Dialogue & Discourse 14(1) (2023) 1–32 doi: 10.5210/dad.2023.101

Automatic Essay Scoring Systems Are Both Overstable And Oversensitive: Explaining Why And Proposing Defenses

Yaman Kumar Singla[*] YAMANK@IIITD.AC.IN
Adobe Media Data Science Research, IIIT-Delhi, SUNY at Buffalo

Swapnil Parekh[*] SWAPNIL.PAREKH@NYU.EDU
New York University

Somesh Singh[*] F20180175@GOA.BITS-PILANI.AC.IN
IIIT-Delhi

Junyi Jessy Li JESSY@AUSTIN.UTEXAS.EDU
University of Texas at Austin

Rajiv Ratn Shah RAJIVRATN@IIITD.AC.IN
IIIT-Delhi

Changyou Chen CHANGYOU@BUFFALO.EDU
SUNY at Buffalo

Editor: Manfred Stede

Submitted 09/2021; Accepted 01/2023; Published online 04/2023

Abstract

Deep-learning based Automatic Essay Scoring (AES) systems are being actively used in various high-stake applications in education and testing. However, little research has been put to understand and interpret the black-box nature of deep-learning based scoring algorithms. While previous studies indicate that scoring models can be easily fooled, in this paper, we explore the reason behind their surprising adversarial brittleness. We utilize recent advances in interpretability to find the extent to which features such as coherence, content, vocabulary, and relevance are important for automated scoring mechanisms. We use this to investigate the oversensitivity (*i.e.*, large change in output score with a little change in input essay content) and overstability (*i.e.*, little change in output scores with large changes in input essay content) of AES. Our results indicate that autoscoring models, despite getting trained as "end-to-end" models with rich contextual embeddings such as BERT, behave like bag-of-words models. A few words determine the essay score without the requirement of any context making the model largely *overstable*. This is in stark contrast to recent probing studies on pre-trained representation learning models, which show that rich linguistic features such as parts-of-speech and morphology are encoded by them. Further, we also find that the models have learnt dataset biases, making them *oversensitive*. The presence of a few words with high co-occurrence with a certain score class makes the model associate the essay sample with that score. This causes score changes in $\sim$95% of samples with an addition of only a few words. To deal with these issues, we propose detection-based protection models that can detect oversensitivity and samples causing overstability with high accuracies. We find that our proposed models are able to detect unusual attribution patterns and flag adversarial samples successfully.

Keywords: Interpretability in AI, Automatic Essay Scoring, AI in Education

[*] Equal Contribution

©2023 Yaman Kumar Singla, Swapnil Parekh, Somesh Singh, Junyi Jessy Li, Rajiv Ratn Shah, Changyou Chen

This is an open-access article distributed under the terms of a Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/).

1. Introduction

Automatic Essay Scoring (AES) systems are used in diverse settings such as to alleviate the workload of teachers, save time and costs associated with grading, and to decide admissions to universities and institutions. On average, a British teacher spends five hours in a calendar week scoring exams and assignments (Micklewright et al., 2014). This figure is even higher for developing and low-resource countries where the teacher to student ratio is dismal. While on the one hand, autograding systems effectively reduce this burden, allowing more working hours for teaching activities, on the other, there have been many complaints against these systems for not scoring the way they are supposed to (Feathers, 2019; Smith, 2018; Greene, 2018; Mid-Day, 2017; Perelman et al., 2014b).

Test questions on standardized tests elicit *persuasive* and *informative* writing with specific discourse structure. While in persuasive writing, students write their opinions about a topic and try to validate them using *convincing* arguments, informative writing is more *descriptive* and requires students to state their experiences to substantiate their opinions. Both of them adhere to strict discourse strategies (Burstein et al., 2003) which includes an introduction, thesis statements, main and supporting ideas, and finally a conclusion. Several research studies have investigated how finding and scoring discourse from essays helps to provide a better holistic score to essays (McNamara et al., 2014; Graesser and McNamara, 2011; Burstein et al., 2001; Nadeem et al., 2019; Burstein et al., 1998). At the same time, both research studies and empirical evidence have suggested that AES models have repeatedly failed to score discourse and other features important for scoring. For instance, on the recently released automatic scoring system for the state of Utah, students scored lower by writing question-relevant keywords but higher by including unrelated content (Feathers, 2019; Smith, 2018). Similarly, it has been a common complaint that AES systems focus unjustifiably on obscure and difficult vocabulary (Perelman et al., 2014a). While earlier, each score generated by the AI systems was verified by an expert human rater, it is concerning to see that now many of them are scoring independently without any intervention by human experts (O'Donnell, 2020; Singla et al., 2022a). The concerns are further alleviated by the fact that the scores awarded by such systems are used in life-changing decisions ranging from college and job applications to visa approvals (ETS, 2020b; Educational Testing Association, 2019; USBE, 2020; Institute, 2020).

