Domain-Adaptive Pre-training for Automated Short Answer Grading in Conceptual Physics: Reliability, Question-Level Analysis, and Error Reduction

Shirin Deochand Lade shirin.lade@open.ac.uk
Alistair Willis alistair.willis@open.ac.uk
Jonathan Nylk jonathan.nylk@open.ac.uk
Oli Howson oli.howson@open.ac.uk
The Open University
United Kingdom

Abstract

Automated short answer grading (ASAG) has attracted increasing research attention as a potential tool to support teachers in evaluating student responses. However, the usefulness of such systems depends on how reliably they operate under realistic classroom constraints, where only limited numbers of graded responses may be available for each question. This study investigates automated grading of conceptual physics explanations from free-response questions derived from the Force Concept Inventory (FCI), a diagnostic tool for assessing understanding of mechanics concepts.

We compare a standard BERT-based grading model with a version further adapted through subject-specific pre-training on physics textbook materials. Building on prior work showing that BERT representations for ASAG can be improved through continued pre-training on domain resources, we do not propose a new adaptation method; instead, we evaluate its reliability in a classroom-derived conceptual physics setting. The models are evaluated across varying training set sizes to examine how domain adaptation affects performance when labelled data is limited. Our results show that domain-adaptive pre-training provides the largest gains in low-data conditions, improving classification accuracy and reducing grading errors. As more labelled responses become available, the difference between the models decreases. This suggests that subject-specific pre-training primarily improves data efficiency rather than the best achievable performance.

Error analysis shows that the performance gains come from real reductions in grading mistakes. When comparing both models, the improved model corrects substantially more baseline errors than it introduces new ones, with a fixed-to-broken ratio of 4.61, indicating a clear net reduction in grading errors. This suggests that the improvement primarily reflects net error reduction rather than simply shifting mistakes. However, improvements vary across individual assessment questions, highlighting the importance of question-level evaluation when deploying automated grading systems in classroom environments. These findings indicate that, relative to a general BERT baseline, textbook-based continued pre-training can support educational assessment.

1 Introduction

Automated Short Answer Grading (ASAG) systems are increasingly explored as tools for supporting student assessment by reducing marking workload and enabling timely feedback to students (Seneviratne and Manathunga, 2025). In real educational settings, however, automated grading systems operate under constraints that differ substantially from those typically considered in benchmark-driven experimental evaluations (Bonthu et al., 2021). When new questions are introduced or when assessments are administered to small cohorts, teachers often have access to only limited numbers of previously graded responses for each assessment question (Egaña et al., 2023). Under such conditions, automated grading systems must operate reliably with minimal task-specific supervision while still producing grading decisions consistent with human evaluation (Morris et al., 2025).

Several approaches have been proposed to improve model performance under limited supervision, including data augmentation, semi-supervised learning, and transfer learning (Bonthu et al., 2021; Burrows et al., 2015). In this work, we focus on domain-adaptive pre-training (DAPT) (Beltagy et al., 2019; Lee et al., 2020) where a pre-trained language model is further trained on domain-relevant text before task-specific fine-tuning (Gururangan et al., 2020). DAPT is an established method, and its use for ASAG has direct precedent: Sung et al. (2019) showed that BERT for short

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 635–650

answer grading can be improved by updating the pre-trained model using domain resources such as textbooks and question-answer data before supervised fine-tuning. Our contribution is therefore not a new adaptation algorithm. Instead, we examine whether textbook-based DAPT improves grading reliability relative to a general BERT baseline in a classroom-derived conceptual physics setting, especially when only limited labelled responses are available per question. The evaluation focuses on binary grading of held-out responses to the same assessment questions, rather than partial-credit scoring or transfer to unseen questions. We further analyse whether improvements correspond to net reductions in grading errors, how gains vary across questions, and whether the adapted model provides better calibrated confidence estimates.

In educational assessment, grading performance must be evaluated not only by predictive accuracy but also by reliability, as automated systems support decisions that influence feedback and student evaluation (Holmes et al., 2021). When comparing automated grading models, it is important to determine whether improved performance reflects a net reduction in grading errors, rather than a redistribution of errors across responses. Furthermore, grading reliability may vary across assessment questions due to differences in conceptual difficulty or response distributions, making question-level analysis important for evaluating automated grading systems in practice (Zhu et al., 2022).

This study investigates automated grading of conceptual physics explanations from free-response questions based on the FCI, a widely used diagnostic instrument in physics education (Hestenes et al., 1992; Hestenes and Halloun, 1995). Many ASAG studies rely on proprietary datasets, which can limit reproducibility and comparability across studies. In this work, we instead use FCI-based conceptual free-text questions, which are widely recognised and explicitly designed to probe common misconceptions in mechanics (Parker et al., 2023; Pathak et al., 2026). This established conceptual framework enables a more meaningful analysis of how domain-adaptive language modelling interacts with disciplinary reasoning tasks.

We address the following research questions:

- RQ1. When only limited numbers of graded student responses are available for the same assessment questions, does textbook-based DAPT improve binary ASAG reliability relative to a general BERT baseline?
- RQ2. When model performance improves, do these gains reflect consistent reductions in grading errors across questions, or do they arise from improvements in some questions offset by declines in others?

This study makes three contributions:

- We provide a reliability-focused evaluation of textbook-based DAPT for ASAG in a classroom-derived conceptual physics setting, using free-response questions derived from the Force Concept Inventory.
- We analyse how textbook-based DAPT changes grading behaviour relative to a general BERT baseline under limited labelled data by comparing learning curves, agreement metrics, calibration, and error-overlap patterns across 26 training-set sizes and 10 random seeds.
- We examine question-level variation in grading behaviour to identify whether improvements are consistent across conceptual physics questions or concentrated in particular items.

2 Related Work

2.1 Automated Short Answer Grading

Early ASAG systems relied on rule-based approaches, lexical overlap, or semantic similarity measures (Burrows et al., 2015). Subsequent work introduced supervised learning models trained on labelled student responses (Bonthu et al., 2021). More recent approaches use transformer-based models such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), which achieve strong performance on benchmark datasets (Takano et al., 2022). Recent work has also examined large language models for ASAG, with mixed evidence on whether few-shot LLM prompting outperforms fine-tuned encoder models in educational grading tasks (Grévisse et al., 2024; Kakarla et al., 2025). Despite these advances, it remains unclear how improvements in model architecture translate into practical grading behaviour, particularly whether they lead to more reliable and consistent decisions on real student responses.

