

FINE-TUNING DOMAIN-SPECIFIC PRETRAINED TRANSFORMERS FOR AUTOMATED SHORT ANSWER GRADING AND FEEDBACK GENERATION

Nguyen Nguyen Nam* and Pham Dinh Phong

University of Transport and Communications, Hanoi, Vietnam

*Corresponding author: Nguyen Nguyen Nam, e-mail: namnn.111199@gmail.com

Received August 11, 2026. Revised September 11, 2026. Accepted September 30, 2026.

Abstract. Automatic Short Answer Grading (ASAG) can return immediate, consistent scores and explanations at a scale that individual instructors cannot match, yet a single general-purpose encoder cannot produce open-ended feedback at all. This paper fine-tunes separate pretrained models for each task: MathBERT, a mathematics-education encoder (110M parameters), for binary grading, and FLAN-T5-base, an instruction-tuned encoder-decoder (248M parameters), for feedback generation, reporting the Matthews Correlation Coefficient (MCC) alongside accuracy and macro-F1 to ensure reliable evaluation under class imbalance. We evaluated the same experimental pipeline on BERT-base and SciBERT across three seeds to test whether mathematics-specific pretraining improves performance. It does not: averaged over seeds, MathBERT ranks last of the three on five of six test splits and ties on the sixth, and on the LaTeX-dense PRM800K dataset, where it should demonstrate the strongest performance, it trails BERT-base by 0.034 in macro-F1 (0.599 vs. 0.634). Performance is usable on answers to seen questions (macro-F1 0.864 on Beetle and 0.764 on SciEntsBank) and collapses on unseen questions and domains (0.508-0.555). A zero-shot Qwen2.5-1.5B-Instruct baseline outperforms the fine-tuned encoders precisely where they collapse, surpassing them by 0.114 and 0.105 macro-F1 on unseen questions and unseen domains while losing by up to 0.329 on seen questions. Additionally, it outscores the fine-tuned feedback branch (ROUGE-L 23.82 vs. 18.75). Composing the two branches end to end on PRM800K routes 67% of explanations to steps that were already correct. These findings suggest that, for these benchmarks, the choice of pretraining corpus matters far less than the difficulty of the generalization setting.

Keywords: automatic short answer grading, MathBERT, pretrained transformers, feedback generation, educational assessment, model generalization.

1. Introduction

Automatic Short Answer Grading (ASAG) - evaluating automatically whether a free-text student response is correct - enables intelligent tutoring systems to return immediate, consistent scores and feedback at a scale instructors cannot match, and ASAG for

Vietnamese classrooms remains comparatively unexplored. Existing systems either hand-engineer lexical and syntactic features in ensembled pipelines (e.g., [1]) or fine-tune a single general-purpose pretrained encoder such as BERT (e.g., [2]); however, neither approach provides explanatory feedback to the learner, and whether a general-purpose encoder is handicapped by mathematical notation remains an empirical question. Prompting a large language model (LLM) such as GPT-4 [3] or LLaMA [4] is a third option, yet model size, deployment costs, and closed weights pose substantial practical challenges for local institutional deployment. To address these limitations, this study evaluates two compact Transformer architectures for ASAG - coupling a classification model for binary grading with a generative model that provides explanatory feedback for incorrect responses.

The Transformer [5] replaces sequential recurrent neural networks (RNNs) and long short-term memory (LSTM) networks with self-attention mechanisms over query, key, and value representations. Multi-Head Attention executes several such attention operations in parallel, enabling the model to simultaneously capture diverse semantic relationships within a short response.

BERT [2] leverages a Transformer encoder-only architecture and prepends a [CLS] token whose final hidden state serves as an aggregated sequence representation mapped by a classification head to a target level - aligning with the ASAG task structure, where a (question, response) pair is mapped to a binary correctness label. Continuing its pretraining on domain-specific corpora yields domain variants such as SciBERT for scientific text [6] and BioBERT for biomedical text [7]. In mathematics education, MathBERT [8] is a BERT variant pretrained from scratch on a mathematics-education corpus spanning kindergarten through graduate school, which raises the question this paper measures: does domain-specific vocabulary and mathematical notation translate into an ASAG advantage, and across which grade levels and subjects?

Because encoder-only models cannot generate free-text explanations, an encoder-decoder framework is required for feedback generation. T5 [9] frames every NLP task as a text-to-text problem, and FLAN-T5 [10] enhances these models through tuning, thereby improving cross-task generalization without architectural modifications. FLAN-T5-base (248M parameters) was selected for this work because text generation mandates an autoregressive decoder, instruction tuning improves output quality prior to domain fine-tuning, and its moderate parameter scale remains computationally efficient for deployment on a single mid-range graphics processing unit (GPU).

2. Related work

For pure mathematics, the MATH dataset [11] collects 12,500 competition problems with step-by-step reference solutions; PRM800K [12] adds 800,000 human-annotated step-level correctness labels, collected to train process reward models (PRMs) that supervise the reasoning process itself - assigning one label per step rather than per solution.

