	0.760	0.737
Mamba-causal	0.668	0.723	0.737	0.744	0.718
Mamba-bi	0.713	0.741	0.751	0.755	0.740

Table 2: Mean test QWK by architecture $\times$ training size, averaged over 7 prompts $\times$ 5 replicates.

Architecture	Train=128	Train=256	Train=384	Train=512
ModernBERT	71%	86%	83%	83%
Longformer	54%	66%	71%	80%
Mamba-causal	46%	74%	83%	86%
Mamba-bi	69%	80%	86%	86%

Table 3: Percentage of configurations achieving green/yellow deployment status by architecture $\times$ training size.

Architecture	KL	JSD	EMD
ModernBERT	0.138	0.008	0.100
Longformer	0.220	0.013	0.141
Mamba-causal	0.190	0.013	0.137
Mamba-bi	0.087	0.007	0.101

Table 4: Distributional fidelity (mean; lower is closer to the human score distribution).

most faithful. The small remaining gap between the two leaders (e.g., ModernBERT vs. bidirectional Mamba on EMD at 512) is within run-to-run variation across replicates, so we treat them as effectively tied on distributional fidelity rather than ranking one above the other.

5.5 Intersectional fairness (RQ4)

On the four largest prompts, the worst subgroup QWK gap across gender $\times$ race/ethnicity cells is 0.169, with a mean of 0.133 across architecture/size cells. Pooled across architectures, per-cell QWK ranges from 0.626 to 0.758 (Table 5). These gaps are comparable across all four architectures and do not widen systematically when the training set drops from 512 to 128. The stability across the size cut is the result that bears on practice, because a program that trains on fewer essays to save labeling cost does not appear to trade away subgroup parity to do so. Subgroup disparity here therefore tracks the task and labels rather than the backbone, and a program that wants to narrow these gaps should look to the data and the rubric before it looks to the model. A subgroup-

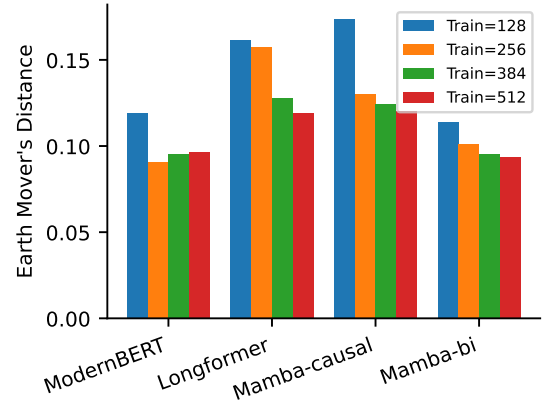


Figure 2: Earth Mover's Distance by architecture and training size.

level caveat is warranted. Williamson et al. (2012) treats a subgroup $|SMD| > .10$ as a fairness violation, a stricter bar than the .15 overall bar we use for status. Under that subgroup bar the M Black cell ($|SMD| = 0.163$) would be flagged, and the next-largest cells approach it, indicating a localized calibration offset concentrated in the mean rather than in rank agreement. This is a mean-shift that a per-subgroup affine adjustment could in principle address, and we flag it as a target for monitoring rather than a property that the size cut induces.

6 Discussion

Data quantity outranks backbone choice. The three leading backbones differ in attention pattern, position encoding, and even sequence-mixing paradigm, yet their mean QWK falls within a single confidence interval. What moves quality is

Subgroup	n	QWK	SMD	Resid.
M White	105	0.758	0.058	+0.010
F White	83	0.749	0.069	+0.030
M Hispanic	74	0.733	0.081	+0.050
F Black	56	0.713	0.094	-0.015
F Hispanic	77	0.709	0.086	+0.048
M Black	52	0.626	0.163	+0.008

Table 5: Per-subgroup results (gender $\times$ race/ethnicity), pooled across architectures and training sizes on the four largest prompts. n is the mean per-cell test count; QWK, |SMD|, and mean residual are cell averages.

the crossing of the label-quantity threshold, where green/yellow deployment attainment rises from 60% (128) to 84% (512). A program that chooses infrastructure should therefore pick the cheapest adequate backbone and spend its budget on labels rather than on architecture search. This advice runs against a common instinct to chase the newest architecture, and our evidence suggests that the instinct is misplaced for data-efficient essay scoring. The marginal label buys more measured quality than the marginal architecture change, at least across the backbones and sizes we study.