However, Ley et al. (2023) identify remaining concerns regarding the reliability, cross-question consistency, and robustness of ASAG systems under limited training data in real educational environ-

ments. Many evaluations rely on curated datasets rather than authentic classroom data (Padó, 2022; Ferrara and Qunbar, 2022). Several benchmark datasets have been developed for ASAG research, including the SemEval Student Response Analysis dataset (Dzikovska et al., 2013) and the ASAP short-answer dataset (Mohler et al., 2011). Automated grading systems are generally proposed as decision-support tools to assist instructors by prioritising responses for review rather than replacing human judgement entirely (Weegar and Idestam-Almquist, 2024). In this context, overall benchmark accuracy is insufficient; evaluation must also consider reliability and error behaviour, including consistency across questions and whether models reduce rather than shift errors (Jung et al., 2025). Although instructors grade student responses during routine assessments, newly introduced questions often have few-or no-labelled responses available to train automated grading systems (Camus and Filighera, 2020; Zhang et al., 2022). The challenge of performing automated short-answer grading with only a small number of labelled student responses is compounded by the lack of systematic investigation into how ASAG systems behave across varying training set sizes, especially under realistic classroom constraints.

2.2 Domain-Adaptive Pre-training (DAPT)

Domain-adaptive pre-training (DAPT) extends the training of a language model by continuing pre-training on domain-relevant corpora prior to task-specific fine-tuning (Gururangan et al., 2020). This approach allows models to better capture specialised vocabulary and conceptual structures that may not be well represented in general web-based training data. DAPT has been shown to improve performance in specialised domains such as scientific text processing (Beltagy et al., 2019) and biomedical text analysis (Gu et al., 2021).

Sung et al. (2019) showed that BERT for ASAG can be improved by further pre-training on domain resources such as textbooks and question-answer pairs. More broadly, Gururangan et al. (2020) showed that DAPT improves performance across domains and resource settings. Our work therefore does not propose a new adaptation method; instead, it evaluates textbook-based DAPT in classroom-derived conceptual physics ASAG, focusing on low-data reliability, calibration, error overlap, and question-level variation. We do not compare against stronger science-domain encoders, physics-domain encoders, modern LLM graders, or same-size non-physics continued pre-training controls; therefore, our claims are limited to the benefit of textbook-based DAPT over a general BERT baseline, rather than superiority over alternative encoder, LLM, or corpus-control baselines.

Recently, Hellert et al. (2024) explored domain-adapted models for scientific and technical domains, including physics-related NLP tasks. Pre-training language models on educational physics materials may therefore enable models to better represent the conceptual language used in student explanations within Force Concept Inventory (FCI) (Hestenes et al., 1992) assessments (Zhu et al., 2022). However, studies such as Henkel et al. (2024) and Gao et al. (2024) focus primarily on performance, providing limited insight into how domain-adaptive pre-training affects automated grading behaviour.

2.3 Reliability, Agreement, and Question-Level Variation in ASAG

Reliability is a fundamental requirement in educational assessment and measurement theory (Nitko and Brookhart, 2014; Schneider et al., 2023). Agreement metrics such as Cohen's κ (Cohen, 1960) are commonly used to measure agreement between automated grading systems and human raters (del Gobbo et al., 2023). However, such metrics do not explain how errors are distributed, making it important to examine the types of mistakes systems make. Because grading decisions directly affect how students are assessed, improvements in model performance must be interpreted not only in terms of overall accuracy but also in terms of the types of errors made and the degree of agreement with human grading (Li et al., 2023). Two systems may achieve similar overall accuracy while producing different error patterns, and some grading errors may have more serious consequences for students than others.

Analysing error overlap (Dietterich, 1998) between models helps determine whether improved performance reflects a genuine reduction in grading mistakes. A model may appear more accurate because it correctly grades responses that were previously misclassified. However, it may also introduce new errors on other responses. As long as more errors are corrected than introduced, overall accuracy will increase, even if errors are redistributed across responses rather than uniformly reduced, which may lead to inconsistent or unfair grading

for certain students (Grévisse et al., 2024). In addition, grading performance may vary across assessment questions due to differences in conceptual difficulty or response distributions, meaning that some questions may be graded less reliably than others, potentially leading to unfair outcomes for students. Examining automated grading behaviour at the question level therefore provides important insight into system reliability, yet such analyses remain relatively limited in ASAG research.

2.4 Conceptual Physics Assessment and the Force Concept Inventory

This study focuses on conceptual physics assessment using free-text responses derived from the FCI (Parker et al., 2023; Pathak et al., 2026). The FCI is a widely used diagnostic instrument designed to measure students' understanding of fundamental Newtonian mechanics concepts and identify common misconceptions (Hestenes and Halloun, 1995; Hestenes et al., 1992). Yasuda et al. (2023) and Parker et al. (2022) explored short-answer reformulations of the Force Concept Inventory that capture students' reasoning more directly.

Grading conceptual explanations in free text presents particular challenges because students may express the same underlying idea using different wording, or provide responses that mix correct reasoning with misconceptions (Rainey et al., 2022; Dzikovska et al., 2013). This makes it difficult for automated grading systems to distinguish between fully correct, partially correct, and incorrect answers based on surface-level features alone. As a result, conceptual physics assessment provides a suitable setting for investigating whether DAPT on physics materials can improve grading performance by better capturing domain-specific terminology and response patterns (Trewartha et al., 2022; Sung et al., 2019). Furthermore, because the FCI targets specific conceptual dimensions, it enables fine-grained analysis of how automated grading performance varies across questions targeting different mechanics concepts.

3 Methodology and Experimental Setup

This section describes the task formulation, dataset construction, DAPT procedure, model configurations, and experimental design. The experiments are designed to evaluate the effect of DAPT while keeping the training data, model architecture, and evaluation procedure consistent across conditions.

3.1 Task and Dataset

Assessment Instrument: This study investigates the use of ASAG on free-response questions from the Newtonian Mechanics Quiz (NMQ) (Pathak et al., 2026), a modified version of the widely used Force Concept Inventory (FCI) (Hestenes et al., 1992) incorporating free-response questions (Parker et al., 2022).

Both the FCI and NMQ are conceptual diagnostic instruments designed to assess students' understanding and identify common misconceptions. The free-response questions integrated into the NMQ, requiring students to construct their own response, can provide deeper insight into students' comprehension of Newtonian mechanics than the multiple-choice questions of the original FCI alone (Pathak et al., 2026), Parker et al. (2022) and Rebello and Zollman (2004).

