Investigating Universal Adversarial Attacks Against Transformers-Based Automatic Essay Scoring Systems

Igor Cataneo Silveira(✉), André Barbosa, Daniel Silva Lopes da Costa, and Denis Deratani Mauá

Universidade de São Paulo, São Paulo, Brazil
{igorcs,abarbosa,ddm}@ime.usp.br, dslcosta2016@usp.br

Abstract. Automatic Essay Scoring promises to scale up student feedback on written input, addressing the excessive cost and time demand associated with human grading. State-of-the-art automatic scorers are based on Transformers-based neural networks. While such models have shown impressive results in reasoning tasks, learned models often produce answers that arise from statistical clues in datasets and are misaligned with human objectives. Such systems are thus potentially fragile for scenarios where users are incentivized to deceive the system, as in a classroom setting. In this work, we evaluate the susceptibility of state-of-the-art automatic scorers to attacks made by non-expert users, such as students interacting with an automatic grader. We develop a methodology to simulate such student attacks and test them against scorers based on BERT, Phi-3 and Gemini models. Our findings suggest that (i) a BERT-based grader can be deceived using simple feature-based attacks; (ii) although Google's Gemini has a solid agreement with graders, it can assign undeservedly high grades for small sentences; (iii) a Phi-3-based grader was less susceptible than BERT, but it still assigned relatively high grades to some of our attacks.

Keywords: Automatic Essay Scoring · Adversarial Attack · Natural Language Processing

1 Introduction

Automatic Essay Scoring (AES) systems ease the burden on teachers while enhancing student learning by providing meaningful, timely and personalized feedback [32]. The technology is currently mature enough to be employed in high-stakes standardized exams such as the GRE and TOEFL [3,5].

Like many Machine Learning based solutions, AES systems are prone to malicious usage and are sensitive to spurious correlations in the training data [22]. This is particularly relevant in educational and scoring settings, where users might be motivated to exploit the system's vulnerability, and where a correct prediction for the wrong reason might be as harmful as an incorrect prediction.

A. Paes and F. A. N. Verri (Eds.): BRACIS 2024, LNAI 15413, pp. 169–183, 2025.
https://doi.org/10.1007/978-3-031-79032-4_12

Unintended uses of predictive systems are known in the Machine Learning literature as adversarial attacks [21]. The more recent literature cares specifically about attacks that are carried out by changing an existing input in a human imperceptible way and such that the prediction shifts towards or away from some desired target value [13,44]. For example, in AES that might amount to determining words or expressions in an essay that, if changed, maintain the text's overall characteristics as perceived by a human while leading the predictive model to raise the score significantly. Such attacks usually require knowledge and access to the system's characteristics. Thus, they are usually difficult to be performed by regular users of AES systems.

A more straightforward approach is the universal attack, which consists of finding input-unrelated rules that cause the predictive model to increase scores. For example, it has been noticed that essay length is a good predictor of essay quality [26]. A malicious user can exploit this fact by appending unrelated text to an essay to increase its score. Such an attack does not require deep knowledge of the system and can be formulated by repeated use.

Designers of AES systems have long been aware of the pitfalls of malicious usage. For instance, authors have discussed whether to remove certain predictive features such as text length in order to improve the system's robustness against attacks [42]. The matter is complicated by the fact that textual features are often correlated, so that removing a feature might hurt performance without actually making the system less susceptible to attacks.

Following the trend in Natural Language Processing, state-of-the-art AES systems are based on Transformer deep neural networks [27]. While in principle such models can learn rich representations and enable logical and commonsense reasoning, there is evidence that learning algorithms and spurious correlations in data often lead to predictions being based on simple statistical cues [43]; hence such systems might be as exposed to attacks as traditional feature-based systems.

In this work, we investigate whether **Transformers-based AES systems are susceptible to universal adversarial attacks that might occur in a classroom setting**. In particular, we consider a fictitious (but realistic) case where users (i.e., students) interact with the system by submitting an essay (input) and receiving a score (the output), while having no knowledge of the predictive model being used (including parameters, derivatives, etc.). The hypothetical system is non-adaptive (i.e., no learning after deployment) and students use the system repeatedly (e.g. grading different essays along the academic term or being able to submit an improved version) and share information about usage among themselves. Given that cheating in the classroom is prevalent [20,41], some students are expected to exploit any vulnerability to improve their grade without necessarily improving the quality of their texts.

To simulate how such non-expert users might learn vulnerabilities, we train an interpretable predictive model (a linear regressor) based on handcrafted features and select the most predictive features to compose attacks, such as inserting lists of adjectives of adverbs, employing more adjectives and adverbs than usual, and repeating the same sentence multiple times. The rationale is that by repeated interaction and information sharing, users might develop a simplified model of the system's inner work and use such a model to come up with such simple attacking strategies.