Transformer encoders became the standard ASAG approach soon after the instruction of BERT: Sung et al. [13] improved grading by continuing BERT's pretraining on domain textbooks - the closest precedent to the question asked in this study - Camus and Filighera [14] found RoBERTa to be the strongest across ASAG datasets, cross-

encoder and sentence-embedding formulations such as SBERT [15] score an answer against a reference, and Bonthu et al. [16] surveyed the pre-LLM period. Since 2023, the competing approach has been prompting instruction-tuned large language models (LLMs), and empirical results remain mixed: Ferreira Mello et al. [17] find GPT-4 with prompt engineering competitive but not uniformly superior to fine-tuned models, SAS-Bench [18] finds current LLMs score short answers unevenly across subjects, and Cong et al. [19] report that LLM grading accuracy on public benchmarks owes more to memorization than to reasoning. Therefore, Section 4.1 evaluates a zero-shot LLM on the same data rather than assuming superiority for either approach.

The work most related to this study is Zhang et al. [20], who used MathBERT in an in-context meta-learning framework to grade open-ended mathematics answers on an unreleased internal dataset. This study differs in four key aspects: (a) data scope extends to natural science and electrical/electronic engineering, as well as a public, LaTeX-dense pure-mathematics dataset (PRM800K); (b) it incorporates feedback generation via a dedicated encoder-decoder architecture; (c) evaluation adheres to the SemEval-2013 Task 7 standard; and (d) grading and generation utilize two task-appropriate model backbones rather than one shared encoder (Section 3.2).

Together, these studies reveal a critical gap: [20] evaluates MathBERT exclusively on an unreleased dataset, while the SemEval-2013 systems [1] that report public scores predate pretrained Transformers; moreover, neither generates explanatory feedback, and no existing study empirically evaluates whether a general-purpose encoder can perform comparably to a specialized domain model. To address this gap, this paper trains MathBERT, BERT-base, and SciBERT on the public splits using a unified experimental pipeline, and finds that the mathematics-specific model is not the top performer among the three (Section 4.1).

3. Methodology

3.1. Architectural characteristics of MathBERT relevant to grading

MathBERT utilizes a BERT-base backbone (12 encoder layers, 12 attention heads, and a 768-dimensional hidden space) but is pretrained from scratch on a mathematics-education corpus. Consequently, its specialization stems entirely from that corpus rather than any architectural modification - making it a clean benchmark to test whether that corpus helps. The underlying hypothesis tested here is that an embedding space optimized for worked-solution phrasing should grade quantitative student responses more accurately than one pretrained on general prose.

Because the pretrained corpus includes quantitative natural-science content, this study hypothesizes that MathBERT holds an advantage on STEM ASAG datasets (SciEntsBank and Beetle) and notation-dense pure mathematics problems (PRM800K) relative to an encoder pretrained on general prose. Section 4 tests this hypothesis by training BERT-base and SciBERT under the same experimental protocol, and reveals that this expected advantage does not materialize.

3.2. System architecture and model selection

To address the literature gap identified in Section 2 - specifically, that no existing MathBERT grading study generates explanatory feedback and that a single shared encoder cannot optimize both tasks simultaneously—this study proposes a dual-backbone architecture, where a specialized pretrained model is fine-tuned for each respective task (Figure 1). MathBERT is fine-tuned for binary grading, while FLAN-T5-base is independently fine-tuned for feedback generation on the PRM800K dataset, both following a unified fine-tuning methodology (Section 3.4). The rationale for this decoupling is structural: MathBERT's classification head consists of a single dense layer that projects the 768-dimensional [CLS] representation to two output logits, followed by a softmax function. In contrast, FLAN-T5-base's generative decoder processes the encoder's hidden states through a cross-attention mechanism to generate a vocabulary probability distribution at each autoregressive step. Because a two-class classification head cannot generate natural language explanations, this architectural asymmetry stems directly from the distinct objective of each task.

FLAN-T5-base was selected because instruction-tuning [10] enhances baseline output quality prior to domain fine-tuning, while its 248M parameters scale remains computationally tractable for execution on a single mid-range GPU. Both model backbones are implemented using the Hugging Face Transformers library; detailed hyperparameter configurations are given in Section 3.4.

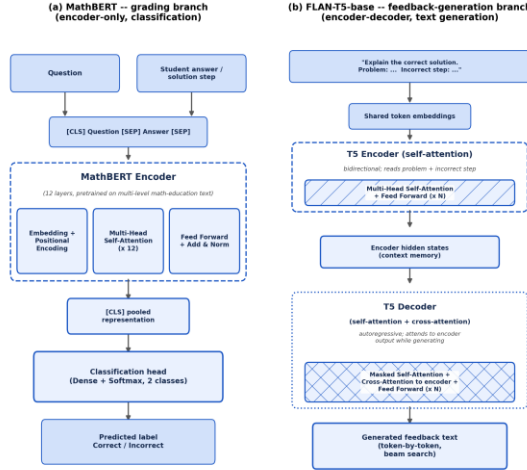


Figure 1. Two independent backbones by task: MathBERT for binary grading, FLAN-T5-base for feedback generation

Two additional design choices were implemented. First, the classification branch incorporates a class-weighted cross-entropy loss function to mitigate severe class imbalance, as the majority (correct) class predominates across all three benchmarks (Table 1). An unweighted loss function would cause the model to collapse into a majority-class predictor - a failure mode analyzed in Section 3.5. Second, the generation branch needs no loss reweighting, as its objective is to generate a natural language reference solution rather than predict a categorical label. Furthermore, the two branches differ in their data formulation: the classification branch processes a (question, response) pair paired with a single binary label, whereas the generation branch operates on a (problem,

incorrect step, reference solution) triple. The latter provides rich step-level supervision, enabling faster model convergence in fewer training epochs.