We use 15 free-response questions from the Newtonian Mechanics Quiz (NMQ), each designed to assess a specific conceptual aspect of Newtonian mechanics. For example, one question asks students to compare the time taken for two objects of different masses to reach the ground when dropped simultaneously, targeting understanding of gravitational acceleration and the independence of mass in free fall. These questions are conceptually aligned with Force Concept Inventory (FCI)-style items in that they probe underlying physical principles rather than requiring numerical calculation.

Student Response Dataset: The dataset analysed in this work consists of student responses collected during administrations of the NMQ at 6 UK higher education institutions in the 2024/25 academic year, either in class or as an out-of-class activity. A total of 674 quiz attempts were collected from students enrolled on introductory undergraduate physics modules. The unit of analysis for modelling is a question-response instance rather than a complete quiz attempt. Since each quiz attempt can contribute responses to multiple free-response questions, the 674 quiz attempts yield multiple labelled question-response instances across the 15 questions; the fixed test set of 1,566 therefore refers to held-out question-response instances, not 1,566 separate students or quiz attempts. Data collection was conducted via an online platform.

Each student response was independently labelled as correct or incorrect by four physics experts following agreed marking guidelines. Disagreements were resolved by a panel of three re-

searchers, who determined the final label. All responses were anonymised prior to analysis to protect student privacy. An anonymised version of the dataset has been released via The Open University repository with DOI 10.21954/ou.rd.32190003 (Pathak et al., 2026).

Split protocol: For each of the 15 questions, labelled response instances were split into a training pool and a fixed test set, with approximately 80% assigned to training and 20% to testing. All scaling subsets were sampled only from the training pool, and test responses were never used during training or subset construction. The split was performed at the question-response level rather than at the student or institution level; therefore, the evaluation measures generalisation to unseen responses for the same questions, not student-disjoint, institution-disjoint, or unseen-question generalisation.

Task Formulation: We study ASAG as a binary classification task. Given a question and a student's free-text response, the model predicts whether the response should be marked correct or incorrect. This formulation matches the available expert labels and enables controlled analysis of agreement and error behaviour, but it does not model partial-credit or rubric-based scoring.

Each instance consists of:

- the question text,
- a student-written response,
- a binary correctness label assigned by human graders.

Table 1 reports the distribution of correct and incorrect responses per question. Class distributions vary substantially across questions, reflecting realistic classroom assessment conditions in which some concepts are more difficult than others and certain misconceptions appear more frequently.

3.2 DAPT on LibreTexts Physics

Following prior work on continued pre-training for domain adaptation and ASAG, we apply textbook-based DAPT before task-specific fine-tuning (Sung et al., 2019; Gururangan et al., 2020). The goal is not to introduce a new pre-training objective, but to test whether exposure to undergraduate physics language improves grading reliability for conceptual physics responses under low-data conditions.

Pre-training Corpus: The domain-adapted model undergoes continued pre-training on six openly available physics textbooks from LibreTexts (Dourmashkin, 2020; Gea-Banacloche, 2019; Cline, 2019), covering foundational topics in Newtonian mechanics. LibreTexts was selected because it provides openly licensed, comprehensive undergraduate-level physics material that reflects terminology and conceptual explanations aligned with FCI-style content. Continued pre-training uses the masked language modelling objective (Devlin et al., 2019). This stage is entirely unsupervised and does not involve student responses or grading labels, ensuring that any downstream performance differences are not due to additional supervised grading labels.

3.3 Model Configurations

We compare two models that share identical architecture and task-specific training procedures, differing only in whether DAPT is applied.

3.3.1 Baseline Model

The baseline model consists of a BERT encoder (Devlin et al., 2019) fine-tuned directly on the short answer grading task.

For each instance:

- The question and student response are concatenated into a single input sequence.
- The final hidden state corresponding to the [CLS] token is passed to a linear classification layer.
- Binary predictions are produced using cross-entropy loss (Cui et al., 2019).

To simulate realistic low-resource classroom conditions, we adopt a **partial fine-tuning strategy** (Howard and Ruder, 2018; Sun et al., 2019):

- The lower eight transformer layers are frozen.
- Only the top four layers and the classification head are updated.

Fine-tuning instability in small datasets has been widely documented in prior work by Dodge et al. (2020); Mosbach et al. (2021); Tinn et al. (2023).

3.3.2 Domain-Adapted Model

The domain-adapted model follows a two-stage procedure:

- Continued pre-training on the LibreTexts physics corpus using MLM (Devlin et al., 2019).
- Fine-tuning on the ASAG task using the same architecture, layer-freezing configuration, and optimization settings as the baseline model.

Holding the base architecture, fine-tuning procedure, training subsets, and evaluation protocol constant isolates textbook-based continued pre-training relative to general BERT, but not whether

Question	Correct (%)	Incorrect (%)	N	Imb. Ratio
Q1	72.3	27.7	517	0.723
Q2	52.0	48.0	244	0.520
Q3	83.9	16.1	509	0.839
Q4	70.0	30.0	496	0.700
Q5	75.2	24.8	508	0.752
Q6	83.3	16.7	497	0.833
Q7	77.0	23.0	252	0.770
Q8	40.5	59.5	487	0.405
Q9	64.2	35.8	246	0.642
Q10	84.2	15.8	493	0.842
Q11	57.5	42.5	496	0.575
Q12	96.0	4.0	496	0.960
Q13	58.4	41.6	243	0.584
Q14	91.9	8.1	258	0.919
Q15	53.9	46.1	490	0.539

Table 1: Per-question class distribution. Percentages indicate the proportion of correct and incorrect responses.

gains arise from physics-specific content rather than continued pre-training more generally.

3.4 Training Procedure

Both models are trained on the Newtonian Mechanics dataset using supervised learning. As described in Section 3.1, the dataset consists of anonymised student responses collected during classroom administrations of free-response questions derived from the FCI. Because class distributions vary across questions (see Table 1), we use class-weighted cross-entropy loss (Cui et al., 2019), with weights computed based on inverse class frequency within each training subset. This mitigates bias toward majority classes and maintains sensitivity to less frequent response types.

All hyper-parameters, including optimizer configuration, learning rate schedule, batch size, number of epochs, and early stopping criteria, are kept identical across models. Given that small training sets can lead to unstable training dynamics, including divergence during the fine-tuning (Dodge et al., 2020), we apply early stopping (Song et al., 2020) and repeat each experimental condition using multiple random seeds.