The effectiveness of the attacks is evaluated with respect to BERT, Phi-3 and Gemini-based AES systems on a benchmark of human annotated (Brazilian) Portuguese essays [34]. We show that all models are vulnerable to simple attacks.[1] For example, on a scale of 0 to 1000, where higher numbers are better, the BERT and Phi-3 based models assigned, respectively, 800 and 560 points to a universal attack (i.e., a fabricated essay) based on repeating a generic conclusion employing many adverbs and adjectives. The Gemini-based model showed resilience against attacks based on repetition, but for instance assigned a relatively high score of 640 points to an essay consisting of only a small list of adjectives.

In the rest of the document, we present related work in Sect. 2, discuss our methodology in Sect. 3, and show the experimental results in Sect. 4. We conclude with a summary and final remarks in Sect. 5.

2 Related Work

The evolution of AES systems mimics that of other NLP tasks. The initial systems were based on handcrafted features, which had limited predictive power and require extensive expertise to be designed [32]. Those systems were incrementally superseded by systems that learned representations directly from data. First, by approaches that extracted shallow representations such as word embeddings [37]. More recently, by approaches that learn representation by deep models such as Transformer neural networks [27].

Several AES systems for (Brazilian) Portuguese have been proposed, mostly focusing on grading essays in the format required for the Brazilian national entrance exam (ENEM) [2,4,12,15,28,34]. An early work used words as features of a Naïve Bayes [4]. Subsequent approaches ranged from defining handcrafted features to combining word-embedding with Recurrent Neural Networks [2,15,29]. Our study intersects with the existing literature by evaluating the performance of previously unused models, specifically the Phi-3 and Gemini-based graders. Our goal is not pursuing high agreement, but presenting how susceptible these novel models are to adversarial attacks. A recent work compared several (Brazilian) BERT-based graders [34]. We incorporate these models into our study, alongside new state-of-the-art models, for comparative analysis.

Different AES systems were assessed based not on their QWK performance, but rather on their oversensibility and overstability [21]. The authors investigated how perturbations in the input affected the models' output. The ASAP challenge dataset [18] was used to test five models, ranging from a feature-based to a BERT-based grader. They found that, except for the feature-based and one of

[1] The models are available at: https://github.com/kamel-usp/aes_enem/.

the word-embedding-based, all models were overstable. Moreover, increasing the disturbance percentage lead to increased standard deviation in the grades. Their work investigated whether adding, removing or changing content of an essay caused it to be graded differently. This can be seen as students rewriting their essays after grading. Our scope is different, for they do not consider universal attacks, and we consider ill-intended attacks that could be done by students with minimal effort.

Without any scenario restrictions, there are possibly infinite adversarial attack setups. We rule out possible attacks by restricting ourselves to a very specific scenario. Firstly, students do not have access to any parts of the model, so they are not allowed to attack the embedding space [16], the gradient [11], nor the calculated features, to name a few. Secondly, they do not know the underlying model, which prevents them from using the pre-trained version of the model to generate attacks, such as [25]. Lastly, the students are not too invested in breaching the system, so attacks such as [14,40] are not applicable. The first requires (possibly) testing combinations of the whole vocabulary used by the tokenizer of the underlying model. This is ruled out of our scenario as they do not have access to this information, some models do not have a vocabulary, and Gemini's vocabulary has 256k tokens, so it would require at least this many attacks. Similarly, the second attack requires changing the whole essay to have the same word, which results in a massive number of needed attacks. Thus, the attack we devise here is much closer to what one could expect to happen at a classroom level and has not been used yet in the literature.

The task of Automatic Short Answer Grading (ASAG) is very similar to AES, the difference is that, instead of (long) essays, its input is a relatively small text. Adjectives and adverbs have already shown to be useful to cheat ASAG systems based on Transformers, such as BERT and T5 [13]. The scenario of their study is similar to ours. The differences are: they are concerned with generating answers that cannot be identified by humans, and that their labels are: correct, incorrect or contradictory. By inserting adjectives and adverbs in the sentences, they could fool the models to change the answer from incorrect to correct between 8 and 22% of the cases. Showing that they are strong candidates to fool the model. Additionally, their investigation found that humans did not find the adversarial examples suspicious, but they were not natural either.

To evaluate the robustness of available models against such adversarial attacks, we rely on the LMSys benchmark [9] to choose candidates. LMSys is a comprehensive benchmarking platform designed to evaluate the performance of large language models across a wide array of tasks, including natural language understanding, generation, and interaction. By providing standardized evaluation metrics and diverse benchmarks, it enables a consistent comparison of model capabilities, thus serving as an invaluable resource for researchers and developers. As of the current date, the 10 best-ranked models available there are proprietary and cannot be easily fine-tuned due to their restricted access. Among the available ones, Nemotron [31] and the top-performing models from the Llama family [39] are notably expensive to run, limiting their practicality for widespread use.

