Identifying Vulnerabilities and Fortifying Defenses in Automated Short-Answer Grading

Anonymous submission

Abstract

Exploitable weaknesses in automated scoring models can greatly impact the responsible use of AI systems in education. This paper identifies several potential adversarial attacks and defense strategies in the context of a transformer-based short-answer scoring system intended for use in medical education. Attacks were designed to resemble tactics that test takers might employ when they are not certain of the correct answer to a given test question. Initial results corroborate previous findings on the susceptibility of transformer-based grading systems to such attacks and show that adversarial training can significantly improve a system's robustness.

Introduction

Vulnerabilities in automated scoring models may allow test takers to exploit a model's behavior to manipulate assessment outcomes. The consequences of such manipulations range from unmerited credit to the erosion of trust in automated grading (Filighera et al. 2023) and thus have important implications for the responsible use of AI systems in educational assessment, especially when such assessments are used to make high-stakes decisions.

The Automated Short Answer Grading (ASAG) literature distinguishes two scoring approaches: an *instance-based method* where a system predicts scores for new responses after training on a portion of the data, and a *similarity-based approach* that assigns new responses the label of a matched annotated response (Bexte, Horbach, and Zesch 2022). With the advent of the transformer model, neural similarity-based ASAG techniques have shown promise by yielding improved accuracy, interpretability, and reduced data annotation requirements. However, these advancements also bring new challenges, making the system susceptible to adversarial attacks such as categorizing random strings of letters as a correct answer (Ding et al. 2020) or misclassifying answers that include specific irrelevant words (Filighera et al. 2023).

This paper identifies several potential exploitations (also known as "gaming strategies") and defensive countermeasures applicable in the context of Automated Short Answer Grading (ASAG) in medical education. Experiments were undertaken using the ACTA system (Analysis of Clinical Text for Assessment; (Suen et al. 2023)), a transformer-based ASAG system designed to classify short responses to medical questions as correct or incorrect (examples of such short responses include "plantar fasciitis", "dermatomyositis", or "Administer corticosteroids then do arterial biopsys"). To achieve this, ACTA utilizes sentence BERT (Reimers and Gurevych 2019) and contrastive learning. When presented with a new response, ACTA matches it to the most similar response within a known training set of human-scored responses and assigns it the matched response's label (correct or incorrect), provided their similarity exceeds a given operational threshold (for a detailed description of the ACTA system see (Suen et al. 2023)). ACTA achieves near human-level performance with a binary F1 score of .98; however, previously reported weaknesses of transformer-based grading systems require the examination of ACTA's susceptibility to gaming. To our knowledge, this is the first investigation of adversarial attacks for ASAG in the context of medical education.

Data

The dataset comprises 71 short-answer questions (SAQs) with 36,735 responses from 24,235 examinees. An example of a SAQ is presented in Table 1. Responses were collected during the administration of Medicine Clinical Science subject exam developed at *anonymous institution* and distributed to a large number of medical schools in the US and Canada to use as a subject exam at the end of a semester.

Gaming Strategies

We simulate three gaming strategies meant to resemble how students *without* the requisite knowledge of a correct answer might nevertheless respond to an item. Each item includes a description of a clinical scenario and the first strategy simulates responses by randomly sampling words (excluding stop words) from a given item's description. Variations of this strategy include consecutive words, non-consecutive words, and samples of words that appear in both the item description *and* a generic list of medical terms. The second strategy uses of a summary of the clinical scenario as a response. Summaries were obtained using chatGPT. Finally, the third strategy uses "mixed" responses that combine both correct and plausible incorrect answers into a single response, which, following operational guidelines, should be scored as incorrect.

These strategies – particularly strategy 1 – have the potential to generate an impractically large number of responses.