3.5 Experimental Design: Controlled Scaling Study

We run a controlled scaling study over 26 nested training-set sizes to reflect classroom data availability, from early deployment to larger cohorts. After constructing the fixed per-question test set, all remaining responses form the training pool. For each question and training size d_x, we sample up to x correct and x incorrect responses from the training pool. Thus, d_x represents a target per-class sampling cap, not a guarantee that every question contains x examples from each class. When one class has fewer than x available responses, both classes are capped at the smaller available count. This is particularly relevant for highly imbalanced questions such as Q12, where the number of incorrect responses is small. Larger subsets are nested by adding newly sampled responses to the previous subset where available. We train both models with 10 random seeds, resulting in 26 sizes $\times$ 10 seeds $\times$ 2 models = 520 runs. The fixed test set contains 1,566 question-response instances.

The fixed test set enables paired comparisons between models. Experiment-level differences across matched training sizes and random seeds are evaluated using paired tests over run-level metrics, while fixed/broken error counts are reported descriptively because the same test responses are evaluated repeatedly across conditions. We also report correlation analyses relating training size and baseline question performance to observed DAPT gains. A summary of statistical procedures is provided in Appendix Table 5.

For each configuration, we report Accuracy, Macro F1, Balanced Accuracy, and Cohen's κ (Cohen, 1960) as mean $\pm$ standard deviation across runs, with detailed per-question results provided in Appendix Tables 6 and 7. Increasing the number of repetitions reduces variance in performance estimates and provides a more stable assessment of model behaviour in low-data settings, where single-run results can be highly variable (Dodge et al., 2020). This controlled scaling framework enables analysis of performance differences, data efficiency, saturation effects, and variability under limited labelled data.

4 Results and Discussion

This section evaluates whether DAPT improves the practical reliability of ASAG when applied to classroom-derived student responses, and whether observed improvements reflect consistent reductions in grading errors.

4.1 Overall Performance and Reliability (RQ1)

Across all evaluation metrics, the domain-adapted model (DAPT) shows higher agreement with reference labels compared to the baseline model.

Table 2 summarises the average performance

across all runs and training sizes.

Metric	Baseline	DAPT	Δ
Accuracy	0.906	0.963	+0.057
Macro F1	0.889	0.955	+0.066
Balanced Acc.	0.895	0.954	+0.059
Cohen's κ	0.783	0.910	+0.127

Table 2: Overall average performance across all runs and training sizes.

From a grading perspective, the increase in Cohen's κ suggests improved consistency in automated decisions, which is important for maintaining fairness in assessment.

Wilcoxon Signed-Rank Test (Experiment-Level Comparison): To evaluate whether these differences are consistent across experimental conditions, we conducted a paired Wilcoxon signed-rank test (Wilcoxon, 1945) on accuracy scores, comparing matched runs and training sizes between the two models.

The paired comparison showed a consistent difference across matched conditions:

- $W = 30.5, p < 10^{-43}$

Because training subsets are nested, this result should be interpreted as evidence of a strong repeated pattern across conditions rather than as fully independent experimental replications.

Classroom interpretation: From an assessment perspective, the increase in Cohen's *kappa* indicates stronger agreement between automated predictions and reference labels. This indicates that the domain-adapted model may provide more stable grading decisions under varying data conditions.

4.2 Effect of Training Data Size (RQ1)

Across 260 matched comparisons (10 runs × 26 training sizes, comparing the same run and dataset size across models), the domain-adapted model outperformed the baseline in the large majority of cases:

- Accuracy improved in 257/260 comparisons
- Macro F1 improved in 258/260 comparisons
- Cohen's κ improved in 258/260 comparisons

These results indicate that performance gains are observed across most experimental conditions, rather than being limited to a small subset of runs or training sizes.

Figure 1 presents learning curves across increasing training sizes. The largest performance differences occur in low-data settings:

- d_{10}: Δ Accuracy $\approx +0.28$
- d_{50}: Δ Accuracy $\approx +0.12$
- d_{470}: Δ Accuracy $\approx +0.006$

Figure 1 shows that DAPT provides the largest gains in low-data conditions, with an accuracy improvement of approximately 28 percentage points at d_{10} and 12 points at d_{50}. As training data increases, the gap narrows, reaching approximately 0.6 points at d_{470}. This pattern is supported by a strong negative correlation between training size and DAPT gain (Spearman's $\rho = -0.90$), indicating that the benefit of domain adaptation diminishes as more labelled data becomes available.

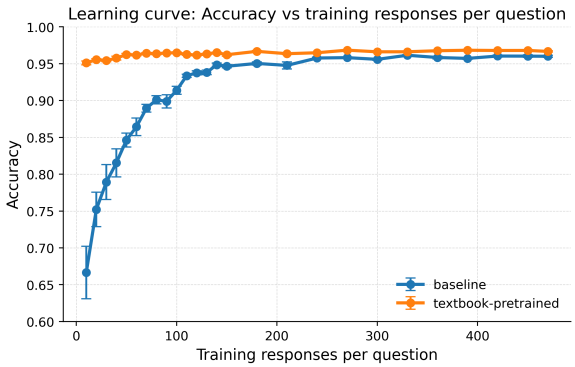

Figure 1: Learning curves across training sizes.

Classroom interpretation: From an assessment perspective, this suggests that domain-adaptive pretraining can provide more reliable grading in early deployment stages, where only a limited number of student responses are available, while offering smaller additional gains once sufficient data has been collected.

4.3 Question-Level Behaviour (RQ2)

While overall performance improves, the magnitude of improvement varies across questions. Larger gains are observed for questions where baseline performance is lower (e.g., Q4, Q5), whereas questions with already high baseline accuracy (e.g., Q1, Q14, Q15) show only marginal improvements.

A strong negative relationship between baseline accuracy and improvement (Spearman's rank correlation coefficient $\rho = -0.94$) indicates a ceiling effect, where questions that are already easy to grade offer limited room for further gains.

Figure 2 presents performance differences at the question level.

These results suggest that DAPT provides the greatest benefit for questions that are less reliably handled by the baseline. A qualitative inspection suggests that lower-gain questions may have more constrained answer spaces, whereas higher-gain questions appear to involve more varied explanations, incomplete causal reasoning, mixed correct and incorrect ideas, or domain-specific phrasing. Textbook-based DAPT may therefore help most where student responses are linguistically heterogeneous. This interpretation is exploratory, and future work should analyse response length, misconception type, and explanation structure systematically.

Classroom interpretation: From a classroom perspective, this indicates that the domain-adapted model may be particularly useful for supporting grading on more challenging questions, where automated decisions are likely to be less reliable. Importantly, we do not observe clear evidence of systematic performance degradation across questions, suggesting that improvements are not achieved at the expense of reduced performance elsewhere.