Concretely, the grading branch formats each input instance as [CLS] question [SEP] student response [SEP], allowing the self-attention mechanism to contextualize every response token relative to the question, while only the pooled [CLS] token representation is fed to the classification head. Conversely, the generation branch receives an instruction prompt (“Explain the correct solution. Problem: ... Incorrect step: ...”) and decodes free text explanations. Consequently, this dual-branch design enables a single student error to be both evaluated and explained, albeit at the computational cost of maintaining two separate model parameter sets (110M and 248M) rather than one.

Table 1. Statistics of the three datasets used in this study

Dataset	Original task	Scale	Train	Test (UA)	Test (UQ)	Test (UD)
SciEntsBank	Short-answer grading	135 questions / 12 domains	4,969	540	733	4,562
Beetle	Short-answer grading	36 questions / 1 domain	3,941	439	819	—
PRM800K	Step-level correctness → reformulated binary grading	800,000 raw labels; 20,000 solutions read, 8,000 sampled	6,821	—	1,179	—

All counts are measured from the released splits used here; the SciEntsBank figures are the SemEval-2013 training partition, not the full corpus.

Table 2. Training cost and epoch selected by early stopping at seed 42 under one identical protocol

Encoder	Dataset	Steps/epoch	Time/epoch	Best epoch	Dev MCC
MathBERT	SciEntsBank	274	~199 s	11/12	0.543
MathBERT	Beetle	217	~156 s	6/12	0.682
MathBERT	PRM800K	376	~280 s	2/12	0.307
BERT-base	SciEntsBank	274	~195 s	4/12	0.608
BERT-base	Beetle	217	~153 s	6/12	0.690
BERT-base	PRM800K	376	~272 s	3/12	0.355
SciBERT	SciEntsBank	274	~200 s	5/12	0.631
SciBERT	Beetle	217	~156 s	10/12	0.699
SciBERT	PRM800K	376	~281 s	3/12	0.341

3.3. Datasets used in this study

** SciEntsBank and Beetle*

SciEntsBank and Beetle are the two primary datasets comprising the SemEval-2013 Task 7 benchmark [1]. Systems are evaluated across three generalization scenarios: (1) Unseen Answers (UA), comprising new responses to questions included in the training set; (2) Unseen Questions (UQ), comprising new questions within a domain present in training; and (3) Unseen Domains (UD, evaluated on SciEntsBank only), representing the most challenging evaluation setting.

The full SciEntsBank corpus [21] contains approximately 10,000 responses to 197 questions across 15 science domains for Grades 3–6 students; however, the training split

utilized in this study corresponds to the subset summarized in Table 1 (4,969 responses, 135 questions, and 12 domains). Therefore, these two datasets are not directly equivalent. Beetle [22] originates from the BEETLE II intelligent tutoring system [23], which focuses on introductory electricity and electronics concepts (comprising over 5,000 responses across 56 questions). Both datasets originally employ a five-way classification scheme, which was collapsed in this work into a binary correct/incorrect classification task, following the benchmark’s official two-way evaluation track.

*** *PRM800K - Step-level math-solution correctness***

To evaluate model performance on notation-dense pure mathematics - a domain not covered by SciEntsBank and Beetle - this study incorporates PRM800K [12], a publicly available (MIT-licensed) dataset derived from the MATH benchmark [11], which contains LaTeX-dense competition problems across algebra, geometry, number theory, and precalculus. The dataset’s step-level annotations (labels: +1 for correct, -1 for incorrect, 0 for neutral) provide an independent evaluation of each step, offering a fine-grained assessment aligned with the hypothesis outlined in Section 3.1, rather than a single holistic score assigned to a solution.

This study reformulates the step-level evaluation task in PRM800K as a binary classification problem. Specifically, we process the first 20,000 solutions, filtering steps annotated with rating $\in \{-1, +1\}$ (discarding neutral steps with rating = 0 discarded) to form (problem, step, label) instances. These instances are then downsampled to a maximum of 8,000 samples to align with the scale of SciEntsBank/Beetle under budget constraints. For the feedback generation branch, the ground truth solution serves as the target text for the fine-tuning FLAN-T5 base. The dataset is split into training (85%) and validation sets (15%) using GroupShuffleSplit grouped by problem text, ensuring that all reasoning steps corresponding to a single problem remain strictly within the same split, thereby preventing data leakage during evaluation.

