Beyond Flesch: Lightweight, Concept-Aware, and Energy-Efficient Difficulty Classification for Educational Texts

Shrishti Srivastava
Santa Clara University
ssrivastava2@scu.edu

Smruthi Danda
Santa Clara University
sdanda@scu.edu

Maneesha Prasanna
Santa Clara University
mprasanna@scu.edu

Abstract

Classifying the difficulty of educational texts by education level is critical to enable adaptive teaching with large language models, but prior work has relied on logistic regression using static and prompt-based metrics. We first replicate the approach of Rooein et al. (2024), who combine 46 static readability features with 63 LLM-derived prompt-based metrics in a multinomial regression to classify ScienceQA texts as elementary, middle, or high school. We then extend their work in three directions. First, we replace the pipeline with a fine-tuned scalar-mixed ELECTRA encoder testing on ScienceQA data without handcrafted features. Second, we apply domain-adversarial training (DANN) to the same encoder, reaching 0.83 - 0.88 macro-F1 on held-out CoQA and CommonLit, where the same architecture trained on synthetic data collapses to near random performance (0.22 - 0.41 macro-F1). Lastly, we fine-tune Phi-3.5-mini with LoRA adapters, improving our reproduced prompt-feature baseline while reducing inference latency by 127x and training energy by 83% lower training energy, and is the only configuration that correctly classifies all surface-hard/concept-easy cases in AdvConcept-50, a small adversarial benchmark we release. Code and data are publicly available at https://github.com/SCU-CSEN346/Beyond-Flesch and https://huggingface.co/datasets/nlpscu/Beyond-Flesch.

1 Introduction

Large language models are increasingly deployed in educational settings, from automated tutoring to content generation. Effective teaching requires matching the difficulty of content to the learner's level of education—a skill teachers develop through training but that current LLMs struggle to perform reliably (Rooein et al., 2024). Reliable measurement of text difficulty is a prerequisite for any system that aims to adapt content to learners.

Classical readability formulas such as Flesch-Kincaid (Flesch, 1948) rely on surface-level statistics like syllable count and sentence length. Recent work by Rooein et al. (2024) introduced PROMPT-BASED metrics—features derived by prompting an LLM with structured questions about a text—and showed that combining these with static features significantly improves difficulty classification on the ScienceQA dataset (Lu et al., 2022), achieving a macro-F1 of 0.95 for three-way grade-level classification. However, the authors explicitly flag generalizability as unaddressed: their experiments are constrained to a single dataset, and they note that ScienceQA is, to their knowledge, the only dataset of its kind.

This motivates our central research question: *Can we build a K–12 text-difficulty classifier that generalizes across corpora, distinguishes curriculum concept difficulty from surface readability, and is cheap enough to deploy?* We answer this question through four contributions:

- We provide the open-source reimplementation of the static and prompt-based difficulty classification pipeline of Rooein et al. (2024), reproducing all 46 static features (Appendix C) and 63 prompt-based metrics (Appendix A) across five LLM backbones on a balanced ScienceQA subset.
- We replace the feature-engineering pipeline with a fine-tuned scalar-mixed ELECTRA encoder, adapting the architecture of Gombert et al. (2024)[1] from biomedical regression to educational classification, and matching reproduction accuracy without handcrafted features.
- We apply domain-adversarial training (DANN) to the same encoder reaching

[1] https://github.com/SGombert/edutec-bea-shared-task-2024

0.83–0.88 macro-F1 on held-out CoQA and CommonLit, where the same architecture trained on synthetic LLM-generated text collapses to near-random, a controlled comparison that isolated the role of training-data diversity in cross-corpus generalization.

- We fine-tune Phi-3.5-mini (3.8B) with LoRA adapters, achieving the reproduction accuracy at 127× lower inference latency and 83% lower training energy, and release AdvConcept-50, a small adversarial benchmark that decouples surface readability from curriculum concept. Our LoRA model is the only configuration that correctly classifies all surface-hard/concept-easy cases in this benchmark.