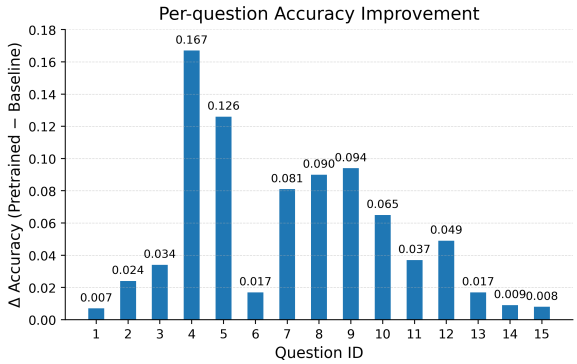

Figure 2: Question-level performance differences (δ accuracy) between baseline and domain-adapted models.

4.4 Error-Level Analysis (RQ2)

To determine whether improvements reflect genuine error reduction rather than redistribution, we conducted a global error-overlap analysis across all questions and training sizes.

In this analysis, *Fixed* refers to cases where the domain-adapted model produces a correct prediction while the baseline model is incorrect, and *Broken* refers to cases where the baseline is correct but the domain-adapted model is incorrect.

Category	Count
Both Correct	362,667
Both Wrong	8,487
Fixed	29,586
Broken	6,420
Net Reduction	23,166
Fixed/Broken Ratio	4.61

Table 3: Error overlap summary between baseline and textbook-pretrained models.

Table 3 shows that the domain-adapted model corrects substantially more baseline errors than it introduces, resulting in a net reduction of 23,166 errors and a Fixed/Broken ratio of 4.61.

Aggregate paired comparison: Because the same fixed test responses are evaluated repeatedly across training sizes and seeds, the pooled fixed/broken counts in Table 3 should be interpreted descriptively rather than as independent observations. The large imbalance between Fixed and Broken cases nevertheless shows a consistent directional pattern: across repeated paired evaluations, the domain-adapted model corrects substantially more baseline errors than it introduces.

4.5 Calibration and Confidence

Table 4 summarises calibration performance.

Metric	Baseline	DAPT
Brier score	0.0668	0.0308
ECE	0.0408	0.0228

Table 4: Calibration metrics for models used

The domain-adapted model achieves lower values on both the Brier score (Brier, 1950) and Expected Calibration Error (ECE) (Guo et al., 2017), suggesting that its predicted probabilities are better aligned with observed outcomes. In particular, the reduction in Brier score reflects improved overall probabilistic accuracy, while the lower ECE indicates that confidence estimates more closely match true correctness likelihoods.

Classroom interpretation: From a grading perspective, improved calibration is important because it allows the system's confidence scores to be interpreted as meaningful indicators of reliability. This enables practical workflows in which low-confidence predictions can be flagged for human review, while high-confidence predictions can be trusted with greater assurance.

4.6 Discussion

These results are consistent with prior work showing that continued pre-training on domain-relevant text can improve downstream performance. Sung et al. (2019) demonstrated this idea for ASAG using domain resources such as textbooks, while Gururangan et al. (2020) showed that DAPT can improve performance across multiple domains and resource settings. The present study extends this line of work by examining reliability, calibration, error overlap, and question-level behaviour in classroom-derived conceptual physics ASAG. Because we do not evaluate LLM graders or a same-size non-physics pre-training control, the results show improvement over general BERT in this setting, but not superiority over LLM-based grading or evidence that gains are uniquely caused by physics-specific content. The evaluation also reflects same-question deployment: models are tested on held-out responses to the same questions, not on unseen questions or unseen institutions.

4.6.1 RQ1: Reliability under limited data:

The results show that domain-adaptive pretraining improves agreement with reference labels across conditions, with the largest gains observed in low-data regimes. Learning curve analysis indicates that these improvements are most pronounced when only limited labelled responses are available, and diminish as training data increases.

4.6.2 RQ2: Nature of improvements:

Improvements are observed across most questions, with larger gains for items where baseline performance is lower. Error-level analysis further indicates that these gains primarily reflect net reductions in grading errors, rather than redistribution of errors across instances.

4.6.3 Implications for practice

From a teaching perspective, these findings suggest that domain-adapted models may:

- improve consistency of automated grading decisions,
- provide the greatest benefit in early deployment scenarios with limited data, and
- support identification of responses requiring human review.

However, such systems should be used to support, rather than replace, human judgement in assessment.

5 Conclusion

We evaluated whether textbook-based DAPT improves binary ASAG reliability relative to a general BERT baseline in a classroom-derived conceptual physics setting. Across 26 training sizes and 10 random seeds, the domain-adapted model showed higher accuracy, macro F1, balanced accuracy, and Cohen's κ, with the largest gains in low-data settings.

Question-level and error-overlap analyses showed that these gains were not only aggregate improvements: the domain-adapted model corrected substantially more baseline errors than it introduced, with larger benefits for questions where baseline performance was lower. Calibration also improved, suggesting that model confidence may better support human-review workflows.

These findings support textbook-based DAPT as a useful low-data strategy for same-question ASAG deployment. However, the system should be treated as a decision-support tool, and broader evaluation is needed for stronger baselines, partial-credit scoring, unseen questions, unseen institutions, and other educational contexts.

6 Limitations

Baseline scope: We compare general BERT with a textbook-adapted version of the same model to isolate the effect of continued pre-training on undergraduate physics materials. We do not compare against science-domain or physics-domain encoders such as SciBERT or PhysBERT, stronger general encoders such as RoBERTa or DeBERTa, modern LLM-based graders, or a same-size non-physics continued pre-training control. Therefore, our results show improvement over general BERT in this setting, but do not establish superiority over stronger baselines or prove that the gains are uniquely due to physics-specific content. Future work should test these comparisons directly.

This study is conducted using free-response conceptual physics questions derived from the Force Concept Inventory (FCI), a widely used diagnostic instrument designed to assess students' conceptual understanding of Newtonian mechanics (Hestenes et al., 1992; Hestenes and Halloun, 1995). This reflects a specific assessment context focused on conceptual reasoning rather than procedural or numerical problem-solving, and the findings should be interpreted within this context.

The grading task is formulated as binary clas-

sification (correct/incorrect), which enables controlled analysis of agreement and error patterns but does not capture partial-credit or rubric-based scoring. This limits direct applicability to classroom settings where teachers assign ordinal scores or apply detailed rubrics. The headline accuracy values should therefore not be directly compared with multi-class, ordinal, or rubric-based ASAG settings, which present a harder grading problem. Future work should examine whether the DAPT advantage observed here transfers to these more complex scoring settings.

