

Math-Specialized or General-Purpose? A Comparison of MathBERT and DeBERTa-v3-large for Automated Scoring of Mathematics-Explanation Items

Gregory M. Jacobs Ahmed H. Bediwy Martha Bellows

Pearson Education

{greg.jacobs, ahmed.bediwy, martha.bellows}@pearson.com

Abstract

Automated scoring of math-explanation items should handle responses that mix natural language with symbolic reasoning. The 2023 NAEP Automated Scoring Challenge showed that fine-tuning pre-trained encoders can reach human-like agreement, but its winning system used a large, general-purpose encoder, while smaller math-specific pre-trained encoders remain a plausible and more efficient alternative. We fine-tune MathBERT (110M parameters) and DeBERTa-v3-large (183M parameters) as single-output regression scorers on 31 math-explanation items from one statewide summative assessment administration, benchmarking every model against human-human reliability and summarizing deployability with a three-tier acceptance criteria. On the 18 items fine-tuned under both encoders with identical datasets, the resulting performance between the two was practically indistinguishable: mean best QWK was 0.931 for MathBERT and 0.933 for DeBERTa-v3-large—a difference of only 0.002—and the models traded top performance at similar rates. Given similar performance, we tested a two-stage approach that used the larger general-purpose encoder only when the smaller math-focused encoder failed acceptance criteria. Extending MathBERT to all 31 items, it reached a mean QWK of 0.926 against a human-human benchmark of $\text{QWK} = 0.940$ and met operational acceptance criteria on 81% of items. Escalating the six items where a fine-tuned MathBERT model failed acceptance to a fine-tuned DeBERTa-v3-large model recovered four of the six, yielding acceptable automated-scoring models for 29 of 31 items. Because the smaller, math-pre-trained encoder matches the larger one at roughly 40% fewer parameters, we recommend fine-tuning a MathBERT encoder as the default and escalating to the larger, general purpose encoder like DeBERTa-v3-large only when a MathBERT model fails when deploying automated scoring models operationally at scale for math-explanation items.

1 Introduction

Constructed-response items that ask students to *explain* their mathematical reasoning yield valuable evidence about conceptual understanding, but hand-scoring them at the scale of a statewide summative assessment is resource intensive and time consuming. Automated scoring seeks to reduce those costs and streamline the scoring timeline. Math explanation items, however, are challenging to score automatically (Hellman et al., 2023; Morris et al., 2025b). A single response can interweave natural-language argumentation, symbolic notation, procedural steps, and domain-specific vocabulary, so a model must reason over a mixture of forms that general-purpose encoders were not designed to represent. The difficulty is compounded by what these items ask for. Students must supply both a mathematically correct solution and an account of the reasoning that produced it, so a scoring model must evaluate both at once (Hellman et al., 2023).

The 2023 NAEP Mathematics Automated Scoring Challenge established an important proof of concept: fine-tuned transformer models can reach *human-like* agreement—model-human quadratic weighted kappa within roughly 0.05 of human-human reliability—on most math-explanation items (Crossley et al., 2023). The winning system was built on DeBERTa-v3-large, a 183M-parameter general-purpose encoder (He et al., 2021, 2023), while the runner-up used a math-specialized RoBERTa variant (NAEP Research and Development Program, 2023). The broader literature highlights that smaller, math-pretrained encoders such as MathBERT (Shen et al., 2021) are potential efficient alternatives. This raises a question the challenge did not resolve: should a smaller math-specialized encoder, rather than a larger general-purpose one, be used to score math-explanation items operationally at scale?

This paper seeks to answer that question. We

fine-tune two encoders, a smaller specialized MathBERT encoder (110M parameters) and a larger general-purpose DeBERTa-v3-large encoder (183M), as single-output regression scorers. To keep the comparison interpretable, we hold the assessment context fixed. The corpus is 31 math-explanation items from one administration of a single statewide summative assessment.

We contribute a head-to-head comparison of a math-specialized and a general-purpose transformer on operational data, under a protocol that references every model against human–human reliability and summarizes deployability with a three-tier agreement-to-standard flag that programs can adopt directly. Concretely, we ask: (RQ1) Does a fine-tuned math-specialized encoder match the performance of a larger general-purpose encoder on math-explanation items? and (RQ2) Does escalating to a larger general-purpose encoder for items where a fine-tuned math-specialized encoder does not meeting acceptance criteria improve performance? Based on the answer to these questions, we propose a training pipeline that can be deployed to fine-tune encoders for automated scoring of operational math explanation items at scale.