2 Methodology

2.1 Reproduction: Static and Prompt-Based Metrics

We first reproduce the pipeline of Rooein et al. (2024). For each text in ScienceQA, we compute two feature sets. The first is 46 STATIC features compiled from prior readability research, covering surface-level properties such as word length, sentence length, and proportion of passive-voice verbs, as well as semantic features such as number of polysemic words and WordNet-based word-sense counts. These are computed with `textstat`, `spaCy`, and `NLTK`. The second is 63 PROMPT-BASED binary features derived from a user study (Rooein et al., 2024): structured yes/no questions are posed to an LLM, and each answer is encoded as 1 or 0. Questions span four categories: education-level suitability (30), readability score (15), topic (10), and lexical or syntactic complexity (8). We collect prompt responses from five open-source models—Gemma-7B-IT, Llama2-7B-Chat, Llama2-13B-Chat, Mistral-7B-Instruct-v0.2, and Qwen2.5-7B—each loaded with 8-bit quantization and temperature set to zero for deterministic outputs.

The two feature sets are fed into a multinomial logistic regression classifier under three conditions: STATIC only, PROMPT only, and COMBO (concatenation of both). Feature selection uses `SelectKBest` with univariate F-tests (`f_classif`), tuning $k \in \{10, 20, \ldots, 100\}$ and regularization strength $C \in \{0.1, 0.3, 1.0, 3.0, 10.0\}$ via grid search. Statistical significance of COMBO over STATIC is assessed by bootstrap resampling ($n = 1{,}000$).

2.2 Transformer Model

We replace the feature-engineering pipeline of Rooein et al. (2024) with an end-to-end transformer encoder. Our backbone is ELECTRA-large (Clark et al., 2020) (24 layers, hidden size 1024), which produces 25 hidden states per passage (embedding plus one per layer). Following Gombert et al. (2024), we combine them with ScalarMix, a learned weighted average:

$$h = \gamma \sum_{k=0}^{L} \mathrm{softmax}(w)_k \cdot h_k, \qquad (1)$$

where $L = 24$, $w \in \mathbb{R}^{L+1}$ are learnable weights, and γ is a global scale. Reading difficulty draws on multiple linguistic levels—vocabulary in lower layers, syntax in middle layers, semantics in upper layers—so the model learns the mixture rather than relying on the top layer alone. We mean-pool across tokens, apply dropout ($p = 0.2$), and pass the resulting 1024-dimensional vector through a two-layer MLP head outputting three logits. The model is trained end-to-end with cross-entropy loss.

2.3 Synthetic Data with Transformer Model

As a controlled test of cross-corpus generalization, we train the architecture of Section 2.2 on synthetic question-and-answer pairs generated by three commercial LLMs (Claude Haiku 4.5, DeepSeek Chat, Mistral Large), then evaluate on held-out real corpora. The architecture and training procedure are unchanged; only the training distribution differs. Using three generators isolates whether any failure is generator-specific or fundamental to LLM-generated training data.

2.4 LLM-as-a-Judge Label Unification

Because the datasets used in our generalization experiments define difficulty in different ways, we used Llama-3.1-8B-Instruct as an LLM judge to assign a unified difficaulty label to each text under one fixed rubric. Each example keeps both its original corpus label and the judge-assigned label, allowing us to compare training on inconsistent source labels versus a shared three-class label space: elementary, middle, and high. We use this setup as a diagnostic for whether cross-corpus failure is caused by incompatible label conventions rather than model capacity alone.

2.5 Domain-Adversarial Training

To force the encoder to learn corpus-invariant features, we apply domain-adversarial neural networks (Ganin et al., 2016). The shared backbone feeds two heads: a *difficulty head* (the task) and a *domain head* that predicts the source corpus. A Gradient Reversal Layer (GRL) between the encoder and the domain head acts as the identity in the forward pass and multiplies gradients by $-\lambda$ in the backward pass. The encoder therefore receives reversed gradients from the domain head and learns to produce representations from which corpus identity cannot be recovered, while the difficulty head receives normal gradients. The combined loss is $\mathcal{L} = \mathcal{L}_{\text{diff}} + \alpha \cdot \mathcal{L}_{\text{dom}}$ with $\alpha = 0.1$. We ramp λ from 0 to $\lambda_{\max} = 0.15$ on a sigmoid schedule, since applying adversarial pressure from step zero destabilizes training.