In contrast to this trend, there is a growing interest in "Small" Language Models, such as Llama3-8B and the Phi-family [17]. The Phi-family has recently released the Phi-3 models [1], which include three distinct model sizes: 3.8B (Phi-3-mini), 7B (Phi-3-small), and 14B (Phi-3-medium). These models, despite their smaller size, are claimed by the authors to outperform some proprietary models, such as GPT-3.5, in various tasks.

We chose the Phi-3 model for our research due to its open-source nature, which allows for easier access and adaptability. Additionally, the Phi-3 model was trained exclusively on English data and has not been extensively tested in other languages. Our study aims to evaluate the performance of the Phi-3 model using Portuguese-only data, contributing to the understanding of its capabilities and limitations in multilingual contexts.

The GPT family (Davinci, 3.5 turbo and 4.0) have been tested in AES and ASAG. The authors of [30] showed that the Davinci model could achieve 0.38 QWK in the TOEFL dataset, while linguistic features could achieve almost 0.6 QWK. GPT 3.5 and 4.0 were used to grade short answers in Finnish [7]. The conclusion was that the models could not be directly employed and that they are more lenient than human graders, assigning fewer failing grades. While OpenAI models are the most famous, their cost is prohibitive, thus we explore here the usage of Gemini's free API and how resilient it is against adversarial attacks.

3 Methodology

We start by selecting the ENEM-AES dataset and pre-trained models [34], which are easily accessible through Hugging Face API. Essays in this dataset are evaluated according to five difference criteria, or competencies, and separated predictors have been trained for each specific criterion. We use the BERT-base Ordinal model available for each competence, as they are relatively small and achieve high performances. Additionally, we used the splits available using the "propor2024" parameter.

We also used the scikit-learn package [33] to train, for each individual competence, a Linear Regressor that uses 72 features calculated by the open API of NILC-Metrix [24]. These features are interesting because they calculate various linguistic levels of written language, such as the ratio of a certain part of speech, how ambiguous the words are, text length, Flesch score and others. Although they might be far from what the ENEM guiding book states that are being evaluated, they might relate closer to what the neural models are evaluating. For this grader, as we did not optimize any hyperparameters, we joined the training and validation split to increase the number of training samples. As this model outputs a real number and the ENEM grades are in steps of 40, we round it to the closest allowed grade.

For fine-tuning a Large Language Model, we chose to fine-tune a Phi-3-medium model, balancing model size with our available hardware. Although smaller than some other models, it still demands significant computational power for training. We had access to an NVIDIA GPU RTX A6000, and we applied several optimizations for fine-tuning [35], including Gradient Checkpointing, Flash Attention V2 [10], and LoRA [19].

The available model is instruction fine-tuned.[2] During our tests, we found it necessary to pass an instruction to the system role before prompting a given essay. The prompt structure is shown in Fig. 1.

```
<|system|>CONCEPT_SYSTEM<|end|>
<|user|>Grade the following essay: {essay_example}<|end>
<|assistant|>
```

Fig. 1. Prompt-structure used for each Concept fine-tuning through Phi-3.

In this structure, `CONCEPT_SYSTEM` is a placeholder for the competence being fine-tuned (e.g., ENEM Competences that vary from 1 to 5), and `essay_example` represents the essay we are interested in fine-tuning.

To define the concepts, we copied the definitions directly from the ENEM 2023 Student's Reference manual[3], which provides clear instructions on how each grade is applied.

Following these optimizations, we added a sequence classification head to the existing model and fine-tuned it according to the hyperparameters outlined in Table 1. It is noteworthy that these models were instruction fine-tuned through supervised fine-tuning and direct preference optimization for English conversations. Fine-tuning capabilities for languages other than English were not investigated prior to the completion of this study [1]. We found it interesting to notice an ability of the final Phi-3 to generalize and outperform models that were trained solely on Portuguese data—the prior model based on BERTimbau [36].

Finally, we employ Google's Gemini chatbot. We chose Gemini over ChatGPT because Gemini offers a free API. The current model available is the Gemini 1.5 Flash [38]. Its size and training set are not publicly disclosed, so we can only speculate about its dimensions. According to the LMSys Leaderboard [9], it is ranked near the LLama3 70B model, suggesting it has at least 70 billion parameters. We do not train this model; we use it exclusively through its API. Additionally, we did not employ any prompt engineering techniques, instead applying a zero-shot learning process [6,8,23], as we only asked it to grade the essays using the ENEM criteria. However, the model produces grades that do not follow the 40-point step, so we round the numbers according to the same criteria as the Linear Regressor.