A previously healthy 26-year-old man is brought to the emergency department because of a tingling sensation in his fingers and toes for 3 days and progressive weakness of his legs. He had an upper respiratory tract infection 2 weeks ago. He has not traveled recently. He was unable to get up from bed this morning and called the ambulance. Temperature is 37.3°C (99.1°F), pulse is 110/min, respirations are 22/min, and blood pressure is 128/82 mm Hg. Pulse oximetry on room air shows an oxygen saturation of 99%. Physical examination shows weakness of all four extremities in flexion and extension; this weakness is increased in the distal compared with the proximal muscle groups. Deep tendon reflexes are absent throughout. Sensation is mildly decreased over both feet. What is the most likely diagnosis?
Sample of correct answers: Guillain-Barré syndrome, acute immune-mediated polyneuropathy

Table 1: An example of a practice SAQ item

Gaming Strategy	FPR Before Adv Training	FPR Adv Training 1	FPR Adv Training 2
Information from the Stem	.061	.006	.048
Clinical Case Summary	.189	0	.016
Mixed Responses	.435	.021	.246

Table 2: False positive rates for the gaming responses before and after adversarial training

To create a set of responses that could be feasibly used as part of an operational process, we randomly sample 5% from each strategy, resulting in 14,657, 573, and 584 simulated responses for strategies 1, 2, and 3. While simulated responses were largely nonce phrases or unequivocally incorrect, 3 simulated responses matched (real) correct responses from the training data. Three misclassifications was deemed tolerable for our purpose thereby allowing us to designate *all* artificial responses as incorrect.

Experiments and Results

We began by evaluating the effect of gaming attempts on ACTA prior to any adversarial training. The model was trained on 70% (26,095) of the real responses and evaluated on the remaining 30% (10, 890) combined with all artificial responses. Since the number of artificial responses may vary across strategies and experiments, we report two separate measures: F1 for real responses and false positive rate (FPR) for artificial responses. ACTA performed well when scoring real data (F1 = .9845); however, the gaming strategies deceived ACTA into misclassifying many of the artificial responses as "correct." FPRs for strategies 1, 2, and 3 were .061, .189, and .435, respectively, demonstrating the vulnerability of this system to examinee gaming (Table 2). Responses from strategy 3 were especially challenging to classify correctly, illustrating the potential for examinees with partial knowledge to game systems that have not been adversarially trained by simply listing as many plausible answers as possible.

Adversarial Training 1: Next, we added 70% of the simulated responses from each strategy to the training data, leaving 30% of the artificial and 30% of the real responses in the test set. Here, F1 remains high (.9818) but FPRs are reduced substantially to .006, 0, and .021, for the gaming strategies demonstrating the potential efficacy of data augmentation. Results for strategy 3 are particularly encouraging, showing that training on combinations of responses generated in advance is an effective countermeasures against the most successful gaming strategy.

Adversarial Training 2: Finally, to evaluate whether data from one strategy can help against tactics that are unknown at the time of training, we perform a 3-fold cross-validation by training on two gaming strategies and evaluating the third. Under these conditions, F1 remains at .98 for the real responses and the FPRs are .048, .016, and .246 for the simulated responses for strategies 1, 2, and 3, respectively.

Discussion

These results add new evidence related to exploitable vulnerabilities in transformer-based grading systems. Despite being artificially generated approximations of potential gaming behaviors, all three gaming strategies were successful in deceiving the non-adversarially trained system. The first group of adversarial training experiments showed that data augmentation is a promising way to fortify ASAG systems against such attacks and the cross-validation experiments showed that it is also beneficial to train on examples across gaming strategies. The substantial decrease in FPR suggests that a transfer of learning occurs between strategies, which holds significant potential to protect against any unforeseen gaming tactic that arises in practice.

Like many other products, automated scoring tools are socio-technical systems, whose impact is determined not solely by their technical capabilities but also by their use or, in some cases, misuse. As the understanding of possible gaming strategies in the context of medical education matures, future work will include the simulation of new adversarial attacks for ASAG systems that are more closely aligned with human behaviors as well as further experimentation with adversarial training techniques. Other crucial fairness evaluation aspects include analyses of scoring bias such as differential functioning of the system for users with different backgrounds, as well as the extent to which human and machine scores are interchangeable.

References

Bexte, M.; Horbach, A.; and Zesch, T. 2022. Similarity-based content scoring-how to make S-BERT keep up with BERT. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, 118–123.

Ding, Y.; Riordan, B.; Horbach, A.; Cahill, A.; and Zesch, T. 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*, 882–892.

Filighera, A.; Ochs, S.; Steuer, T.; and Tregel, T. 2023. Cheating Automatic Short Answer Grading with the Adversarial Usage of Adjectives and Adverbs. *International Journal of Artificial Intelligence in Education*, 1–31.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Suen, K. Y.; Yaneva, V.; Mee, J.; Zhou, Y.; Harik, P.; et al. 2023. ACTA: Short-Answer Grading in High-Stakes Medical Exams. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, 443–447.