Investigating Adversarial Robustness in LLM based Automated Essay Scoring

Renjith P Ravindran
ETS Assessment Services
rpravindran@ets.org

Ikkyu Choi
ETS Research Institute
ichoi001@ets.org

Abstract

Automated Essay Scoring (AES) is one of the most widely studied applications of Natural Language Processing (NLP) in education and educational measurement. Recent advances with pre-trained Transformer-based large language models (LLMs) have shifted AES from feature-based modeling to leveraging contextualized language representations. These models provide rich semantic representations that substantially improve scoring accuracy and human–machine agreement compared to systems relying on handcrafted features. However, their robustness towards adversarially crafted inputs remains poorly understood. In this study, we define adversarial input as any modification of the essay text designed to fool an automated scoring system into assigning an inflated score. We evaluate a fine-tuned DeBERTa-based AES model on such inputs and show that it is highly susceptible to a simple text duplication attack, highlighting the need to consider adversarial robustness alongside accuracy in the development of AES systems.

1 Introduction

Automated Essay Scoring (AES) is one of the earliest applications of Natural Language Processing (NLP) to educational assessment, with roots dating back to the 1960s (Page, 1967). Over the decades, AES systems have evolved from statistical models with shallow surface-level features to highly sophisticated neural architectures (Beigman Klebanov and Madnani, 2020). Traditional approaches often relied on handcrafted features designed to approximate lexical diversity, syntactic complexity, discourse organization, and stylistic control. For example, the use of connectives such as "therefore" or "in conclusion" could serve as proxies for argumentative structure, while measures such as type–token ratio or average sentence length are aimed at capturing lexical richness (Chodorow and Burstein, 2004). These approaches, although effective to some extent, are inherently limited: they depend heavily on feature engineering and are vulnerable to superficial manipulation by test takers (Powers et al., 2001; Chodorow and Burstein, 2004; Perelman, 2020).

The advent of deep learning (Goodfellow et al., 2016), and more recently pre-trained Transformer (Vaswani et al., 2017) based large language models (LLMs), has reshaped the AES landscape. Transformer-based models such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2020) learn contextual representations of text that capture lexical, syntactic, and semantic information simultaneously. When fine-tuned on essay scoring datasets, these models substantially increase the agreement between machine predictions and human raters, often measured using Quadratic Weighted Kappa (QWK) (Li and Ng, 2024). This leap in performance could lead to growing enthusiasm towards operational deployment of LLM-based AES in high-stakes testing environments.

Yet, the question of robustness remains underexplored (Ding et al., 2020; Kabra et al., 2022). Accuracy gains in typical test settings do not guarantee resilience under adversarial conditions. Adversarial attacks in NLP — ranging from synonym substitution in sentiment analysis (Zhou et al., 2021) to input perturbations in machine translation (Michel et al., 2019) — have shown that state-of-the-art models can be surprisingly fragile. In educational contexts, this fragility has serious implications. Unlike sentiment classification or translation, AES models directly influence student outcomes. If models can be "fooled" by trivial manipulations, such as artificially inflating essay length or inserting irrelevant but sophisticated-sounding sentences or words, the integrity of automated scoring is jeopardized. This is particularly concerning given the high stakes of standardized assessments, where

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 3: Coordinated Session Papers, pages 86–91

even a one-point increase in an essay score can affect admissions or scholarship decisions.

Prior work has begun to highlight these vulnerabilities. Ding et al. (2020) showed that content scoring systems can be misled by adversarial strings of meaningless characters. Kabra et al. (2022) proposed toolkits for systematically probing AES robustness, underscoring the need for adversarial evaluation. Jeon and Strube (2021) demonstrated that essay length continues to exert disproportionate influence on neural AES models, echoing concerns that date back to earlier systems (Chodorow and Burstein, 2004). Collectively, this line of work suggests that LLM-based AES models, despite their sophistication, may inherit structural weaknesses from both feature-based and neural predecessors.