[2] Specifically, microsoft/Phi-3-medium-128k-instruct on the Hugging Face Hub: https://huggingface.co/microsoft/Phi-3-medium-128k-instruct.

[3] https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/a_redacao_no_enem_2023_cartilha_do_participante.pdf.

Table 1. Phi-3 Fine Tuning Hyperparameters

Hyperparameter	Value
Warmup steps	10
Gradient Checkpoint	True
Num train epochs	12
Learning rate	5e–5
Weight decay	0.01
BFloat16 use	True
LoRA Attention Dimension	256
LoRA Alpha Scaling	32
LoRA Dropout	0.15
LoRA Task Type	Text Classification
LoRA Target Modules	All Linear

Using these graders, we test them—except for the BERT one, as its performance has already been published in [34]—and compare them against the baselines. Then, we use the weights of the Linear Regression to identify the relevant features. Finally, we use them to create possible universal adversarial attacks that students could create.

There are no limits to the number of adversarial attacks that can be created. In our investigation, we restricted ourselves to a scenario where (1) the students do not know the underlying AES system; (2) they do not have access to any part of the system except for an interface where they can type the essay and see their final score; (3) students can share their intuitions with each other; (4) they do not want to properly write the essay, but still they want the best possible grade. Having these restrictions, we devise some attacks and test them against all graders.

4 Experiments

We now discuss the empirical findings of our investigation.

4.1 Testing the New Models

The first step before devising our adversarial attacks is comparing the performance of an interpretable model against the neural ones. We train the models presented in the previous section, namely a Linear Regressor (LR), a Phi-3 and a Gemini-based grader. Their performances, measured in Quadratic Weighted Kappa (QWK), are presented in Table 2 and compared against the BERTs available and presented in [34].

The first thing to notice is that their sizes have different orders of magnitude, but the performances are still not that distant. The Linear Regressor was

Table 2. Performance comparison of different automatic graders, as measured by QWK.

Model	Size	C1	C2	C3	C4	C5
LR	72	0.23	0.40	0.47	0.34	0.22
Phi-3	14B ($\approx$ 892M Trainable)	0.46	0.35	0.52	0.29	0.61
Gemini	$\geq$70B?	0.41	0.40	0.40	0.36	0.35
BERTs	110M–330M	0.29–0.37	0.23–0.37	0.42–0.50	0.28–0.42	0.26–0.53

expected to have the worst performance across all competences, but it happened only in two out of the five. When it had the lowest performance, it was only by a margin of 0.07 and 0.04 compared to the BERT baseline. The similar performance to the others may point out that the neural models are exploring simple features of the essays.

The Phi-3 model performs best in three competences, with a surprising 0.61 in the fifth competence. Notably, this model loses twice to the Linear Regression model, but it is always at least competitive with the BERT baseline.

Finally, Gemini is the largest model, with an unknown size and training set. It drew with the Linear Regressor for the best performance in Competence 2 and had the worst performance in Competence 3. In general, this model is the most stable one, with performances between 0.41 and 0.35. Its relatively poor performance must be seen with a grain of salt, as it is the only model that was not fine-tuned to actually grade the essays.

Having shown that the performances between the different Transformers-based graders are not that different from the Linear Regressor, we check which features receive the largest weights. The names of the five more important (positive) features are presented in Table 3. Although named with a different pattern, *adverbs* and *verbs* are features that measure the ratio of such parts of speech. *Content words* and *Function words* also account for the ratio of words. The former aggregates nouns, verbs, adjectives, adverbs, and discourse markers. The latter accounts for articles, conjunctions, interjections, numerals, pronouns, prepositions, conjunctive and subordinating adverbs. Related to the previous two, content density is the ratio of *content words* to *functional words*. Finally, *cau neg conn ratio* is the ratio of negative causal connectives.

Competences 1 to 3 have the same set in the same order, which does not match what is expected from the official grading, as each competence is supposed to evaluate a different aspect of the text. Competence 5 has the same set as the previous three but in a different order. Competence 4 is supposed to evaluate the proper usage of connectives, so having *function words* and *cau neg conn ratio* between the top 5 makes sense.

With this information, we can assume that students will come up with intuitions like (1) it is important to have a lot of adverbs, as it is the most important feature in 3 competencies, it is part of the top feature in competence 4 and is the third most important in competence 5; (2) it is important to have a lot

Table 3. The four most important features for each competence according to the Linear Regressor.

Competence 1	Competence 2	Competence 3	Competence 4	Competence 5
adverbs	adverbs	adverbs	content words	adjective ratio
adjective ratio	adjective ratio	adjective ratio	function words	verbs
noun ratio	noun ratio	noun ratio	cau neg conn ratio	adverbs
verbs	verbs	verbs	content density	noun ratio

of adjectives, as it is the most important in one competence, the second most important in three, and part of the top feature in the other.