2 Related Work

Automated scoring of constructed responses:

Established evaluation frameworks in educational measurement hold that an engine’s usefulness should be judged not by raw accuracy but by its agreement with human raters *relative to* the agreement humans achieve with one another, together with fairness and score-distribution considerations (Bennett and Zhang, 2011; Pearson and ETS, 2015; Williamson et al., 2012; Yan et al., 2020). Two properties of operational data shape our method. First, rubric scores are ordinal, so regression under a mean-squared-error objective can match or exceed cross-entropy classification because it respects the ordering of score points (Morris et al., 2025a). Second, severe class imbalance is endemic (Crossley et al., 2023; Yan et al., 2020).

Transformer encoders and domain-adaptive pre-training:

The transformer architecture (Vaswani et al., 2017) and its bidirectional encoder instantiation BERT (Devlin et al., 2019) underpin most modern scoring systems. DeBERTa introduced disentangled attention (He et al., 2021), and DeBERTa-v3 improved pre-training with an ELECTRA-style objective (He et al., 2023). In parallel, encoders

pretrained on mathematical corpora—MathBERT (Shen et al., 2021) and math-specialized RoBERTa variants (NAEP Research and Development Program, 2023)—aim to capture the notation and discourse of mathematics with far fewer parameters. Our comparison is a controlled test of the tradeoff: MathBERT embodies the domain-pretraining approach and DeBERTa-v3-large the scale approach, fine-tuned under an identical harness so that only the encoder choice differs.

The NAEP challenge and operational deployment:

The top-performing systems in the NAEP challenge were optimized for peak accuracy on a fixed benchmark, using extensive data-augmentation and ensembling pipelines (Crossley et al., 2023; NAEP Research and Development Program, 2023). Operational programs face a different objective: sustaining human-like agreement at acceptable performance metrics across many items and administrations in an efficient manner, with score distributions close enough to those of human raters that reported results stay comparable year over year. We focus explicitly on identifying the optimal training pipeline for operational programs.

3 Corpus

3.1 Data

Every statistic in this paper is computed from a single administration of 31 math-explanation items from one statewide summative assessment. The items span from grades 3 through 8 and apply rubrics with scorepoint ranges that span from two (0-1) up to five (0-4) integer values.

3.2 Human scoring

Responses were scored under an operational double-read procedure whereby 10% of responses received a second score to measure human-to-human reliability. For both training and evaluation, we take the target label to be `best_human`: the program’s final reported score, determined by a predefined hierarchy in which an adjudication or resolution score takes precedence over the first and second reads. Every metric we report is computed on a held-out validation set of 2,500 responses per item. Human–human reliability is taken from the first and second reads on those same set of responses, so the human metrics are calculated on the smaller 10% of those responses that received a second human score (approximately 250 validation responses per item).

3.3 Response pre-processing

Operational responses arrive in a semi-structured delivery format rather than as clean text, and they carry markup that a language model should not see verbatim. Before fine-tuning, we apply a deterministic cleaning pipeline: we unwrap the JSON delivery format, decode HTML entities and strip presentational tags, convert embedded MathML into a plain-text rendering so that symbols, fractions, and operators survive as readable tokens, and normalize Unicode. The pipeline is identical across all items and both encoders, so no difference between models can be attributed to pre-processing.

4 Modeling Methodology

4.1 Data splits

For each item we drew a random sample of 7,500 responses and stratified it into 5,000 for training—five folds of 1,000 for cross-validated hyperparameter selection—and 2,500 for validation. Stratification preserves the score-point distribution across splits, and all models for an item share the same splits, so comparisons are paired.

4.2 Transformers implementation

Both encoders use transformers (Wolf et al., 2020) with each model’s native tokenizer, AdamW (Loshchilov and Hutter, 2019) under a linear warmup schedule, a physical batch size of 4 with four gradient accumulations (effective batch size of 16), and early stopping on validation QWK. The harness is shared, so architectural rather than harness differences drive any performance gap.

4.3 Hyperparameter grid search

Hyperparameter values are identified using the training dataset to perform a five-fold stratified cross-validation grid search, varying learning rate, weight decay, and label-smoothing α parameter values and evaluating each by mean validation QWK. Once the optimal hyperparameter values are identified through the grid search, a final model is trained using those optimal parameter values.