For state-space models, scan bidirectionally. The one architectural decision that demonstrably matters is directionality. Bidirectional Mamba matches the transformers everywhere and beats causal Mamba by +0.022 QWK, and the gain concentrates where data is scarcest. The size-matched 1.30m checkpoint is the same for both Mamba variants, so this contrast is not confounded with capacity. It is a clean attribution to scan direction. The practical lesson is narrow but firm. A team that adopts a state-space backbone for scoring should pay the modest cost of a backward scan, because the alternative inherits a left-to-right bias that the scoring task does not share.

Distribution-matching favors the bidirectional state space. Bidirectional Mamba is the most distribution-faithful model on the divergence measures (lowest KL and JSD), and it ties ModernBERT for the lowest EMD, even though it does not lead on rank agreement. Where operational reporting depends on reproducing the human score *distribution* (e.g., aggregate score-band statistics), the bidirectional state space is a sensible default. Its distributional edge over ModernBERT is modest and not always statistically separable (the two are effectively tied on EMD), so the practical case rests as much on cost as on accuracy. It trains in

roughly half the wall-clock of Longformer. A program that cares most about rank order can pick any of the three leaders, and a program that must reproduce the score distribution has a reason to prefer the bidirectional state space or ModernBERT over Longformer.

Fairness. Subgroup QWK gaps across gender $\times$ race/ethnicity cells are non-trivial (max 0.169, mean 0.133) and comparable across all four architectures. A cut in training data from 512 to 128 does not systematically widen them. Fairness in this setting is thus more a property of the task and labels than of the backbone. It also redirects fairness effort toward the data and the rubric, where the gaps appear to originate.

7 Conclusion

We compared transformer and state-space backbones for data-efficient essay scoring under matched and explicitly small label budgets. No single architecture dominated, and the choice between dense attention, sparse attention, and a bidirectional state space mattered far less than the quantity of labeled data. The one architectural decision that mattered was scan direction within the state-space family, where a tied backward scan closed the gap to the transformers in the data-scarce regime. The bidirectional state space also reproduced the human score distribution most faithfully on the divergence measures, which makes it a sensible default when operational reporting depends on distributional fidelity. For a program that must commit infrastructure before it collects data, the practical recipe is to pick a cheap adequate backbone, scan bidirectionally if that backbone is a state-space model, and invest the remaining budget in labels.

8 Limitations

Several limits constrain these results. We study a single dataset (ASAP-2), holistic rather than multi-trait scoring, and one discourse mode (argumentative), so the findings may not transfer to narrative prompts or to analytic rubrics that score traits separately. We test four training sizes (128, 256, 384, and 512), which trace the low-data regime in some detail but stop at 512, so the scaling curve beyond that point remains uncharacterized. The SSM fine-tuning recipe is also less established than the transformer recipe, a risk the gating pilot mitigates but does not remove. Finally, we report results at a

single model scale, and a larger or smaller checkpoint could shift the balance between the backbones. ASAP-2 also provides a single human score per essay, so we cannot compute human–human (h1–h2) agreement as a reference ceiling for the model–human QWK or evaluate the relative degradation criterion noted in Section 4; benchmarking model–human against human–human agreement on double-scored data is a natural direction for follow-up work. We see each of these as a direct invitation for follow-up work rather than as a threat to the central finding, which is that label quantity governs deployable quality more than backbone choice.

References

- Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2623–2631.
- Justin O Barber, Michael P Hemenway, and Edward Wolfe. 2025. When does active learning actually help? empirical insights with transformer-based automated scoring. In *Proceedings of AIMECon*.
- Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.
- Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.
- Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations (ICLR)*.
- Michael P Hemenway, Justin O Barber, and Martha Bellows. 2026. Systematic hyperparameter exploration for minimal training data in transformer-based automated scoring models. In *Proceedings of NCME*.
- Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 6300–6308.
- Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. 2018. Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 2796–2804.
- Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.
- Jennifer Lewis and Jodi M. Casabianca. 2026. Limitations of qwk in evaluating automated and human scoring systems. *Educational Measurement: Issues and Practice*, 45(1).
- Jianhua Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.
- Christopher M. Ormerod. 2026. Paying attention to mamba: Long-sequence scoring beyond transformers. <https://www.ncme-aime.org/events.html>. Invited presentation, NCME AIME Webinar Series, June 10.
- Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 2000. The earth mover’s distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121.
- Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1882–1891.
- Masaki Uto, Yikuan Xie, and Maomi Ueno. 2020. Neural automated essay scoring incorporating hand-crafted features. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 6077–6088.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, and 1 others. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547.
- David M Williamson, Xiaoming Xi, and F Jay Breyer. 2012. A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1):2–13.