*** *Cross-dataset comparison and roles***

Table 1 summarizes the core characteristics of all three datasets. SciEntsBank and Beetle provide natural-science and engineering context across three standardized generalization levels (UA, UQ, and; whereas PRM800K adds a notation-dense pure mathematics benchmark. This setup allows for a unified experimental evaluation of MathBERT across increasing levels of mathematical notational density. Collectively, these three datasets yield approximately 20,000 grading instances, with the correct label constituting the majority class across all datasets.

3.4. Training and evaluation procedure

The grading branch fine-tunes MathBERT (tbs17/MathBERT) separately on each dataset’s training split using a class-weighted cross-entropy loss (weight $w_c = N / 2n_c$ for class c ; focal loss [24] was considered but ultimately omitted). All experimental runs share a uniform hyperparameter configuration: the AdamW [25], a learning rate of $2e-5$, a linear schedule with a 10% warmup ratio, a batch size of 16, a maximum sequence length of 256 tokens, and a gradient clipping threshold of 1.0. Models are trained for a maximum of 12 epochs with an early stopping patience of 4 epochs, and the checkpoint that achieves the highest validation Matthews Correlation Coefficient (MCC). The learning rate corresponds to the lowest value within the standard BERT range $\{2e-5, 3e-5, 5e-5\}$ [2]

and is held fixed across all datasets to prevent test-set information from influencing model selection. The generation branch fine-tunes FLAN-T5-base on PRM800K using AdamW with a learning rate of $3e-4$ and an effective batch size of 16. Model performance is evaluated using ROUGE-1, ROUGE-2, and ROUGE-Lsum metrics [26].

All three datasets are processed via a standardized pipeline using modular code execution. A validation split comprising 12% of each training set is held out for model selection, ensuring that the test set remains unexposed throughout training. A single held-out validation split is maintained because the SemEval-2013 benchmark specifies fixed partitions for the UA, UQ, and UD settings. The 12-epoch upper bound exceeds the point at which early stopping was triggered during preliminary experiments; thus, total training duration is determined dynamically by early stopping, with the optimal checkpoint evaluated on the test set based on the highest validation MCC.

To support the experimental analyses presented in Section 4, three additional protocols were implemented.

(i) Replication Protocol: Every encoder-dataset combination is retrained across three distinct random seeds ($\{42, 1337, 2024\}$), while holding all other hyperparameters constant. Results in Table 3 are reported as the mean $\pm$ standard deviation.

(ii) Zero-Shot Baseline: To benchmark encoder performance against prompting strategies [17] under identical data conditions. Qwen2.5-1.5B-Instruct [27], an instruction-tuned decoder-only large language model (LLM) of comparable deployable scale, is evaluated in a zero-shot setting without task-specific training. The model is prompted to assign a correctness verdict on a sample of 400 items per split and to generate explanations for 200 held-out PRM800K problems.

(iii) End-to-End System Evaluation: the grading and feedback generation branches are integrated sequentially: every reasoning step identified as incorrect by the grading module is routed directly to the generation module. End-to-end performance is evaluated using grouping precision and recall, alongside ROUGE metrics computed over correctly routed steps.

3.5. Proposed evaluation metrics

Cohen's Kappa [28] measures inter-annotator or model-target agreement by adjusting for chance agreement, and its quadratically weighted variant (QWK) is standard for ordinal scoring. However, on binary classification tasks, quadratic weights degenerate mathematically, causing QWK to coincide exactly with unweighted Kappa and providing no additional information beyond the Matthews Correlation Coefficient (MCC). Consequently, QWK is omitted from this study and would become relevant only if the ASAG task were extended to ordinal or multi-level scoring scales.

The Matthews Correlation Coefficient (MCC), computed directly from all four quadrants of the confusion matrix [29] and bounded within $[-1, 1]$, provides a more stable evaluation metric than $F1$ -score or Accuracy under severe class imbalance, where high overall accuracy can obscure poor minority-class performance. On PRM800K, MathBERT's Macro- $F1$ score (0.599) aligns closely with Accuracy while maintaining a positive MCC value (0.245), thereby preventing majority-class bias. Given its mathematical stability under class-imbalanced distributions, MCC was selected as the primary early-stopping criterion for all experimental training run.

4. Results and discussion

4.1. Full training results (measured) on all three datasets

Table 2 summarizes the training computational cost and the optimal epochs determined by early stopping for each encoder-dataset pair at random seed 42. Across all 27 runs, the selected optimal epoch ranges from 2 to 12 (out of a maximum of 12 epochs), demonstrating that the predefined epoch upper bound rarely constrains model convergence. Subsequently, Table 3 presents the corresponding test-set metrics evaluated at each selected optimal epoch.

Because a single fine-tuned model cannot isolate the specific impact of the pretraining corpus from the underlying training configuration, the experimental protocol was executed identically across three Transformer encoders featuring an identical backbone architecture, varying solely in their pretraining initializations: MathBERT (mathematics-specific pretraining), SciBERT (general scientific domain specialization), and BERT-base (general-purpose reference).