4.4 Model fine-tuning

Each encoder is fine-tuned with a single-output regression head, which respects the ordinal structure of rubric scores (Morris et al., 2025a). To counter imbalance, we minimize a sample-weighted mean-squared-error loss with class-balanced label smoothing. Let $\hat{y}_i$ be the model’s

continuous prediction for response i , let y_i be that response’s human score on a rubric with score points $\{0, \dots, K - 1\}$, and let f_k be the empirical frequency of score point k in the training set. With per-response weight w_i and smoothing coefficient $\alpha \in [0, 1]$,

$$w_i = (1 - \alpha) \frac{1/f_{y_i}}{\frac{1}{K} \sum_k 1/f_k} + \alpha, \quad (1)$$

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N w_i (\hat{y}_i - y_i)^2,$$

which reduces to inverse-frequency (class-balanced) weighting at $\alpha = 1$ and to uniform weighting at $\alpha = 0$, up-weighting rare high-score responses without destabilizing training on the dominant low-score class. Smoothing α is treated as a hyperparameter because the right amount of correction varies with how skewed a item’s score distribution is.

4.5 Evaluation and deployability

Models are evaluated with four complementary statistics, each benchmarked against the corresponding human–human value so that “good” always means “close to what two humans achieve.” *Quadratic weighted kappa* is the primary statistic (Cohen, 1968): disagreements are weighted by squared distance, which suits ordinal scores. *Exact agreement* is the proportion of responses assigned the same score point by model and human. The *standardized mean difference* captures systematic leniency or severity. Finally, *score-point distribution match* compares model and human marginal frequencies, guarding against a model that agrees well while distorting the reported distribution and breaking year-over-year comparability.

These four statistics are combined into a three-tier deployability flag (Figure 1). A *green* flag requires that (i) model-minus-human QWK ≥ -0.10 , (ii) the exact-agreement gap ≥ -0.0525 , (iii) $|SMD| \leq 0.15$, and (iv) every score-point frequency falls within four percentage points of the human distribution. *Yellow* means the agreement metrics pass but a score-point frequency falls outside that band; *red* means at least one agreement metric is not met. The thresholds are *deltas* against each item’s own human–human values, so an item on which humans agree strongly is held to a stricter standard.

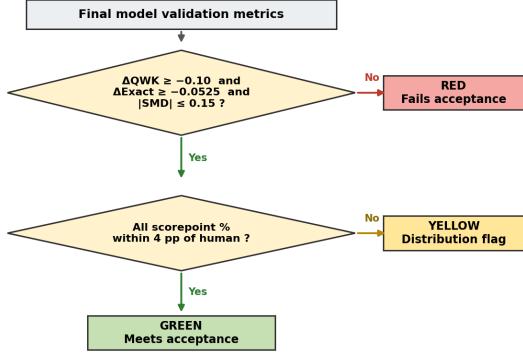


Figure 1: The three-tier agreement-to-standard flag. Green requires human-like agreement on all metrics and a close score-point distribution match.

These model evaluation metrics are considered industry standard and are consistently applied within the automated scoring field (Pearson and ETS, 2015; Williamson et al., 2012).

5 Initial Modeling Results

The initial round of modeling involved fine-tuning both MathBERT and DeBERTa-v3-large models, using a robust hyperparameter grid search, on only a subset of 18 of the 31 items to conserve resources. The purpose of this initial experimentation was to (1) evaluate the effectiveness of the robust hyperparameter grid search to determine whether a more limited search could be used at scale, and (2) compare the performance metrics between the two encoders to determine which one is more effective at scoring math-explanation items.

5.1 Hyperparameter grid search

For each encoder, we elected to search over learning rate, weight decay, and label-smoothing α and set the values for each based on the best performing fine-tuned models surveyed during the 2023 NAEP Mathematics Automated Scoring Challenge referenced above. Table 1 summarizes the hyperparameter values used in the grid search.

The main findings from this grid search results were that (1) the only hyperparameter that potentially should be tuned at scale for both encoders is learning rate, and (2) that the number of epochs should be increased with early stopping to prevent wasted resources. The "best" performing models for both encoders used the largest learning rate value a majority of the time (albeit with a marginally better QWK value), suggesting that a

Hyperparameter	Values searched
Learning rate	1e-5, 2e-5 (MathBERT); 1e-5, 1.5e-5, 2e-5 (DeBERTa)
Weight decay	0.01, 0.05 (MathBERT); 0.1, 0.3 (DeBERTa)
Smoothing α	0.5, 0.7
Batch size	4 (with 4 grad. accumulation)
Epochs	2
Scheduler	linear with warmup

Table 1: Stage 1 search space on 18 items. Grids yielded 8 (MathBERT) and 12 (DeBERTa-v3-large) configurations over 5 folds.

grid search with larger learning rate values could be implemented out of an abundance of caution.