4.2 Defining the (Universal) Adversarial Attacks

Bearing the previous intuitions in mind, we can devise three goals: increase the number of adverbs, increase the number of adjectives, and increase both. We consider that although the feature is the ratio of such classes, it is not trivial to identify that the importance is the ratio and not the total number.

To increase the number of one of the classes, we can consider some strategies, such as (a) writing a list of words belonging to this class, (b) copying and pasting the previous list so that it becomes similar to the four paragraphs structure; (c) similar to the previous, but with more copying and pasting and using a memorized sentence that uses many of the desired part of speech.

We have 3 goals, and each goal can be reached by 3 strategies, so we can define 9 adversarial attacks:

Attack 1a: We devise a list of adverbs, such as: “Well, badly, enormously, smally, certainly, wrongly, rapidly, slowly, fairly, unfairly”.

Attack 1b: We repeat the previous list four times, one in each paragraph.

Attack 1c: We create a sentence that uses more adverbs than usual, such as: “Undeniably, progressing slowly, leisurely, carefully, silently while deeply breathing and thinking intensively about the given problem”, copy it 10 times in each paragraph, in 4 paragraphs.

Attack 2a: We devise a list of adjectives, such as: “Good, bad, big, small, best, worst, right, wrong, last, first, fair, unfair”.

Attack 2b: We repeat the previous list four times, one in each paragraph.

Attack 2c: We create a sentence that uses more adjectives than usual, such as: “The constant and innovative development of modern technology has brought significant and deep changes to this issue, creating diverse and thrilling opportunities for a more promising and sustainable future.” then we copy it 10 times in each paragraph, in 4 paragraphs.

Attack 3a: We concatenate the sentences of 1a and 2a.

Attack 3b: We repeat the previous list four times, one in each paragraph.

Attack 3c: We create a sentence that uses more adjectives and adverbs than usual, such as: "Consequently and undeniably, the constant innovative development slowly and progressively brought some significant and deep changes to constantly needed problems", then we copy it 10 times in each paragraph, in 4 paragraphs.

4.3 Testing the Attacks

We now use the sentences created for each attack and submit them to the models. The grades assigned by each model are displayed in Table 4. Schools in Brazil usually require students to get at least 6 or 7 to be approved, so we can assume that an adversarial student would be happy with getting a 600 or higher with the lowest possible effort, and this will count as fooling the model.

Table 4. Grades assigned by all models, in all competences, to each attack.

Att.	Model	C1	C2	C3	C4	C5	Total	Att.	C1	C2	C3	C4	C5	Total	Att.	C1	C2	C3	C4	C5	Total
1a	LR	200	200	200	200	0	800	2a	200	200	200	200	0	800	3a	200	200	200	200	0	800
	BERT	120	80	40	120	0	360		120	80	40	120	0	360		120	120	40	120	0	400
	Gemini	120	120	80	120	120	560		120	120	120	120	160	640		120	80	80	120	120	520
	Phi-3	0	40	40	0	0	80		80	120	40	0	0	240		80	40	40	0	0	160
1b	LR	200	200	200	200	0	800	2b	200	200	200	200	0	800	3b	200	200	200	200	0	800
	BERT	120	120	80	120	0	440		120	120	80	120	0	440		120	120	120	120	0	480
	Gemini	80	80	80	80	80	400		80	80	80	80	80	400		80	40	40	40	80	280
	Phi-3	80	120	80	120	0	400		80	120	120	120	0	440		80	120	80	120	0	400
1c	LR	200	200	200	200	200	1000	2c	120	200	200	200	200	920	3c	200	200	200	200	200	1000
	BERT	160	160	160	160	0	640		160	160	160	160	0	640		160	160	160	160	40	680
	Gemini	40	40	0	40	40	160		0	0	0	0	0	0		40	0	0	0	0	40
	Phi-3	80	120	40	80	0	320		80	40	40	80	0	240		80	40	40	80	0	240

We can see that our simplest attacks, the ones with suffix *a*, are sufficient to fool a simple Linear Regressor on all but the Competence 5. This is not surprising, as the attacks were defined according to most correlated predictor variables. Changing from one sentence to four paragraphs (i.e., changing an attack from type *a* to type *b*) did not increase the score, while changing from type *b* to type *c* did improve the score. We note that either changing from type *a* to type *b* or from type *b* to *c* leads to a significant increase of the text length. The number of words and paragraphs are features available to the model and have a positive coefficient, so increasing them naturally increases the grade. Additionally, as this Linear Regressor has a limit of 2k words, it is easy to get a full grade by increasing certain word frequencies.