Table 3. Macro-F1 and MCC (mean $\pm$ standard deviation over seeds) of three fine-tuned encoders under one identical protocol, a zero-shot LLM, and the SemEval-2013 two-way systems [1]

Dataset	Split	n	MathBERT	BERT-base	SciBERT	Qwen 0-shot	med. [1]	best [1]
SciEntsBank	UA	540	0.726 ± 0.018 0.455 ± 0.037	0.764 ± 0.011 0.528 ± 0.022	0.737 ± 0.008 0.476 ± 0.017	0.617 0.303	0.705	0.768
SciEntsBank	UQ	733	0.555 ± 0.009 0.124 ± 0.009	0.552 ± 0.004 0.111 ± 0.005	0.536 ± 0.008 0.097 ± 0.026	0.669 0.353	0.629	0.737
SciEntsBank	UD	4,562	0.508 ± 0.039 0.078 ± 0.058	0.542 ± 0.017 0.129 ± 0.029	0.546 ± 0.037 0.149 ± 0.042	0.651 0.380	0.666	0.705
Beetle	UA	439	0.827 ± 0.012 0.662 ± 0.014	0.851 ± 0.007 0.706 ± 0.013	0.864 ± 0.010 0.729 ± 0.020	0.535 0.112	0.778	0.833
Beetle	UQ	819	0.685 ± 0.015 0.372 ± 0.029	0.696 ± 0.029 0.394 ± 0.057	0.720 ± 0.010 0.441 ± 0.019	0.517 0.057	0.666	0.720
PRM800K	held-out	1,179	0.599 ± 0.020 0.245 ± 0.018	0.634 ± 0.014 0.286 ± 0.025	0.619 ± 0.031 0.285 ± 0.033	0.439 -0.106	—	—

Note: In Table 3, the upper value in each cell reports the macro-F1 score, while the lower value indicates the Matthews Correlation Coefficient (MCC). Values following $\pm$ symbols represent the standard deviation calculated across three distinct random seeds {42, 1337, 2024} (comprising 27 fine-tuning runs in total). The “Qwen 0-shot” baseline denotes the performance of the Qwen2.5-1.5B-Instruct model evaluated zero-shot on 400 stratified sampled instances per data split, because this model was evaluated without task-specific fine-tuning; seed-based variance is not applicable. The final two columns report the median and best macro-F1 scores achieved by the nine original SemEval-2013 Task 7 two-way classification systems [1]. Note that the PRM800K dataset was not included in the original SemEval 2013 benchmark suite.

Table 3 presents the central empirical findings of this study. Averaged across random seeds under a unified experimental protocol, MathBERT ranks lowest among the three Transformer encoders on five out of six splits and is ahead on the sixth by only 0.003 in

macro-F1 score (SciEntsBank UQ, 0.555 against 0.552 for SciBERT) -a margin smaller than its inherent run-to-run variance, effectively representing a statistical tie. Conversely, SciBERT achieves the highest performance on three splits (SciEntsBank UD 0.546, Beetle UA 0.864, Beetle UQ 0.720), whereas BERT-base leads on the remaining two (SciEntsBank UA 0.764, PRM800K 0.634). The performance deficits observed for MathBERT range consistently from 0.034 to 0.039 in macro-F1 score across highly heterogeneous datasets.

The reported standard deviations in Table 3 make it possible to say which of these performance deficits are larger than the seed noise. The seed-to-seed variance ranges from 0.009 to 0.039 in macro-*F1* score for MathBERT, 0.004-0.029 for BERT-base, and 0.008-0.037 for SciBERT. When measured relative to the pooled standard deviation between the models under comparison, MathBERT's deficit corresponds to 3.5 standard deviations on Beetle Unseen Answers, 2.7 on Beetle Unseen Questions, 2.6 on SciEntsBank Unseen Answers, and 2.0 on PRM800K, so four of the five deficits exceed two pooled standard deviations of the seed noise. Only on SciEntsBank Unseen Domains does the 0.039 gap represent a 1.0 standard deviation margin, which coincides with the split exhibiting the high cross-seed instability across all encoders. Furthermore, re-evaluating the training process under random seed 42 yielded variance between 0.002 and 0.020 in macro-*F1* score, confirming that fixing the random seed does not fully eliminate run-to-run variability.

Table 3 additionally incorporates the median and maximum macro-*F1* scores from the nine original SemEval-2013 Task 7 two-way classification systems [1] (which relied on feature-engineered, pre-Transformer pipelines). On the Beetle dataset, the fine-tuned pretrained encoders demonstrate competitive performance: SciBERT reaches a macro-*F1* score of 0.864 on Unseen Answers – exceeding the 2013 benchmark maximum of 0.833- and achieves 0.720 on Unseen Questions, matching the 2013 benchmark best score. On SciEntsBank, however, this performance advantage reverses when evaluating unseen questions: the top-performing encoder achieves macro-*F1* scores of 0.555 on Unseen Questions and 0.546 on Unseen Domains, compared to the 2013 median scores of 0.629 and 0.666, respectively. On SciEntsBank, therefore, the 2013 feature-engineered pipelines still generalize further than any of the three encoders, whatever corpus each was pretrained on.