We train in two phases: phase 1 freezes the encoder and trains only the heads, phase 2 unfreezes the top 8 encoder layers and activates the GRL. We save two checkpoints—best validation macro-F1 (`best_val`) and the final post-DANN weights (`phase2_final`) because in-distribution validation consistently underestimates the OOD benefit of adversarial training.

2.6 LoRA-Phi-3.5

The final model is a lightweight decoder-based classifier. LoRA-Phi-3.5 (Pillar B+) finetunes `microsoft/Phi-3.5-mini-instruct` using LoRA (Hu et al., 2022). The base model is frozen, and only about 50M LoRA parameters (rank 32) are trainable. Each text is wrapped in a fixed instruction:

> *Classify the text by the curriculum grade level required to understand its concepts, not just its reading complexity. Answer with one letter: E (elementary, grades 1–5), M (middle, 6–8), or H (high school, 9–12).*

The prediction is read from the next-token logits at the answer position as the $\arg\max$ over the {E, M, H} tokens, requiring a single forward pass with no autoregressive generation.

This design removes the expensive 63-prompt feature extraction stage. Instead of asking many surface-correlated questions and training a separate classifier, LoRA-Phi-3.5 uses the language model's pretrained knowledge directly and learns a compact mapping from text to curriculum level. Training uses 16,849 multi-corpus examples re-labelled into three levels by an LLM judge. Because the levels are imbalanced (elementary 1,493; middle 11,026; high 4,330), we apply class-weighted sampling that draws roughly 3,500 examples per class per epoch, preventing the model from defaulting to the majority middle class. We train for two epochs in bf16 precision and track energy with CodeCarbon (Lacoste et al., 2019).

3 Evaluation Setup

3.1 Synthetic Data with Transformer Model

For each generator, we split the synthetic data 80/20 into training and held-in test partitions and train an independent model. After deduplication, the training pools contain 1,206 (Claude), 1,054 (DeepSeek), and 647 (Mistral) examples. Out-of-distribution evaluation uses three real corpora the model never saw, the ScienceQA test set, OneStopEnglish, and combined RACE middle+high.

3.2 Domain-Adversarial Training

We train the DANN model on the `llm_as_a_judge` splits. The training set contains 20,414 texts drawn from 15 corpora including ScienceQA, OneStopEnglish, RACE-middle, RACE-high, and CNN/DailyMail. The validation set contains 2,099 texts from the same corpora as training, held out for checkpoint selection. Out-of-distribution evaluation uses the held-out test split of 5,050 texts from four corpora the model never saw during training: XSum (2,000), CoQA (2,000), WeeBit (500), and CommonLit (500). We report the phase2-final checkpoint for all OOD results in Table 2, since in-distribution validation underestimates the OOD benefit of adversarial training.

3.3 LoRA-Phi-3.5

We evaluate LoRA-Phi-3.5 (Pillar B+) on three kinds of data. First, ordinary held-out splits (validation and test) measure in-distribution accuracy. Second, out-of-distribution corpora (RACE, OneStop) measure cross-corpus transfer. Third, AdvConcept-50 measures concept-vs-surface robustness. AdvConcept-50 is a hand-curated adversarial benchmark of 50 examples where surface readability and curriculum difficulty disagree, spanning three categories: surface-easy/concept-hard, surface-hard/concept-easy, and surface-matches-concept. Every label is grounded in an official

Prompt Model	STATIC	PROMPT	COMBO
Gemma-7B-IT	0.8005	0.6836	0.8290
Llama2-7B-Chat	0.8005	0.7045	0.8368
Llama2-13B-Chat	0.8005	0.7595	0.8405
Mistral-7B-Instruct	**0.8005**	**0.7431**	**0.8416**
Qwen2.5-7B	0.8005	0.7634	0.8394

Table 1: Macro-F1 on the ScienceQA test set ($n = 910$) across three feature conditions and five LLM backbones. STATIC is identical across all rows since it does not use an LLM. Best overall result (Mistral-7B COMBO) highlighted in bold. Bootstrap significance: COMBO vs STATIC $p = 0.031$.