More interestingly, the BERT grader was not completely fooled by the simplest attacks; it assigned attacks *1a* and *2a* with 360 points, and attack *3a* with 400 points. This suggests that BERT-based predictors are either sensitive to diversity in parts of speech or that their inner representation considers word frequencies. Changing from attack type *a* to *b* increased scores, which could point to the fact that BERT is sensitive to the four paragraph structure—this could be further emphasized by Competence 3's scores going up. But since BERT does not have a token for line break, it is not clear whether that effect is due to the presence of longer sentences or due to the increased word frequency. Finally, we see that going from attacks of type *b* to *c* resulted in an increase of 200 points. This was the only attack that received a score above zero in Competence 5 with this model.

Gemini, on the other hand, seems to recognize attacks that increase essay length. The simplest attacks *1a* and *2a* were assigned scores of 560 and 640 points, respectively, by this model, which are relatively high and way higher than expected. However, the relatively longer essays produced by attacks of type *b* received lower scores, showing that the Gemini is sensitive to repetition. This is further evidenced by the fact that is assigns only 40 points to attack *3c*, which is the one with most repetitions. This model was the only one to identify one attack (*2c*) and it almost identified attack *3c*.

Finally, the Phi-3-based model was the most resilient to the simplest attacks od type *a*. Even though attacks *1a* and *2a* are very similar, the model was more sensitive to the latter one. Attack *3a*, which somehow combines attack *1a* and *2a*) received an intermediary score, suggesting that the model is not composing the representation to produce scores. Phi-3's tokenizer has a token for the line break, which can make it sensitive to the four paragraphs structure. This appears to be the case, since attacks of type *a* received smaller scores than attacks of type *b*. Finally, the attacks of type *c*, which are based on repetition of sentences were not as effective as the attacks of type *b* which are based of repetition of lists. We highlight that Phi-3 achieved the highest QWK score for Competence 5, across all models and competences, and was not fooled by any attack—that is, it scored all attacks with a zero.

Seeing that Competence 5 is the one that most often receives a zero, it is possible to imagine that students would try to target it. The fifth competence is responsible for evaluating the conclusion. Accordingly, we devised the **Attack 4** based on the previous results: a sentence that resembles a conclusion using adjectives and adverbs, namely: "Consequently, it is up to the fair and democratic Federal Government to rapidly approve laws that rapidly reduce the occurrence of these horrendous problems. Following, the dear Brazilian population must abide by the undeniable laws, and the fast police must arrest those that committed any inhuman crime", copied in 7 paragraphs. The grades assigned by each model for this attack are presented in Table 5.

Surprisingly, the Linear Regressor was the only grader not deceived in the fifth competence. Nonetheless, the model as a whole assigned 800 points, a high score. The grade can be further boosted by adding more repetitions and thus achieving a perfect score.

As intended, the BERT grader assigned a high grade in Competence 5, demonstrating that it is not only using word frequency. Still, this attack uses a

Table 5. Grades assigned by each model when grading the Attack 4.

Attack	Model	C1	C2	C3	C4	C5	Total
4	LR	200	200	200	200	0	800
4	BERT	160	160	160	160	160	800
4	Gemini	40	40	0	40	40	160
4	Phi-3	160	120	120	120	40	560

lot of words, which could cause lead to other competences receiving high scores. It is noteworthy that when trained, this grader did not have access to many training instances of perfect grade, which causes it to rarely (if ever) assign 200 points in a competence. Thus, 800 points is the highest possible grade in this method, which means that it was completely deceived by this attack.

If Gemini is sensitive to repetition and we repeat the same sentence 7 times, then we would expect a low grade assigned by this model. That is exactly what happens. The zero assigned in Competence 3 is rounded from 20. In our rounding we always round down, but maybe systems used in classrooms would round up to motivate students. If we had rounded it up, this attack would have been the attack based on repetition that fooled Gemini the most. As an additional test made only for this grader, we tested asking it to grade only the first paragraph of Attack 4; it assigned a total of 440 points.

Finally, this attack was the one in which Phi-3 assigned the highest grades, for the first time assigning points in the fifth competence. Overall, Phi-3 was more resilient than BERT and the Linear Regressor but less than Gemini. This might suggest that being an order of magnitude pays off by having higher-level features. Nonetheless, the grade assigned by this model was surprisingly high.

5 Conclusion

In this work we investigated whether Transformers-based Automatic Essay Scorers are susceptible to universal adversarial attacks that might occur in a classroom setting.

To simulate realistic student behavior, we trained a Linear Regressor based on standard NLP features to predict scores, then used the most important features to craft simple essays that should receive low scores according to grading guidelines yet might fool statistical AES systems. Such essays varied from simple list of adjectives and adverbs to repetitions of template texts unrelated to the required prompt.