In this preliminary study, we take a focused step toward systematically evaluating adversarial robustness of an LLM-based AES model. Specifically, we examine the behavior of a DeBERTa-based scoring system fine-tuned on the Persuade 2.0 corpus (Crossley et al., 2024). We design and test three adversarial scenarios that are both simple to implement and highly plausible in real testing conditions:

- Appending high-impact words, where the test taker simply append few words that are likely to be found in high scoring essays. If an automatic scoring model is overly relying on uni-grams such essays could see a boost in score.

- Fancy-language injection, where a short paragraph of complex, topic-agnostic sentences are appended to the essay to mimic advanced vocabulary and sentence structure.

- Text duplication, where a test taker repeats their essay once or twice to artificially inflate length. Scoring models often pick up essay length as a proxy to essay quality, duplication of text is the easiest way to increase essay length.

To provide additional context for robustness, we also examine noise-based perturbations such as scrambling words or sentence spans. These manipulations allow us to probe the model's reliance on lexical coherence versus discourse-level organization.

Findings from our preliminary study are twofold:

- We demonstrate that a DeBERTa-based AES model, while achieving strong baseline accuracy (QWK = 0.87), is highly vulnerable to text duplication, with systematic and substantial score inflation.

- We show that the model is relatively robust to high-impact lexicon and sentence insertions, suggesting that sophisticated vocabulary and structuring without semantic relevance does not easily fool the system.

Taken together, these findings highlight a central tension in AES research: while LLMs improve accuracy, they do not automatically confer robustness. Even trivial adversarial strategies can yield large score changes, raising fairness and validity concerns. We argue that adversarial robustness should be treated as a primary design criterion for AES, alongside scoring accuracy and reliability, and we hope this work stimulates further research in this direction.

2 Experiment Setup

For our experiments we use the Persuade Corpus 2.0 (Crossley et al., 2024), a large-scale dataset of approximately 25,000 student essays written by grades 6–12 in response to argumentative writing prompts. Each essay in this corpus has been scored holistically by human raters on a six-point ordinal scale (1 = weakest, 6 = strongest), reflecting overall writing quality rather than individual analytic traits. The dataset is particularly suitable for adversarial evaluation because it is both large enough to fine-tune LLMs effectively and realistic in content, covering authentic student writing with diverse levels of proficiency. In addition, Persuade 2.0 is a recent corpus explicitly designed to advance AES research, which makes it a valuable benchmark for studying not only predictive performance but also model robustness. For training and evaluation, we adopt the official splits, which contain 15,528 items in the training set and 10,356 items in the test set.

Our AES model is built on DeBERTa (He et al., 2020), a Transformer-based large language model that improves upon BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) through disentangled relative attention mechanisms. Specifically we used DeBERTa V3 base (He et al., 2021) via the Hugging Face Transformers Python Module. An important parameter that is relevant for this particular study is the token limit, which limits the size

transform	mean change (sd)
scramble-words	-2.10 (1.0)
scramble-sents	-0.06 (0.2)
add-low-words	-0.13 (0.2)
add-high-words	-0.09 (0.1)
add-smoke	-0.05 (0.2)
length-2x	0.93 (0.3)
length-3x	1.28 (0.6)

Table 1: Mean score change (SD) per transform (score range 1-6).

of the input text. Thanks to relative positioning bias in DeBERTa, the maximum number of tokens is given by $(2k-1)l$, where l is the number of layers and k is the maximum relative distance allowed between tokens. Since in DeBERTa base $l = 12$ and $k = 512$, we can have a maximum of 24,528 tokens in the input text. This is lower than the thrice the number of tokens in the longest essay (1902 tokens [1]) in the dataset. To adapt DeBERTa for essay scoring, we attach a simple regression head on top of the [CLS] token embedding to predict continuous essay scores in the range 1–6. The head consists of a two-layer feed-forward network trained jointly with the DeBERTa encoder, so that the model learns both task-specific features and general linguistic representations. Training is performed with mean squared error (MSE) loss, and model quality is evaluated using Quadratic Weighted Kappa (QWK), a standard metric for measuring agreement with human raters in AES research. When evaluated on the test-set our model gives a QWK of 0.87.