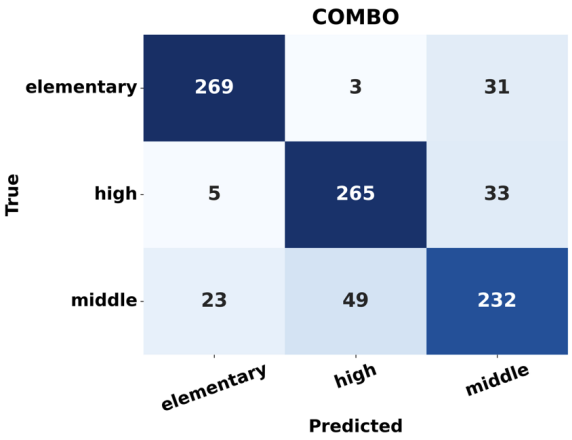

Figure 1: Confusion matrix for Mistral-7B COMBO ($k = 90$, $C = 3.0$) on the ScienceQA test set ($n = 910$). Diagonal entries show correct predictions. Middle school texts show the highest confusion rate, with errors concentrated at adjacent grade boundaries.

US K-12 curriculum standard (NGSS or Common Core). The benchmark is small, but it directly tests the failure mode that motivated the final architecture: separating what a text is *about* from how it *reads*. All macro-F1 figures are reported in Table 3.

4 Results

4.1 Reproduction Results

Across all five LLM backbones, STATIC features alone achieve a macro-F1 of 0.8005—a strong baseline that does not vary with LLM choice, since no LLM is involved in static feature computation. PROMPT-only performance ranges from 0.68 (Gemma-7B) to 0.76 (Qwen2.5-7B), consistently below STATIC. COMBO outperforms both conditions for every model, reaching a peak of 0.8416 with Mistral-7B (`SelectKBest` $k = 90$, $C = 3.0$). Bootstrap significance testing confirms that the COMBO improvement over STATIC is statistically significant ($p = 0.031$, improvement $+0.0412$).

Table 1 summarises macro-F1 across all conditions. Per-class analysis on the best model (Mistral-7B COMBO) shows elementary (F1 $= 0.897$) and high (F1 $= 0.855$) are classified reliably, while middle school texts are the hardest class (F1 $= 0.773$), with errors concentrated at adjacent grade boundaries. Figure 1 shows the confusion matrices across all three conditions for Mistral-7B. COMBO reduces middle-class errors most substantially, with misclassifications dropping from 97 (STATIC) to 72 (COMBO) for the middle class. These results confirm the central finding of Rooein et al. (2024): prompt-based features add complementary signal beyond static readability metrics, but only in combination.

4.2 Transformer Model

The transformer baseline reaches a macro-F1 of 0.89 on ScienceQA, matching the COMBO performance of the prompt-based pipeline without handcrafted or prompt-derived features. Performance saturates after one epoch and shows mild overfitting at the second, indicating that the in-distribution task is largely solved by the architecture itself. This motivates our focus on out-of-distribution evaluation in the remainder of the paper.