We found that the BERT-based grader is susceptible to both long sentences and to the presence of adjectives and adverbs, similarly to reported in the literature [13].

The Gemini-based grader was the only model to completely identify an attack, meaning that it assigned 0 to such essays (as a human grader would). The more repetition an essay had, the lower was the score assigned by that model.

This contrasts with the BERT-based model, which assigned higher scores to longer (and more repetitive) texts. Despite that relative improvement, this type of model still assigned undeserving high grades to most of the attacks. This is aligned with the findings of the ChatGPT-based grader [7], which found that the model was more lenient than a human-grader. It is however interesting to note that this model assigned the highest score to essays consisting only of a list of adjectives or adverbs.

The Phi-3-based model, which is the best performing model for most competences, performed similarly to the other models. In particular, it was more susceptible than Gemini to an attack that included a unrelated but sensible template conclusion to essay. This is in line with findings that fine-tuned models improve performance at the expense of being more susceptible to attacks [40]. Interestingly, although the model was originally trained only texts in English and fine-tuned with texts in Portuguese, it achieved the best performances, even when compared to models trained exclusively with Portuguese texts.

This work shows that even sophisticated AES systems can be fooled by simple attacks, which should be of great concern to real-world systems deployed. Future work should better investigate whether standard techniques such as data augmentation and loss function shaping can mitigate such an effect.

Acknowledgments. The authors of this work would like to thank the Center for Artificial Intelligence (C4AI-USP) and the support from the São Paulo Research Foundation (FAPESP grants #2019/07665-4 and 2022/02937-9) and from the IBM Corporation. This work was also supported by CNPq grant no. 305136/2022-4 and CAPES Finance Code 001.

References

1. Abdin, M., Jacobs, S.A., Awan, A.A., Aneja, J., Awadallah, A., et al.: Phi-3 technical report: a highly capable language model locally on your phone (2024)
2. Amorim, E., Veloso, A.: A multi-aspect analysis of automatic essay scoring for Brazilian Portuguese. In: Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics, pp. 94–102 (2017)
3. Attali, Y., Burstein, J.: Automated essay scoring with e-rater v.2. J. Technol. Learn. Assess. **4**(3) (2006)
4. Bazelato, B.S., Amorim, E.: A bayesian classifier to automatic correction of Portuguese essays. In: Conferência Internacional sobre Informática na Educação (TISE), vol. 18, pp. 779–782 (2013)
5. Beigman Klebanov, B., Madnani, N.: Automated evaluation of writing – 50 years and counting. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7796–7810 (2020)
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., et al.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901. Curran Associates, Inc. (2020)
7. Chang, L.H., Ginter, F.: Automatic short answer grading for finnish with chatgpt. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 23173–23181 (2024)

8. Chang, M.W., Ratinov, L., Roth, D., Srikumar, V.: Importance of semantic representation: dataless classification. In: Proceedings of the 23rd National Conference on Artificial Intelligence, vol. 2, pp. 830–835 (2008)
9. Chiang, W.L., et al.: Chatbot arena: an open platform for evaluating llms by human preference (2024)
10. Dao, T.: Flashattention-2: faster attention with better parallelism and work partitioning (2023)
11. Ebrahimi, J., Rao, A., Lowd, D., Dou, D.: HotFlip: white-box adversarial examples for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, vol. 2: Short Papers, pp. 31–36 (2018)
12. Filho, A.H., Concatto, F., do Prado, H.A., Ferneda, E.: Comparing feature engineering and deep learning methods for automated essay scoring of Brazilian national high school examination. In: International Conference on Enterprise Information Systems (2021)
13. Filighera, A., Ochs, S., Steuer, T., Tregel, T.: Cheating automatic short answer grading with the adversarial usage of adjectives and adverbs. Int. J. Artif. Intell. Educ. **34**(2), 616–646 (2024)
14. Filighera, A., Steuer, T., Rensing, C.: Fooling automatic short answer grading systems. In: Bittencourt, I.I., Cukurova, M., Muldner, K., Luckin, R., Millán, E. (eds.) AIED 2020. LNCS (LNAI), vol. 12163, pp. 177–190. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-52237-7_15
15. Fonseca, E.R., Medeiros, I., Kamikawachi, D., Bokan, A.: Automatically grading Brazilian student essays. In: Proceedings of International Conference on Computational Processing of the Portuguese Language, pp. 170–179 (2018)
16. Gao, H., Oates, T.: Universal adversarial perturbation for text classification. arXiv preprint arXiv:1910.04618 (2019)
17. Gunasekar, S., et al.: Textbooks are all you need (2023)
18. Hamner, B., Morgan, J., Lynnvandev, M., Ark, T.: The hewlett foundation: automated essay scoring. Kaggle (2012)
19. Hu, E.J., et al.: Lora: low-rank adaptation of large language models (2021)
20. Klein, H., Levenburg, N., McKendall, M., Mothersell, W.: Cheating during the college years: how do business students compare? J. Bus. Ethics **72**, 197–206 (2007)
21. Kumar, Y., Bhatia, M., Kabra, A., Li, J.J., Jin, D., Shah, R.R.: Calling out bluff: attacking the robustness of automatic scoring systems with simple adversarial testing. CoRR arxiv:2007.06796 (2020)
22. Kumar, Y., Parekh, S., Singh, S., Li, J.J., Shah, R.R., Chen, C.: Automatic essay scoring systems are both overstable and oversensitive: explaining why and proposing defenses. Dial. Discour. **14**(1), 1–33 (2023)
23. Larochelle, H., Erhan, D., Bengio, Y.: Zero-data learning of new tasks. In: Proceedings of the 23rd National Conference on Artificial Intelligence, vol. 2, pp. 646–651 (2008)
24. Leal, S.E., Duran, M.S., Scarton, C.E., Hartmann, N.S., Aluísio, S.M.: Nilc-metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. Lang. Res. Eval. **58**(1), 73–110 (2024)
25. Li, L., Ma, R., Guo, Q., Xue, X., Qiu, X.: BERT-ATTACK: adversarial attack against BERT using BERT. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pp. 6193–6202 (2020)
26. Lim, C.T., Bong, C.H., Wong, W.S., Lee, N.K.: A comprehensive review of automated essay scoring (aes) research and development. Pertanika J. Sci. Technol. **29**(3), 1875–1899 (2021)