2.1 Attacks

2.1.1 Adding High Scoring Words

This attack tests whether the scoring model is relying on uni-grams to score the essay. High scoring essays are likely to have impactful vocabulary. Thus test takers may add such words out of context in the hope of triggering the scoring model to award a higher score. To find such words, we split the dataset (train) into two, a high scoring set which has essays having scores 4, 5, 6 and a low scoring set with essays having score 1, 2, 3. Now for each word we find its log-odds ratio of probability of the word occurring in the high scoring set over the probability of the word occurring in the low scoring set. This allows us to rank words based on how likely they are to be found exclusively in high scoring essays. The attack then is to append a sequence of 10 words sampled randomly from the top 100 of these words to each of the test essays. As this is a preliminary study the choice of 10 is informed intuitively as the likely number of words test takers may add. We defer testing a range of numbers to future work. Here is such a random sample: *dependence*, *traditional*, *theatre*, *platforms*, *extracurriculars*, etc. This attack is referred to as add-high-words.

2.1.2 Adding Fancy Language

Our next attack is to test the impact of adding a paragraph with impactful sentence structure and vocabulary. This simulates the situation where test takers may memorise a piece of fancy sounding text that could be added to any essay in order to trigger the machine to give a higher score. To study the effect of fancy-language injection, we transform the essay texts by adding the following to the end of each essay: "*Conceptual dynamics often emerge through the oscillations of undefined frameworks. This interaction, while nebulous, suggests a layered intentionality. Consequently, abstraction persists as both method and outcome.*". This is an arbitrary piece of text intended to add a dose of potentially high-impact vocabulary and sentence structure. As this is a preliminary study, we do not attempt to quantify what is meant by high-impact, and also limit the study to considering only a single instance. We refer to this attack as add-smoke.

2.1.3 Inflating Essay Length

Essay scores are often correlated to its length. One of the easiest ways a test taker can game this feature, without adding out of context text is to simply duplicate their essays. To study the effect of text-duplication we transform the essay text by duplicating it once, and twice, referred to as length-2x

[1] tokens here refer to lexical units after tokenisation

and length-3x, respectively.

2.1.4 Baselines

To understand the general robustness of our model we add two more transformations, scramble-words: in which words in the essay are scrambled, and scramble-sents: in which the spans of text separated by newlines are scrambled (note that this is not perfect sentence scrambling). To contrast with the add-high-words attack we include a add-low-words where we append words that are exclusively found in low scoring essays. The intention is to test if adding such words can lower the essay scores. A random sample from add-low-words : *luke*, *electrol*, *presendent*, *thay*, *negitive*, etc. We find that most of these are typos, and therefore can be expected to bring down scores when added to essays.

3 Experiments and Results

3.1 Average Score Change

Table 1 gives the mean and standard deviation of the score change induced by each transformation. The first clear observation is that scrambling words devastates performance (–2.10 average). This is expected: scrambling disrupts local coherence, making essays nonsensical.

In contrast, scrambling sentences produces almost no change (–0.06). This suggests that the DeBERTa-based AES model may be largely insensitive to discourse-level ordering of sentences. Although discourse coherence is a key aspect of human evaluation, our results imply that the model's reliance on the [CLS] embedding fails to adequately capture paragraph-level or argumentative flow. This insensitivity could become problematic if test takers deliberately manipulate essay structure while maintaining superficial lexical quality.

The more striking pattern emerges with duplication attacks. Doubling essay length (length-2x) increases average scores by +0.93, and tripling (length-3x) by +1.28. These are substantial gains considering the total score range is only 1–6. The effect size rivals the difference between adjacent holistic score levels as judged by human raters. Put differently, a mediocre essay rated 3 could be artificially boosted into the "proficient" range (4–5) simply by repetition.

Interestingly, adding high-scoring words or high impact vocabulary and sentence structure out of context doesn't increase the scores, instead marginally decreases the scores. This contrasts with anecdotal expectations that "sophisticated" vocabulary could fool models. Instead, the AES model appears somewhat robust to this type of lexical padding, possibly because embeddings capture topical mismatch between the appended text and the main essay body.

3.2 Score Change at each Human Score Level

Table 2 disaggregates score changes by human-assigned scores. This analysis yields three notable insights.

scramble-words degrades higher-quality essays more severely. Essays originally scored 6 lose over 4 points, while those scored 1 lose less than 1.