4.3 Synthetic Data with Transformer Model

Models trained on synthetic data learn the in-distribution task well: held-in macro-F1 reaches 0.91, 0.86, and 0.77 for Claude, DeepSeek, and Mistral respectively. The learned signal does not transfer to real text: macro-F1 collapses to the 0.14–0.36 range on ScienceQA and OneStopEnglish across all three generators. The prediction distributions reveal the failure mode: the *middle* label is assigned to the vast majority of real passages regardless of true label (e.g., 8,618 of 8,732 ScienceQA predictions for the Claude-trained model). The encoder has learned what LLMs *generate* as elementary or high-school text—stylistically exaggerated relative to real text—rather than what real text at each level looks like. The DeepSeek macro-F1 of 0.69 on RACE is an artifact of the two-class RACE split rather than a partial success: on the three-class ScienceQA and OneStop evaluations DeepSeek collapses similarly. This negative result motivates the domain-adversarial approach: in-

OOD corpus	Macro-F1
CoQA	0.83
CommonLit	0.88
OneStopEnglish (held-out)	0.48
WeeBit	0.31

Table 2: Macro-F1 on held-out OOD corpora for the DANN-trained ELECTRA model (`phase2_final` checkpoint).

creasing training-data diversity alone is insufficient if the model can exploit corpus-style shortcuts.

4.4 Domain-Adversarial Training

Domain-adversarial training substantially improves cross-corpus generalization (Table 2). On held-out CoQA and CommonLit, macro-F1 reaches 0.83 and 0.88—a three-to-five-fold improvement over the synthetic-data baseline on comparable real-text evaluations. The `phase2_final` checkpoint outperforms `best_val` on OOD corpora despite slightly lower validation macro-F1, confirming that in-distribution selection underestimates adversarial training's transfer benefit.

Error structure is informative. On CoQA, the confusion matrix shows zero elementary-to-high misclassifications in either direction across 2,000 examples; all confusion is concentrated at adjacent boundaries. The encoder has learned an ordinal representation of difficulty—when it errs, it errs by one level, not by two. Generalization is uneven: performance drops on OneStopEnglish (0.48) and WeeBit (0.31), both of which contain substantially shorter passages than the training corpora. This reflects sentence-length distribution shift in our training mixture rather than a fundamental property of the method.

4.5 LoRA-Phi-3.5

LoRA-Phi-3.5 (Pillar B+) is the final deployable system. It replaces the Rooein-style prompt-feature pipeline with one Phi-3.5-mini instruction prompt and a LoRA adapter. The final comparison in Table 3 shows that the gain is not limited to the ordinary ScienceQA test split. LoRA-Phi-3.5 improves the in-distribution test score from 0.840 to 0.872 macro-F1, improves RACE-high and OneStop transfer, and gives the strongest result on the concept-focused adversarial benchmark.

The most important result is the AdvConcept-50 behavior. DANN performed well on some held-out corpora, but AdvConcept-50 tests a different failure

Metric	Base	DANN	LoRA-Phi-3.5
Val macro-F1	0.880	–	**0.953**
Test macro-F1	0.840	–	**0.872**
Balanced test	0.856	–	**0.869**
Synth vs judge	0.918	–	**0.925**
Synth vs gen	0.940	–	0.893
RACE-middle	0.595	–	0.561
RACE-high	0.479	–	**0.546**
OneStop	0.415	**0.480**	0.460
AdvConcept F1	0.534	0.488	**0.775**
Hard/easy acc.	0.000	0.000	**1.000**

Table 3: Accuracy and robustness comparison. Base is the Rooein-style prompt-feature pipeline; DANN is the domain-adversarial model, evaluated where directly comparable. Rows are macro-F1 except where marked accuracy.

mode: whether the model can separate curriculum concept difficulty from surface readability. Domain invariance helps with source shift but does not by itself solve this surface-vs-concept problem.

On the AdvConcept-50 surface-hard/concept-easy category, which contains long or wordy sentences about elementary concepts, the prompt-feature baseline scores 0/6 while LoRA-Phi-3.5 scores 6/6. This is the exact case where readability formulas and surface-correlated prompt features should fail: the text looks hard, but the curriculum concept is easy.