The final column of Table 3 introduces a zero-shot baseline using Qwen2.5-1.5B-Instruct, which exhibits a notable crossover effect rather than uniform underperformance. On splits where question contexts were exposed during training, fine-tuned encoders are far ahead of the zero-shot LLM (e.g., Beetle UA 0.864 vs. 0.535, SciEntsBank UA 0.764 vs. 0.617). Conversely, on splits where fine-tuned encoders experience performance degradation due to unseen questions, the untuned LLM achieves superior generalization: scoring 0.669 vs. 0.555 on Unseen Questions and 0.651 vs. 0.546 on Unseen Domains -- both exceeding the 2013 benchmark median. This demonstrates that task-specific fine-tuning optimizes question-specific fitting rather than broader domain generalization. The notable exception occurs on PRM800K, where the LLM achieves an overall Accuracy of 69.25% but yields a macro-*F1* score of 0.439 and a negative MCC (-0.106). In this setting, the LLM degenerates by predicting the majority class for nearly all instances, resulting in agreement worse than chance - a diagnostic failure mode effectively exposed by the MCC metric (Section 3.5).

4.2. Discussion of measured PRM800K results

The PRM800K dataset serves as the critical testbed where the initial hypothesis outlined in Section 3.1 yields its most direct prediction: input sequences rich in mathematical notation should inherently favor an encoder specialized through domain-specific pre-training on mathematical text. However, our empirical findings directly contradict this expectation. Across three random seeds, MathBERT underperforms relative to the other two encoder baselines (achieving a macro- $F1$ score of 0.599 ± 0.020 , an MCC 0.245 ± 0.018), trailing behind both BERT-base (0.634 ± 0.014 / 0.286 ± 0.025) and SciBERT (0.619 ± 0.031 / 0.285 ± 0.033). Models pretrained on general web prose and broader scientific literature consistently evaluate mathematical step correctness more effectively than an architecture tailored specifically to mathematics-education corpora. Crucially, MathBERT ranks last across every individual experimental run, with its 0.034 macro- $F1$ deficit relative to BERT-base spanning two pooled standard deviations.

Furthermore, PRM800K yields the lowest validation MCC scores across all evaluated models (0.272-0.355). Models trigger early stopping within two to six training epochs – a pattern consistent with filtering out neutral-rated step annotations, which effectively restricts training to approximately 20,000 out of about 75,000 available solution sequences. This behavior indicates that the observed performance bottleneck stems from data under-utilization rather than being intrinsically linked to MathBERT’s representation or dataset-specific characteristics. Re-incorporating the discarded neutral annotations into a three-class formulation presents the most direct avenue for improving MCC performance.

Table 4. Feedback-generation results on PRM800K (ROUGE, held-out test set)

System	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
FLAN-T5-base, fine-tuned	21.84	8.32	18.75	18.82
Qwen2.5-1.5B-Instruct, zero-shot	36.68	14.20	23.82	30.73

Because no published baselines currently exist for step-level mathematical feedback generation on the PRM800K dataset, the zero-shot baseline serves as the primary point of comparison. For evaluation, the zero-shot baseline was scored on a stratified sample of 200 sampled held-out problems, whereas the fine-tuned models were evaluated across the complete split.

4.3. Best-suited grade levels and subjects

These mappings are inferred from the source material rather than empirically validated - as there are no Vietnamese datasets or classroom studies, this serves as a hypothesis rather than a definitive finding. SciEntsBank aligns with primary and early-secondary levels, Beetle corresponds to upper-secondary Physics or introductory university Electrical Engineering, and PRM800K competition-prep mathematics.

4.4. Key contributions relative to prior work

To address the research gaps identified in Section 2, this study provides four principal contributions:

(1) Controlled Combined Comparison: We perform a rigorous, three-seed comparative evaluation of MathBERT, SciBERT, and BERT-base across the official

SemEval-2013 benchmark splits within a unified training pipeline – a controlled setup absent in prior studies such as [20]. (ii) Extension to Notation-Dense Mathematics: We extend this standardized protocol to step-level mathematical reasoning using the LaTeX-dense dataset, demonstrating that the relative performance ranking among the three encoders remains unchanged in LaTeX-dense context. (iii) LLM baseline and End-to-End Evaluation: We establish a zero-shot decoder LLM baseline across both task branches and conduct a holistic end-to-end evaluation of the integrated dual backbone system; (iv) Class-Imbalance Diagnostic Metric: We adopt the Matthews Correlation Coefficient (MCC) as an imbalance-robust model selection metric, addressing diagnostic vulnerabilities caused by severe class imbalance during early stopping.

4.5. Feedback-generation quality

On PRM800K's held-out problems, fine-tuned FLAN-T5-base (248M) achieves the ROUGE scores in Table 4 against ground-truth references, alongside the zero-shot LLM baseline (Section 3.4).

Although these automated metrics confirm lexical overlap, they do not guarantee reasoning correctness or pedagogical quality. Qualitative analysis shows that FLAN-T5 base often repeats prompt fragments verbatim, serving as an empirical baseline rather than a deployable feedback generator. Notably, zero-shot Qwen2.5-1.5B-Instruct outperforms fine-tuned FLAN-T5-base across all ROUGE variants (ROUGE-L 23.82 vs. 18.75, ROUGE-Lsum 30.73 vs. 18.82) without task-specific training. Because ROUGE rewards surface-level textual similarity rather than instructional validity, these metrics reflect surface similarity only.