Duplication benefits mid-range essays the most. For length-2x and length-3x, the largest gains occur at human scores 3–4. For example, a 3-rated essay rises on average by +1.42 under length-3x. This reflects the model's tendency to conflate length with quality in borderline cases. Such vulnerabilities are particularly concerning because many operational decisions hinge on distinguishing "adequate" from "proficient" performance in this mid-range.

add-smoke and add-high-words has negligible effects across all bins. The consistency of near-zero changes suggests that superficial stylistic padding does not easily exploit this model.

3.3 Score Change Distribution

Average changes alone can obscure practical impact. Figure 1 therefore examines the distribution of rounded score differences under duplication.

For length-2x, nearly 80% of essays increase by at least +1 point, and around 10% gain +2 points. Such shifts could materially alter student outcomes: an essay initially rated 3 (marginal) may be reclassified as 4 (proficient).

For length-3x, the effects are even more dramatic: 50% of essays gain +1 point and 40% gain +2. In practice, this means almost every duplicated essay is rewarded, with a non-trivial fraction jumping two score categories.

Very few essays decrease in score, confirming that duplication is a high-reward, low-risk adversarial strategy.

These findings underscore the operational significance of duplication: if undetected, test takers can consistently and predictably exploit the scoring system.

transform	1	2	3	4	5	6
scramble-words	-0.39 (0.4)	-1.08 (0.4)	-1.86 (0.5)	-2.69 (0.5)	-3.59 (0.5)	-4.29 (0.3)
scramble-sents	-0.01 (0.1)	-0.04 (0.1)	-0.07 (0.2)	-0.07 (0.2)	-0.05 (0.2)	-0.02 (0.1)
add-low-words	-0.07 (0.1)	-0.10 (0.1)	-0.13 (0.2)	-0.17 (0.2)	0.17 (0.2)	0.09 (0.1)
add-high-words	-0.06 (0.1)	-0.08 (0.1)	-0.09 (0.1)	-0.12 (0.2)	0.08 (0.2)	0.00 (0.1)
add-smoke	-0.05 (0.1)	-0.07 (0.1)	-0.08 (0.2)	-0.07 (0.2)	0.03 (0.2)	0.09 (0.1)
length-2x	0.82 (0.4)	0.91 (0.3)	0.99 (0.3)	1.04 (0.3)	0.78 (0.4)	0.27 (0.3)
length-3x	1.12 (0.6)	1.42 (0.4)	1.49 (0.5)	1.33 (0.5)	0.71 (0.6)	0.01 (0.6)

Table 2: Mean score change (SD) at each human score level.

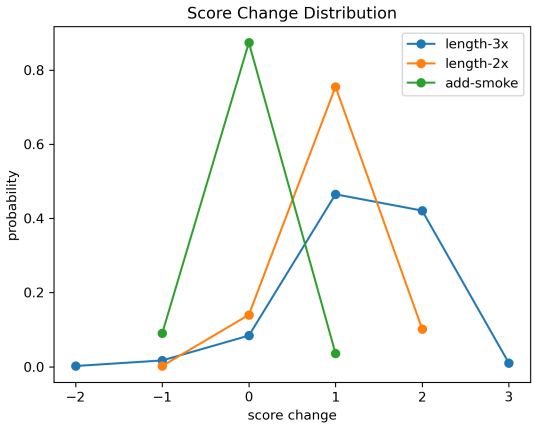

Figure 1: Score change distribution.

4 Conclusion

This study has examined the adversarial robustness of an LLM-based AES system trained on the Persuade 2.0 corpus. While the baseline DeBERTa model achieved strong agreement with human raters (QWK = 0.87), our experiments reveal that high scoring accuracy alone does not guarantee robustness to adversarially crafted responses. The most striking finding is the model's vulnerability to duplication: repeating an essay once or twice almost always leads to inflated scores, with gains of one or even two points on a six-point scale. Because such changes occur consistently across a large portion of the test set, they represent a genuine threat to the validity of AES in operational settings. Even if duplication is easy to detect with simple preprocessing, the fact that a trivial manipulation yields such predictable benefits underscores the importance of evaluating AES systems against adversarial input.