5.2 Head-to-head comparison

Table 2 compares the encoders on the 18 items where both models were run using identical splits, scoring each by its *best* configuration. Mean best QWK was 0.931 for MathBERT and 0.933 for DeBERTa-v3-large—a negligible difference of 0.002. The models traded wins at similar rates, 8 for MathBERT to 10 for DeBERTa-v3-large, with a mean absolute per-item exact agreement gap of only 0.004. The model results on SMD was similarly equivalent with DeBERTa-v3-large winning by 0.04. The fine-tuned MathBERT models did generate a lower mean SMD value and passed the acceptance criteria at a marginally larger rate, suggesting that it can be fine-tuned to score math-explanation items with the same, if not greater, success rate as the larger DeBERTa-v3-large model.

6 Final Training Results

6.1 Proposed Operational Training Pipeline

Based on the initial training results, we executed an operational training pipeline across all 31 items using a revised, more targeted hyperparameter grid search focused on the learning rate hyperparameter and the number of training epochs (with early stopping implemented). Because in our initial experimentation both encoders performed slightly better at higher learning rate values, we chose to expand the grid search at the upper bounds. For MathBERT, we used learning rate values of 2e-5, 3e-5, 4e-5, and 5e-5, and allowed training to run for up to 6 epochs. For DeBERTa-v3-large, we used learning rate values of 1.5e-5, 2e-5, 3e-5, and 4e-5, and allowed training to run for up to 4 epochs.

Additionally, we implemented a two-stage training approach. For stage 1, we fine-tuned only Math-

Configuration	Items	Mean QWK	Mean exact	Mean $ SMD $	% green
Human-human reference	31	0.940	0.902	—	—
<i>Initial Modeling Results (18 items)</i>					
MathBERT	18	0.931	0.852	0.036	78%
DeBERTa-v3-large	18	0.933	0.856	0.056	67%
<i>Final Modeling Results (31 items)</i>					
MathBERT	31	0.926	0.872	0.015	81%
DeBERTa-v3-large	6	0.917	0.836	0.013	67%

Table 2: Mean engine-human agreement relative to human-human reliability.

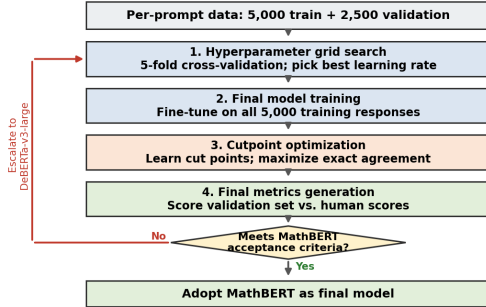


Figure 2: Per-item modeling pipeline. A model is adopted only if it meets the acceptance criteria; otherwise the program escalates to the larger encoder.

BERT models for all 31 items and evaluated performance. We selected MathBERT over DeBERTa-v3-large as our preferred encoder because it performed on par with the larger model in our initial experimentation, it can be fine-tuned more efficiently, and it is specifically designed to evaluate text that includes math notation. If a fine-tuned MathBERT model passed the acceptance criteria, we selected it as the model to be used operationally.

If a fine-tuned MathBERT model was unable to pass the acceptance criteria for an item, a second round of training was performed where we fine-tuned a DeBERTa-v3-large using the same training pipeline. If a fine-tuned DeBERTa-v3-large model passed the acceptance criteria, we selected it as the recommended model to be used operationally for the item. Figure 3 provides a visual representation of this workflow.

6.2 Final Modeling Performance

Fine-tuning a MathBERT model to meet acceptance criteria was successful for 25 of the 31 items, producing an 81% success rate. For the six items where a fine-tuned MathBERT model failed to pass acceptance criteria, all six failed because of the requirement by producing a model-to-human score

	Human-human	MathBERT	DeBERTa
Item 1	0.949	0.858	0.871
Item 2	0.822	0.762	0.738
Item 3	0.852	0.796	0.820
Item 4	0.874	0.798	0.863
Item 5	0.878	0.814	0.841
Item 6	0.909	0.850	0.884
Mean	0.881	0.813	0.836

Table 3: Exact agreement on the six items MathBERT failed to earn green, showing human-human reliability alongside MathBERT and the escalated DeBERTa-v3-large model agreement rates.

agreement rate greater than 0.0525.