27. Liu, J., Xu, Y., Zhu, Y.: Automated essay scoring based on two-stage learning (2019)
28. Marinho, J., Anchiêta, R., Moura, R.: Essay-br: a Brazilian corpus of essays. In: Anais do III Dataset Showcase Workshop, pp. 53–64 (2021)
29. Marinho, J., Cordeiro, F., Anchiêta, R., Moura, R.: Automated essay scoring: an approach based on enem competencies. In: Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional, pp. 49–60 (2022)
30. Mizumoto, A., Eguchi, M.: Exploring the potential of using an AI language model for automated essay scoring. Res. Methods Appl. Linguist. **2**(2), 100050 (2023)
31. Nvidia, Adler, B., Agarwal, N., Aithal, A., Anh, D.H., et al.: Nemotron-4 340b technical report (2024)
32. Page, E.B.: The imminence of... grading essays by computer. In: The Phi Delta Kappan, pp. 238–243 (1966)
33. Pedregosa, F., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
34. Silveira, I.C., Barbosa, A., Mauá, D.D.: A new benchmark for automatic essay scoring in Portuguese. In: Proceedings of the 16th International Conference on Computational Processing of Portuguese, vol. 1. pp. 228–237 (2024)
35. Singh, A., Pandey, N., Shirgaonkar, A., Manoj, P., Aski, V.: A study of optimizations for fine-tuning large language models (2024)
36. Souza, F., Nogueira, R., Lotufo, R.: BERTimbau: pretrained BERT models for Brazilian Portuguese. In: Cerri, R., Prati, R.C. (eds.) BRACIS 2020. LNCS (LNAI), vol. 12319, pp. 403–417. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-61377-8_28
37. Taghipour, K., Ng, H.T.: A neural approach to automated essay scoring. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1882–1891. Association for Computational Linguistics (2016)
38. Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., et al.: Gemini 1.5: unlocking multimodal understanding across millions of tokens of context (2024)
39. Touvron, H., et al.: Llama: open and efficient foundation language models (2023)
40. Wangkriangkri, P., Viboonlarp, C., Rutherford, A.T., Chuangsuwanich, E.: A comparative study of pretrained language models for automated essay scoring with adversarial inputs. In: IEEE Region 10 International Conference TENCON, pp. 875–880 (2020)
41. Whitley, B.E.: Factors associated with cheating among college students: a review. Res. High. Educ. **39**(3), 235–274 (1998)
42. Woods, B., Adamson, D., Miel, S., Mayfield, E.: Formative essay feedback using predictive scoring models. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2071–2080 (2017)
43. Zhang, H., Li, L.H., Meng, T., Chang, K.W., Van den Broeck, G.: On the paradox of learning to reason from data. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, pp. 3365–3373 (2023)
44. Zhang, J., Wang, J., Luo, X., Ma, B., Xiong, N.: Imperceptible and reliable adversarial attack. In: Cao, C., Zhang, Y., Hong, Y., Wang, D. (eds.) FCS 2021. CCIS, vol. 1558, pp. 49–62. Springer, Singapore (2022). https://doi.org/10.1007/978-981-19-0523-0_4