At the same time, the results also highlight areas where the model appears more robust. The insertion of sophisticated but irrelevant sentences ("smoke text") produced negligible effects, and the more systematic attempt to append vocabulary disproportionately associated with high- or low-scoring essays also failed to move predictions in a meaningful way. These negative results suggest that the model does not simply reward isolated lexical items, even when those items are correlated with writing quality in the training data. Instead, it appears to integrate vocabulary in context, discounting out-of-place words. This robustness to shallow lexical padding contrasts with the severe susceptibility to length manipulation, pointing to a specific structural weakness rather than a general fragility.

It is to be noted that these results are from our preliminary study along these lines. A major limitation of this study is that we have evaluated only one kind of model. A comprehensive evaluation is being planned as future work with multiple AES models, and to address other limitations.

More generally future studies should pursue two directions in parallel: developing systematic taxonomies of adversarial risks in AES (including semantic drift, coherence disruption, and targeted vocabulary injection), and exploring defenses that go beyond heuristic filters. Possibilities include explicit modeling of discourse, normalization against essay length, and the integration of adversarial training protocols.

Ultimately, if AES systems are to be trusted in high-stakes testing, adversarial robustness must be evaluated alongside accuracy and fairness. Our results provide early evidence that while certain manipulations are resisted, others remain alarmingly effective. Robustness cannot be assumed from model sophistication alone; it must be deliberately measured and built into the design of future AES systems.

References

Beata Beigman Klebanov and Nitin Madnani. 2020. Automated evaluation of writing – 50 years and counting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7796–7810, Online. Association for Computational Linguistics.

Martin Chodorow and Jill Burstein. 2004. Beyond essay length: Evaluating e-rater®'s performance on toefl® essays. *ETS Research Report Series*, 2004(1):i–38.

Scott Andrew Crossley, Y Tian, P Baffour, Alex Franklin, Meg Benner, and Ulrich Boser. 2024. A large-scale corpus for assessing written argumentation: Persuade 2.0. *Assessing Writing*, 61:100865.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yuning Ding, Brian Riordan, Andrea Horbach, Aoife Cahill, and Torsten Zesch. 2020. Don't take "nswvtnvakgxpm" for an answer –the surprising vulnerability of automatic content scoring systems to adversarial input. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 882–892, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Sungho Jeon and Michael Strube. 2021. Countering the influence of essay length in neural essay scoring. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pages 32–38, Virtual. Association for Computational Linguistics.

Anubha Kabra, Mehar Bhatia, Yaman Kumar Singla, Junyi Jessy Li, and Rajiv Ratn Shah. 2022. Evaluation toolkit for robustness testing of automatic essay scoring systems. In *Proceedings of the 5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD)*, CODS-COMAD '22, page 90–99, New York, NY, USA. Association for Computing Machinery.

Shengjie Li and Vincent Ng. 2024. Automated essay scoring: A reflection on the state of the art. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17876–17888, Miami, Florida, USA. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Paul Michel, Xian Li, Graham Neubig, and Juan Pino. 2019. On evaluation of adversarial perturbations for sequence-to-sequence models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3103–3114, Minneapolis, Minnesota. Association for Computational Linguistics.

Ellis B Page. 1967. Statistical and linguistic strategies in the computer grading of essays. In *COLING 1967 Volume 1: Conference internationale sur le traitement automatique des langues*.

Les Perelman. 2020. The babel generator and e-rater: 21st century writing constructs and automated essay scoring (aes). *Journal of Writing Assessment*, 13(1).

Donald E. Powers, Jill C. Burstein, Martin Chodorow, Mary E. Fowles, and Karen Kukich. 2001. Stumping e-rater: Challenging the validity of automated essay scoring. *ETS Research Report Series*, 2001(1):i–44.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yi Zhou, Xiaoqing Zheng, Cho-Jui Hsieh, Kai-Wei Chang, and Xuanjing Huang. 2021. Defense against synonym substitution-based adversarial attacks via Dirichlet neighborhood ensemble. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5482–5492, Online. Association for Computational Linguistics.