Split and generalisation scope: The train/test split is performed at the question-response level. The fixed test set contains held-out responses to the same 15 questions used to construct the training pools, so the study evaluates new responses to existing questions rather than transfer to entirely unseen questions. Because the split is not student-disjoint or institution-disjoint, responses from the same student or institution may appear across training and test partitions through different question responses.

Although the experimental design includes multiple training sizes and repeated runs, the number of assessment questions is limited (n = 15), which constrains the strength of conclusions regarding question-level variation.

Finally, domain-adaptive pre-training is based on a single type of corpus (undergraduate physics textbooks from LibreTexts). Other domain sources may lead to different adaptation effects and remain for future investigation. The present analysis also does not establish that DAPT improves conceptual understanding in the model; gains may partly reflect better handling of physics vocabulary, common phrasings, or surface-level response patterns.

7 Ethics Statement

Collection of the Newtonian Mechanics Quiz student response dataset was approved by the relevant institutional Human Research Ethics Committee. Student responses were collected during authentic classroom assessments as part of ongoing doctoral research, and the anonymised dataset has been released via The Open University repository (Pathak et al., 2026). Participation complied with institutional ethical guidelines, and no personally identifiable information was retained. The ethical approval permits the use of anonymised data for research dissemination.

The physics textbooks used for domain-adaptive pre-training were obtained from *LibreTexts*, an open educational resource platform providing materials under open licenses, including Creative Commons, Public Domain, and similar licenses. These licenses permit reuse, adaptation, and redistribution of content for educational and research purposes, subject to specific conditions such as attribution and, in some cases, non-commercial use or share-alike requirements. All materials were accessed and used in accordance with their respective licensing terms, and appropriate attribution was maintained where required.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Sridevi Bonthu, Rama Sree Sripada, and M. H. M. Krishna Prasad. 2021. Automated short answer grading using deep learning: A survey. In *Machine Learning and Knowledge Extraction (CD-MAKE 2021)*, volume 12861 of *Lecture Notes in Computer Science*, pages 61–78. Springer.

Glenn W. Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25(1):60–117.

Leon Camus and Anna Filighera. 2020. Investigating transformers for automatic short answer grading. In *Artificial Intelligence in Education (AIED 2020)*, volume 12164 of *Lecture Notes in Computer Science*, pages 43–48. Springer.

Douglas Cline. 2019. Variational principles in classical mechanics. University of Rochester. Accessed February 2026.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, Long Beach, California, USA. IEEE.

Emiliano del Gobbo, Alfonso Guarino, Barbara Cafarelli, and Luca Grilli. 2023. GradeAid: a framework for automatic short answers grading in educational contexts—design, implementation and evaluation. *Knowledge and Information Systems*, 65:4295–4334.

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, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Thomas G. Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah A. Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *Preprint*, arXiv:2002.06305.

Peter Dourmashkin. 2020. Classical mechanics. Based on MIT OpenCourseWare materials. Accessed February 2026.

Myroslava O. Dzikovska, Rodney D. Nielsen, and Chris Brew. 2013. Semeval-2013 task 7: Student response analysis. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 263–274, Atlanta, Georgia, USA. Association for Computational Linguistics.

Aner Egaña, Itziar Aldabe, and Oier López de Lacalle. 2023. Exploration of annotation strategies for automatic short answer grading. In *Artificial Intelligence in Education (AIED 2023)*, volume 13916 of *Lecture Notes in Computer Science*, pages 377–388. Springer.

Steve Ferrara and Ann Qunbar. 2022. Validity arguments for AI-based automated scores. *Journal of Educational Measurement*, 59(3):273–294.

Yifan Gao, Carolyn Penstein Rosé, Mingming Fan, and Kenneth R. Koedinger. 2024. Automatic assessment of text-based responses in post-secondary education: A systematic review. *Computers and Education: Artificial Intelligence*, 6:100207.

Julio Gea-Banacloche. 2019. University physics i: Classical mechanics. Accessed February 2026.

Camille Grévisse, Martina Bientzle, Juan Cendan, and Niklas Kühl. 2024. LLM-based automatic short answer grading in undergraduate medical education. *BMC Medical Education*, 24:1026.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Intelligent Systems and Technology*, 12(2):1–23.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360. Association for Computational Linguistics.

Thorsten Hellert, João Montenegro, and Andrea Pollastro. 2024. PhysBERT: A text embedding model for physics scientific literature. *APL Machine Learning*, 2(4):046105.

Markus Henkel, Chris Kedzierski, and Iryna Gurevych. 2024. Can LLMs grade open response reading comprehension questions? an empirical study using the ROARs dataset. *International Journal of Artificial Intelligence in Education*.

David Hestenes and Ibrahim Halloun. 1995. Interpreting the force concept inventory. *The Physics Teacher*, 33(8):502–506.

David Hestenes, Malcolm Wells, and Gregg Swackhamer. 1992. Force concept inventory. *The Physics Teacher*, 30(3):141–158.

Wayne Holmes, Matthew Bialik, and Charles Fadel. 2021. Ethics of ai in education: Towards a community-wide framework. Technical report, UCL Knowledge Lab.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Ji Yoon Jung, Lillian Tyack, and Matthias von Davier. 2025. Towards the implementation of automated scoring in international large-scale assessments: Scalability and quality control. *Computers and Education: Artificial Intelligence*, 8:100375.

Sanjit Kakarla, Conrad Borchers, Danielle R. Thomas, Shambhavi Bhushan, and Kenneth R. Koedinger. 2025. Comparing few-shot prompting of gpt-4 llms with bert classifiers for open-response assessment in tutor equity training. In *Proceedings of the Innovation and Responsibility in AI-Supported Education Workshop*, volume 273 of *Proceedings of Machine Learning Research*, pages 133–140. PMLR.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Tobias Ley, Kairit Tammets, Gerti Pishtari, Pankaj Chejara, Reet Kasepalu, Mohammad Khalil, Merike Saar, Iris Tuvi, Terje Väljataga, and Barbara Wasson. 2023. Towards a partnership of teachers and intelligent learning technology: A systematic literature review of model-based learning analytics. *Journal of Computer Assisted Learning*, 39(5):1397–1417.

Thomas W. Li, Shiyan Hsu, Max Fowler, Zhikai Zhang, Craig Zilles, and Karrie Karahalios. 2023. Am i wrong, or is the autograder wrong? effects of AI grading mistakes on learning. In *Proceedings of the 2023 ACM Conference on International Computing Education Research (ICER 2023)*, pages 159–170, New York, NY, USA. ACM.

Michael Mohler, Razvan Bunescu, and Rada Mihalcea. 2011. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 752–762, Portland, Oregon, USA. Association for Computational Linguistics.