A second round of training to fine-tune DeBERTa-v3-large for these items produced a passing model for four of the six items. So, the proposed operational training pipeline was able to produce acceptable automated scoring models for 29 of 31 items in the subject administration. Table 4 outlines the exact agreement rates for humans and the best MathBERT and DeBERTa-v3-large models.

7 Discussion

We organize this discussion around the two research questions posed in Section 1.

RQ1: Does a fine-tuned math-specialized encoder match the performance of a larger general-purpose encoder on math-explanation items? Yes, if both are fine-tuned under an identical training harness. On the 18 items run under both encoders with identical splits, mean best QWK differed by only 0.002 (0.931 for MathBERT versus 0.933 for DeBERTa-v3-large), the models produced the best performance at similar rates (8 for MathBERT to 10 for DeBERTa-v3-large), and the models performed quite similarly across the other three acceptance criteria metrics (exact agreement, SMD, and scorepoint distribution). We therefore

suggest an approach that starts with fine-tuning the smaller, more efficient, MathBERT encoder that has been pre-trained specifically to interpret text that includes mathematical notation.

RQ2: Does escalating to a larger general-purpose encoder for items not meeting acceptance criteria improve performance? Yes. Of the six items for which a fine-tuned MathBERT encoder failed to meet acceptance criteria, a fine-tuned DeBERTa-v3-large model was able to flip the results to acceptable. This two stage modeling approach therefore worked for all but two of the items where human-to-human agreement was either abnormally high (0.949) or low (0.822), suggesting that the items may be outliers that are not well suited for automated scoring. Table 3 reports the exact agreement performance metrics for these six items.

8 Limitations

One limitation is the scope of this research. The results reported here comes from items modeled from a single administration of a single state-wide assessment. The results accordingly could lack external validity because item styles, rubric conventions, and rater training vary across programs.

Secondly, automated scores often influence high-stakes decisions, so agreement with human raters is necessary but not always sufficient. An engine must also score comparably well across student subgroups. Our analysis establishes aggregate human-like agreement but does not dis-aggregate by subgroup. That analysis should be conducted before any operational use and is an important item of remaining work.

9 Conclusion

Both a math-specialized encoder (MathBERT) and a larger general-purpose encoder (DeBERTa-v3-large) can be fine-tuned to human-like agreement on operational math-explanation items, and their respective performance is practically indistinguishable. MathBERT achieves the same performance at roughly 40% fewer parameters, allowing it to be fine-tuned more efficiently, and has been pre-trained specifically to interpret text that includes mathematical notations. We accordingly recommend it as the default pre-trained encoder to score math-explanation items, and to fine-tune DeBERTa-v3-large models only for items for which a fine-

tuned MathBERT model is unable to meet operational deployment acceptance criteria.

References

- Randy Elliot Bennett and Mo Zhang. 2011. Automated scoring of constructed-response literacy and mathematics items. Technical report, Educational Testing Service.
- Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.
- Scott Crossley, Perpetual Baffour, Tian Yu, Alex Franklin, Meg Benner, and Ulrich Boser. 2023. A large-scale corpus for assessing written argumentation: PERSUADE 2.0.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.
- 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*.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.
- Scott Hellman, Alejandro Andrade, and Kyle Habermehl. 2023. Scalable and explainable automated scoring for open-ended constructed response math word problems. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 137–147. Association for Computational Linguistics.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.
- Wesley Morris, Scott Crossley, Langdon Holmes, and Amber Trumbore. 2025a. Mean squared error versus cross-entropy loss for ordinal automated scoring of constructed responses. *Journal of Educational Data Mining*.
- William Morris, Langdon Holmes, Jinsook S. Choi, and Scott Crossley. 2025b. Automated scoring of constructed response items in math assessment using large language models. *International Journal of Artificial Intelligence in Education*, 35:559–586.
- NAEP Research and Development Program. 2023. [NAEP math automated scoring challenge results and preprint](#).

- Pearson and ETS. 2015. [Research results of parcc automated scoring proof of concept study](#). Technical report, Pearson and ETS. Accessed September 2026.
- Jia Tracy Shen, Michiharu Yamashita, Ethan Prihar, Neil Heffernan, Xintao Wu, Ben Graff, and Dongwon Lee. 2021. MathBERT: A pre-trained language model for general NLP tasks in mathematics education. *arXiv preprint arXiv:2106.07340*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.
- 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.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.
- Duanli Yan, André A. Rupp, and Peter W. Foltz. 2020. *Handbook of Automated Scoring: Theory into Practice*. Chapman and Hall/CRC.