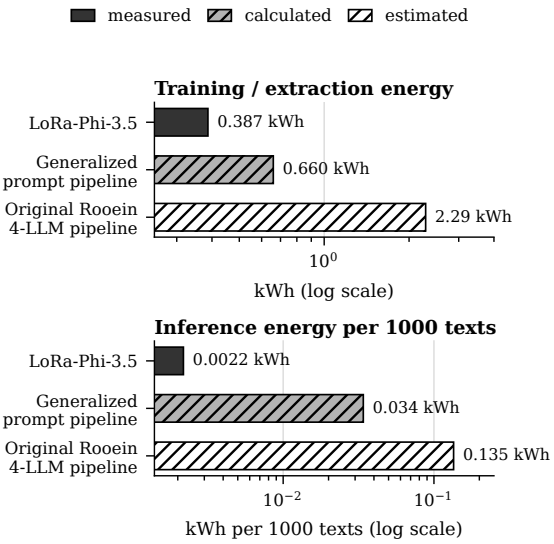

Figure 2: Energy cost for LoRA-Phi-3.5 and the prompt-feature pipelines. Solid bars are measured values; hatched bars are calculated or estimated values. LoRA-Phi-3.5 is cheaper because it removes the repeated prompt-feature extraction stage and uses one model call at inference.

The efficiency result is shown in Figure 2. The original four-LLM prompt pipeline uses an esti-

mated 2.29 kWh for training and feature extraction, compared with 0.387 kWh for LoRA-Phi-3.5. At inference time, the gap is larger: the original pipeline uses about 0.135 kWh per 1000 texts, while LoRA-Phi-3.5 uses about 0.0022 kWh per 1000 texts. In latency terms, this corresponds to roughly 5000 ms/text for the prompt-feature pipeline versus 39 ms/text for LoRA-Phi-3.5.

The final model is not uniformly best on every OOD split. The baseline remains stronger on RACE-middle and on the synthetic-vs-generated split. We therefore do not claim that LoRA-Phi-3.5 solves all cross-corpus difficulty classification. The supported claim is more specific: LoRA-Phi-3.5 gives the best overall tradeoff between accuracy, latency, energy, and concept-vs-surface robustness. For an educational setting where inference cost matters, that tradeoff is more useful than a heavier prompt-feature pipeline that depends on many LLM calls.

5 Conclusion and Future Work

We reproduced Rooein et al.'s prompt-metric pipeline and confirmed that COMBO features improve in-domain ScienceQA performance. The larger finding is that in-domain F1 hides serious transfer failures. Frozen ScienceQA classifiers collapse on OneStopEnglish, and synthetic data alone does not fix the problem. LLM-as-a-judge labels improve OOD transfer, showing that label consistency matters, but they are still only a partial solution.

The final LoRA-Phi-3.5 (Pillar B+) model gives the strongest practical result. It is concept-aware, faster, and more energy-efficient than the prompt-feature pipeline. It also performs best on AdvConcept-50, the benchmark designed to separate surface readability from difficulty of curriculum concepts. This supports our main conclusion: robust educational difficulty classification should evaluate concept difficulty, not only sentence complexity.

Future work should scale AdvConcept-50 to a teacher-validated AdvConcept-500, use a multi-judge labeling panel instead of one LLM judge, and combine the lightweight LoRA-Phi-3.5 (Pillar B+) direction with domain-adversarial training. DANN performed well on several held-out corpora, so a model that combines LoRA-Phi-3.5 (Pillar B+) efficiency with domain-invariant representation learning is a natural next step.

Limitations

The main limitation is data quality. The corpora use incompatible difficulty schemes, and our judge-label experiment shows substantial disagreement between original and unified labels. AdvConcept-50 is also small and should be expanded before making strong deployment claims. Finally, our demo and final model require GPU-oriented dependencies for best performance.

References

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35.

Sebastian Gombert, Lukas Menzel, Daniele Di Mitri, and Hendrik Drachsler. 2024. Predicting item difficulty and item response time with scalar-mixed transformer encoder models and rational network regression heads. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Donya Rooein, Paul Röttger, Anastassia Shaitarova, and Dirk Hovy. 2024. Beyond flesch-kincaid: Prompt-based metrics improve difficulty classification of educational texts. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*.