Wesley Morris, Langdon Holmes, Joon Suh Choi, and Scott A. Crossley. 2025. Automated scoring of constructed response items in math assessment using large language models. *International Journal of Artificial Intelligence in Education*, 35(2):559–586.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*.

Anthony J. Nitko and Susan M. Brookhart. 2014. *Educational Assessment of Students*. Pearson.

Ulrike Padó. 2022. Assessing the practical benefit of automated short-answer graders. In *Artificial Intelligence in Education (AIED 2022)*, volume 13355 of *Lecture Notes in Computer Science*, pages 212–224. Springer.

Mark A. J. Parker, Holly Hedgeland, Nicholas Braithwaite, and Sally Jordan. 2022. Student reaction to a modified force concept inventory: The impact of free-response questions and feedback. *European Journal of Science and Mathematics Education*, 10(3):310–323.

Mark A. J. Parker, Holly Hedgeland, Sally E. Jordan, and Nicholas St. J. Braithwaite. 2023. Establishing a physics concept inventory using computer marked free-response questions. *European Journal of Science and Mathematics Education*, 11(2):360–375.

Ashutosh K. Pathak, Andrew James, Sally Jordan, and Jonathan Nylk. 2026. Student responses to a modified force concept inventory: The newtonian mechanics quiz. Dataset, Version 1.

Katherine D. Rainey, Michael Vignal, and Bethany R. Wilcox. 2022. Validation of a coupled multiple response assessment for upper-division thermal physics. *Physical Review Physics Education Research*, 18(2):020116.

N. S. Rebello and D. A. Zollman. 2004. The effect of distracters on student performance on the force concept inventory. *American Journal of Physics*, 72(1):116–125.

Johanna Schneider, Brian Schenk, and Abraham Bernstein. 2023. Towards trustworthy autograding of short, multi-lingual, multi-type answers. *International Journal of Artificial Intelligence in Education*, 33(1):88–118.

H.M.T.W. Seneviratne and S.S. Manathunga. 2025. Artificial intelligence assisted automated short answer question scoring tool shows high correlation with human examiner markings. *BMC Medical Education*, 25(1146).

Hwanjun Song, Minseok Kim, Dongmin Park, and Jae-Gil Lee. 2020. How does early stopping help generalization against label noise? In *Proceedings of the ICML Workshop on Uncertainty and Robustness in Deep Learning (UDL)*.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? In *Chinese Computational Linguistics*, volume 11856 of *Lecture Notes in Computer Science*, pages 194–206, Cham. Springer.

Chul Sung, Tejas Dhamecha, Swarnadeep Saha, Tengfei Ma, Vinay Reddy, and Rishi Arora. 2019. Pre-training BERT on domain resources for short answer grading. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 6071–6075, Hong Kong, China. Association for Computational Linguistics.

Yuki Takano, Tomoya Mizumoto, and Ryo Nagata. 2022. Automatic scoring of short answers using justification cues and BERT-based models. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 12–22, Seattle, Washington, USA. Association for Computational Linguistics.

Robert Tinn, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2023. Fine-tuning large neural language models for biomedical natural language processing. *Patterns*, 4(4):100729.

Amalie Trewartha, Nicholas Walker, Haoyan Huo, Sanghoon Lee, Kevin Cruse, John Dagdelen, Alexander Dunn, Kristin A. Persson, Gerbrand Ceder, and Anubhav Jain. 2022. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. *Patterns*, 3(4):100488.

Rebecka Weegar and Jakob Idestam-Almquist. 2024. Reducing workload in short answer grading using machine learning. *International Journal of Artificial Intelligence in Education*.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

J. Yasuda, M. M. Hull, and N. Mae. 2023. Visualizing depth of student conceptual understanding using subquestions and alluvial diagrams. *Physical Review Physics Education Research*, 19(2):020121.

Mengxue Zhang, Sami Baral, Neil Heffernan, and Andrew S. Lan. 2022. Automatic short math answer grading via In-Context meta-learning. In *Proceedings of the 15th International Conference on Educational Data Mining (EDM 2022)*.

Xinhua Zhu, Han Wu, and Lanfang Zhang. 2022. Automatic short-answer grading via BERT-based deep neural networks. *IEEE Transactions on Learning Technologies*, 15(3):364–375.

A Additional Per-Question Results

Panel A (Performance Metrics). Panel A reports per-question performance as mean $\pm$ standard deviation across runs and training sizes. We report accuracy, macro F1, weighted F1, and Cohen's κ.

Consistent with the results in Section 4, improvements in accuracy are accompanied by increases in agreement measured by Cohen's κ, indicating that gains are not limited to raw accuracy. Variability is higher for questions with lower baseline performance, reflecting greater instability under limited data conditions, as observed in the learning-curve analysis.

Panel B (Question Characteristics and Error Dynamics). Panel B links question properties to the mechanism of performance changes. N denotes the total number of repeated paired predictions aggregated across training sizes and runs; it should not be interpreted as the number of independent test instances. N_0 and N_1 denote the repeated counts of incorrect and correct labels, respectively. Imbalance Ratio and PosRate describe class prevalence.

While ΔAcc, reported in percentage points, indicates the magnitude of improvement, it does not by itself capture grading behaviour. The Fixed/Broken decomposition provides this insight:

- **Fixed**: baseline errors corrected by the domain-adapted model
- **Broken**: new errors introduced by the domain-adapted model
- **Net**: Fixed $-$ Broken, indicating the directional change in errors

Because the same fixed test responses are evaluated repeatedly across training sizes and random seeds, the fixed/broken counts are interpreted descriptively rather than as independent observations for significance testing. Consistent with the error-overlap analysis in Section 4, all questions show positive Net values, indicating that the domain-adapted model corrects more baseline errors than it introduces. Larger performance gains correspond to larger positive Net values, particularly for questions where baseline performance is lower. Taken together, Panels A and B support the finding that performance gains primarily reflect net reductions in grading errors across questions, rather than improvements in some questions offset by declines in others.

Table 5: Summary of statistical analyses used in the study.