4.6. End-to-end evaluation of the composed pipeline

Because reporting the pipeline branches independently obscures the error compounding inherent to the deployed system, the two models were evaluated end-to-end over the complete PRM800K held-out split (1,179 steps). The integrated pipeline routes 460 steps for feedback generation, of which 152 are genuinely incorrect — yielding a Precision of 0.330, Recall of 0.591, *F1*-score of 0.424. Consequently, 308 of the 460 generated explanations (67%) would be delivered to students whose steps were actually correct, whereas 105 genuinely incorrect steps receive no feedback. On correctly routed steps, the generator achieves ROUGE-1/2/L/LSum scores of 27.36, 8.01, 18.59, and 22.00, respectively. Given this two-thirds false-alarm rate, the practical deployment of this pipeline requires a human-in-the-loop framework, confirming that grading module precision - rather than feedback generation quality - serves as the primary system bottleneck.

5. Conclusions

This study fine-tuned specialized models under a unified experimental protocol to systematically evaluate the foundational assumption supporting domain-specific pretraining in mathematical reasoning. To the best of our knowledge, this work presents the first empirical benchmark of MathBERT on the official SemEval-2013 Task 7 splits, comparing its performance directly against a general-purpose (BERT-base) and scientific (SciBERT) encoders under identical training conditions. Empirical results do not support the assumed benefit of domain-specific pre-training: averaged across three random seeds over 27 training runs, BERT-base and SciBERT outperform MathBERT on five of six

benchmark splits while matching its performance on the sixth. Crucially, four of these five performance deficits exceed two standard deviations of seed-level variance. Furthermore, a zero-shot 1.5B-parameter decoder LLM, still small enough to deploy, generalizes better, outperforming all fine-tuned encoders on unseen problem domains, and surpassing the fine-tuned feedback generator across all four ROUGE metrics. Finally, when evaluated as an integrated end-to-end system, the pipeline misroutes two-thirds of its feedback explanations to student steps that were already correct — highlighting that classifier precision rather than generative capability constitutes the primary constraint for real-world deployment.

Several key limitations of this study should be acknowledged:

(i) Out-of-Domain Performance Collapse: All fine-tuned Transformer encoders exhibit severe performance degradation when evaluated on unseen question contexts—macro-*F1* scores decline from 0.726-0.764 on SciEntsBank Unseen Answers to 0.508-0.555 on Unseen Questions and Unseen Domains, falling below the median benchmark established in 2013.

(ii) Coarse Variance Estimation: While evaluating across three random seeds bounds experimental noise, it yields only a coarse estimate of variance, leaving performance gaps on one split (SciEntsBank Unseen Domains) within the standard deviation margin.

(iii) Run-to-Run Non-Determinism: Re-evaluating experiments under a fixed random seed (seed 42) resulted in macro-*F1* fluctuations of up to 0.020, confirming that fixing the random seed does not by itself make a run reproducible.

(iv) Scope of the Zero-Shot Baseline: The zero-shot-baseline evaluation was constrained to a single instruction-tuned model architecture utilizing a single prompt template on 400 sampled instances per data split;

(v) Surrogate NLG Evaluation Metrics: Automated ROUGE metrics measure surface-level n-gram overlap rather than conceptual accuracy or pedagogical value. Future work will focus on five key directions: (i) expanding evaluation runs to five random seeds per split; (ii) conducting expert human evaluations of feedback quality; (iii) improving grading precision, which is the binding constraint on the pipeline; (iv) reformulating PRM800K step evaluation into a three-class task by incorporating neutral ratings; and (v) constructing localized Vietnamese datasets paired with real-world classroom pilot studies.

Note for contributor:

- Short bio: Nguyen Nguyen Nam is a researcher at the University of Transport and Communications, Vietnam. Pham Dinh Phong is the Head of the Department of Computer Science at the University of Transport and Communications, Vietnam.

- Author's contributions: 1st author: conceptualization, methodology, software, experiments, writing - original draft; 2nd author: supervision, validation, writing - review & editing.

Conflict of interest. The authors declare no conflict of interest

REFERENCES

- [1] M. O. Dzikovska, R. D. Nielsen, C. Brew & et al., “SemEval-2013 Task 7: The joint student response analysis and 8th recognizing textual entailment challenge”,

- Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM 2013)*, Atlanta, GA, USA, pp. 263-274, 2013.
- [2] J. Devlin, M. W. Chang, K. Lee & K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", *Proceedings of NAACL-HLT 2019*, Minneapolis, MN, USA, pp. 4171-4186, 2019.
 - [3] OpenAI, "GPT-4 technical report", arXiv:2303.08774, 2023.
 - [4] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix & et al., "LLaMA: Open and efficient foundation language models", arXiv:2302.13971, 2023.
 - [5] A. Vaswani, N. Shazeer, N. Parmar & et al., "Attention is all you need", *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, Long Beach, CA, USA, pp. 5998-6008, 2017.
 - [6] I. Beltagy, K. Lo & A. Cohan, "SciBERT: A pretrained language model for scientific text", *Proceedings of EMNLP-IJCNLP 2019*, Hong Kong, China, pp. 3615-3620, 2019.
 - [7] J. Lee, W. Yoon, S. Kim & et al., "BioBERT: a pre-trained biomedical language representation model for biomedical text mining", *Bioinformatics*, vol. 36, no. 4, pp. 1234-1240, 2020.
 - [8] J. T. Shen, M. Yamashita, E. Prihar & et al., "MathBERT: A pre-trained language model for general NLP tasks in mathematics education", presented at *the NeurIPS 2021 MATH4ED Workshop*, 2021.
 - [9] C. Raffel, N. Shazeer, A. Roberts & et al., "Exploring the limits of transfer learning with a unified text-to-text transformer", *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1-67, 2020.
 - [10] H. W. Chung, L. Hou, S. Longpre & et al., "Scaling instruction-finetuned language models", *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1-53, 2024.
 - [11] D. Hendrycks, C. Burns, S. Kadavath & et al., "Measuring mathematical problem solving with the MATH dataset", *Proceedings of the NeurIPS 2021 Datasets and Benchmarks Track*, 2021.
 - [12] H. Lightman, V. Kosaraju, Y. Burda & et al., "Let's verify step by step", *Proceedings of the International Conference on Learning Representations (ICLR)*, Vienna, Austria, 2024. github.com/openai/prm800k
 - [13] C. Sung, T. Dhamecha, S. Saha & et al., "Pre-training BERT on domain resources for short answer grading", *Proceedings of EMNLP-IJCNLP 2019*, Hong Kong, China, pp. 6071-6075, 2019.
 - [14] L. Camus & A. Filighera, "Investigating transformers for automatic short answer grading", *Artificial Intelligence in Education (AIED 2020)*, Part II, Ifrane, Morocco, pp. 43-48, 2020.
 - [15] N. Reimers & I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks", *Proceedings of Conference on Empirical Methods in Natural Language Processing - International Joint Conference on Natural Language Processing*, Hong Kong, China, pp. 3982-3992, 2019.
 - [16] S. Bonthu, S. R. Sree & M. H. M. K. Prasad, "Automated short answer grading using deep learning: a survey", *Machine Learning and Knowledge Extraction (CD-MAKE 2021)*, pp. 61-78, 2021.

- [17] R. Ferreira Mello, C. Pereira Junior, L. Rodrigues & et al., “Automatic short answer grading in the LLM era: does GPT-4 with prompt engineering beat traditional models?”, *Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK 2025)*, Dublin, Ireland, 2025.
- [18] P. Lai, K. Zhang, Y. Lin & et al., “SAS-Bench: A fine-grained benchmark for evaluating short answer scoring with large language models”, arXiv:2505.07247, 2025.
- [19] L. Cong, L. Hammerla, S. Hahn, S. Gombert, H. Drachsler & U. Kroehne, “Automatic short answer grading with LLMs: From memorization to reasoning”, *Proceedings of the 16th International Learning Analytics and Knowledge Conference (LAK 2026)*, Bergen, Norway, 2026.
- [20] M. Zhang, S. Baral, N. Heffernan & A. Lan, “Automatic short math answer grading via in-context meta-learning”, *Proceedings of the 15th International Conference on Educational Data Mining (EDM 2022)*, Durham, UK, pp. 102-113, 2022.
- [21] R. D. Nielsen, W. Ward, J. H. Martin & M. Palmer, “Annotating students' understanding of science concepts”, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, 2008. Release used: huggingface.co/datasets/nkazi/SciEntsBank
- [22] M. O. Dzikovska, D. Bental, J. D. Moore, N. B. Steinhauser & G. E. Campbell, “Intelligent tutoring with natural language support in the BEETLE II system”, *Proceedings of the Fifth European Conference on Technology Enhanced Learning (EC-TEL)*, Barcelona, Spain, pp. 620-625, 2010. Release used: huggingface.co/datasets/nkazi/Beetle
- [23] M. O. Dzikovska, N. Steinhauser, E. Farrow & et al., “BEETLE II: Deep natural language understanding and automatic feedback generation for intelligent tutoring in basic electricity and electronics”, *International Journal of Artificial Intelligence in Education*, vol. 24, no. 3, pp. 284-332, 2014.
- [24] T.-Y. Lin, P. Goyal, R. Girshick, K. He & P. Dollár, “Focal loss for dense object detection”, *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, pp. 2980-2988, 2017.
- [25] I. Loshchilov & F. Hutter, “Decoupled weight decay regularization”, *Proceedings of the International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA, 2019.
- [26] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries”, *Text Summarization Branches Out (ACL Workshop)*, Barcelona, Spain, pp. 74-81, 2004.
- [27] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu & et al., “Qwen2.5 technical report”, arXiv:2412.15115, 2024.
- [28] J. Cohen, “Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit”, *Psychological Bulletin*, vol. 70, no. 4, pp. 213-220, 1968.
- [29] B. W. Matthews, “Comparison of the predicted and observed secondary structure of T4 phage lysozyme”, *Biochimica et Biophysica Acta*, vol. 405, no. 2, pp. 442-451, 1975.