Analysis	**Method**	**Unit**	n	**Reported**	**What is assessed**
Overall experiment-level comparison	Wilcoxon signed-rank	Matched run × training-size condition	260	W, p	Whether DAPT shows a repeated performance advantage over the baseline across matched experimental conditions
Global error-overlap comparison	Descriptive fixed/broken analysis	Aggregated repeated paired predictions	407,160	Fixed, Broken, Net, Fixed/Broken ratio	Whether DAPT corrects more baseline errors than it introduces; interpreted descriptively because test responses are repeated across training sizes and seeds
Per-question error behaviour	Descriptive fixed/broken analysis	Question-level aggregated repeated paired predictions	15 questions	ΔAcc, Fixed, Broken, Net	Whether improvements are concentrated in particular questions or reflect positive net error reduction across questions
Training size vs. improvement	Spearman ρ	Training-size condition	26	ρ	Monotonic association between training size and performance improvement (ΔAcc)
Baseline accuracy vs. question gain	Spearman ρ	Question	15	ρ	Monotonic association between baseline question performance and improvement under DAPT

Notes: Fixed/broken counts are interpreted descriptively because the same fixed test responses are evaluated repeatedly across training sizes and random seeds. ΔAcc is reported in percentage points. n denotes the number of analysis units, which differs by row.

Table 6: Appendix per-question results: performance metrics.

Panel A: Per-question metrics

Columns: Acc = accuracy; F1 = macro F1; wF1 = weighted F1; κ = Cohen's kappa.

	Baseline BERT				**Textbook-Pre-trained BERT**			
Q	**Acc**	**F1**	**wF1**	$\boldsymbol{\kappa}$	**Acc**	**F1**	**wF1**	$\boldsymbol{\kappa}$
Q1	0.987 ± 0.055	0.985 ± 0.059	0.987 ± 0.063	0.974 ± 0.088	0.994 ± 0.005	0.993 ± 0.006	0.994 ± 0.005	0.986 ± 0.013
Q2	0.942 ± 0.055	0.940 ± 0.062	0.941 ± 0.062	0.883 ± 0.109	0.966 ± 0.012	0.966 ± 0.012	0.966 ± 0.012	0.933 ± 0.024
Q3	0.954 ± 0.056	0.921 ± 0.082	0.956 ± 0.051	0.845 ± 0.156	0.988 ± 0.009	0.978 ± 0.016	0.988 ± 0.009	0.956 ± 0.032
Q4	0.786 ± 0.202	0.751 ± 0.212	0.773 ± 0.225	0.556 ± 0.329	0.954 ± 0.018	0.944 ± 0.021	0.953 ± 0.017	0.887 ± 0.041
Q5	0.843 ± 0.159	0.799 ± 0.178	0.837 ± 0.173	0.633 ± 0.281	0.969 ± 0.017	0.959 ± 0.023	0.969 ± 0.017	0.917 ± 0.045
Q6	0.946 ± 0.064	0.909 ± 0.075	0.946 ± 0.063	0.823 ± 0.129	0.963 ± 0.010	0.932 ± 0.017	0.962 ± 0.010	0.865 ± 0.035
Q7	0.830 ± 0.133	0.757 ± 0.133	0.821 ± 0.147	0.544 ± 0.181	0.911 ± 0.019	0.870 ± 0.028	0.910 ± 0.019	0.740 ± 0.055
Q8	0.800 ± 0.134	0.784 ± 0.156	0.791 ± 0.152	0.589 ± 0.274	0.890 ± 0.023	0.885 ± 0.023	0.890 ± 0.023	0.771 ± 0.046
Q9	0.833 ± 0.136	0.809 ± 0.161	0.822 ± 0.158	0.645 ± 0.267	0.928 ± 0.040	0.919 ± 0.046	0.926 ± 0.042	0.839 ± 0.090
Q10	0.903 ± 0.129	0.832 ± 0.149	0.901 ± 0.134	0.684 ± 0.236	0.968 ± 0.019	0.934 ± 0.034	0.967 ± 0.018	0.869 ± 0.066
Q11	0.917 ± 0.072	0.913 ± 0.083	0.915 ± 0.079	0.828 ± 0.155	0.953 ± 0.014	0.953 ± 0.014	0.954 ± 0.014	0.905 ± 0.028
Q12	0.922 ± 0.110	0.747 ± 0.143	0.937 ± 0.079	0.518 ± 0.247	0.971 ± 0.017	0.825 ± 0.063	0.972 ± 0.013	0.651 ± 0.124
Q13	0.965 ± 0.034	0.964 ± 0.036	0.965 ± 0.036	0.928 ± 0.067	0.982 ± 0.005	0.982 ± 0.005	0.982 ± 0.004	0.963 ± 0.009
Q14	0.990 ± 0.038	0.975 ± 0.060	0.991 ± 0.030	0.953 ± 0.109	0.999 ± 0.003	0.998 ± 0.009	0.999 ± 0.003	0.997 ± 0.018
Q15	0.987 ± 0.041	0.986 ± 0.049	0.986 ± 0.050	0.974 ± 0.077	0.995 ± 0.004	0.995 ± 0.004	0.995 ± 0.004	0.990 ± 0.009

Table 7: Appendix per-question results: question characteristics and descriptive error dynamics.

Panel B: Per-question characteristics and descriptive error dynamics

Columns: N = total repeated paired predictions aggregated across training sizes and runs; N_0/N_1 = repeated counts of label 0/1; Imb. Ratio = minority/majority class ratio; PosRate = N_1/N; ΔAcc (pp) = pre-trained $-$ baseline; Fixed/Broken = descriptive error changes; Net = Fixed $-$ Broken.

Q	N	N_0	N_1	Imb. Ratio	PosRate	ΔAcc (pp)	Fixed	Broken	Net
Q1	33800	9360	24440	0.383	0.723	0.70	335	98	237
Q2	15860	7540	8320	0.906	0.525	2.45	479	91	388
Q3	33280	5200	28080	0.185	0.844	3.41	1357	221	1136
Q4	32500	9620	22880	0.420	0.704	16.74	6169	727	5442
Q5	33020	8060	24960	0.323	0.756	12.55	4577	432	4145
Q6	32500	5460	27040	0.202	0.832	1.71	1167	610	557
Q7	16380	3640	12740	0.286	0.778	8.07	1765	443	1322
Q8	31720	18980	12740	0.671	0.402	8.96	4230	1388	2842
Q9	16120	5720	10400	0.550	0.645	9.44	2226	704	1522
Q10	32240	4940	27300	0.181	0.847	6.50	2599	505	2094
Q11	32500	13780	18720	0.736	0.576	3.69	1680	480	1200
Q12	32500	1300	31200	0.042	0.960	4.91	2204	609	1595
Q13	15860	6500	9360	0.694	0.590	1.71	291	20	271
Q14	16900	1300	15600	0.083	0.923	0.91	160	6	154
Q15	31980	14820	17160	0.864	0.537	0.82	347	86	261

Notes: Fixed/Broken counts are aggregated over repeated paired evaluations and are interpreted descriptively rather than as independent observations.