Traditionally, autograding systems are built using manually crafted features used with machine learning based models (Kumar et al., 2019; Bamdev et al., 2022). Lately, these systems have been shifting to deep learning based models (Ke and Ng, 2019). For instance, many companies have started scoring candidates using deep learning based automatic scoring (SLTI-SOPI, 2021; Assessment, 2021; Duolingo, 2021; LaFlair and Settles, 2019; Yu et al., 2015; Chen et al., 2018; Singla et al., 2021; Riordan et al., 2017; Pearson, 2019). However, there are very few research studies on the reliability[1] and validity[2] of ML-based AES systems. More specifically, we have tried to address the problems of robustness and validity which plague deep learning based black-box AES models. Simply measuring test set performance may mean that the model is right for the wrong reasons. Hence, much research is required to understand the scoring algorithms used by AES models and to validate them on linguistic and testing criteria. Similar opinions are expressed by Madnani and Cahill (2018) in their position paper on automatic scoring systems.

[1] A reliable measure is one that measures a construct consistently across time, individuals, and situations (Ramanarayanan et al., 2020)

[2] A valid measure is one that measures what it is intended to measure (Ramanarayanan et al., 2020)

With this in view, in this paper, we make the following contributions towards understanding current AES systems:

1) Several research studies have shown that essay scoring models are *overstable* (Yoon et al., 2018; Powers et al., 2002; Kumar et al., 2020; Feng et al., 2018). Even large changes in essay content do not lead to significant change in scores. For instance, Kumar et al. (2020) showed that even after changing 20% words of an essay, the scores do not change much. We extend this line of work by addressing *why* the models are overstable. Extending these studies further (§4.1), we investigate AES overstability from the perspective of discourse, coherence, facts, vocabulary, length, grammar and word choice. We do this by using integrated gradients (§3.2), where we find and visualize the most important words for scoring an essay (Sundararajan et al., 2017). We find that the models despite using rich contextual embeddings and deep learning architectures, are essentially behaving as bag-of-words models. Further, we develop models through which we are able to improve the adversarial attack strength (§4.1.2). For example, for memory networks scoring model (Zhao et al., 2017), we delete 40% words from essays without significantly changing score (<1%), whereas Kumar et al. (2020) observed that deleting a similar number of words resulted in a decrease of 20% scores for the same model.

2) While there has been much work on AES overstability (Kumar et al., 2020; Perelman, 2014; Powers et al., 2001), there has been little work on AES oversensitivity. Building on this research gap, by using adversarial triggers, we find that the AES models are also oversensitive, *i.e.*, small changes in an essay can lead to large change in scores (§4.2). We find that, by just adding 3 words in an essay containing 350 words (<1% change), we are able to change the predicted score by 50% (absolute). We explain the oversensitivity of AES systems using integrated gradients (Sundararajan et al., 2017), a principled tool to discover the importance of parts of an input. The results show that the trigger words added to an essay get unusually high attribution. Additionally, we find the trigger words have usually high co-occurrence with certain score labels, thus indicating that the models are relying on spurious correlations causing them to be oversensitive (§4.2.4). We validate both the oversensitive and overstable sample in a human study (§5). We ask the annotators whether the scores given by AES models are right by providing them with both original and modified essay responses and scores.

3) While much previous research in the linguistic field studies how essay scoring systems can be fooled, for the first time, we propose models that can detect samples causing overstability and oversensitivity (Pham et al., 2021; Kumar et al., 2020; Perelman, 2020). Our models are able to detect both overstability and samples causing oversensitivity with high accuracies (>90% in most cases) (§6). Also, for the first time in the literature, through these solutions, we propose a simple yet effective solution for universal adversarial peturbation (§6.1). These models, apart from defending AES systems against samples causing oversensitivity and overstability, can also inform effective human intervention strategy. For instance, AES deployments either completely rely on double scoring essay samples (human and machine) or solely on machine ratings alone (ETS, 2020a; Singla et al., 2022a). With the developed model, AES deployments can choose to have an effective middle ground by selecting samples for human testing and intervention more effectively. Public school systems, *e.g.*, in Ohio which use automatic scoring without any human interventions can select samples using these models for limited human intervention (O'Donnell, 2020; Institute, 2020). For this, we also conduct a small-scale pilot study on the AES deployment of a major language testing company proving the efficacy of the system (§6.3). Previous solutions for human interventions optimization rely on brittle features such as number of words and content modeling approaches like off-topic

detection (Yoon et al., 2018; Yoon and Zechner, 2017). These models cannot detect adversarial samples like the ones we present in our work.

We perform our experiments for three model architectures and eight unique prompts[3], demonstrating the results on twenty-four unique model-dataset pairs. It is worth noting that our goal in this paper is *not* to argue against AES systems and their applications. Rather, **our goal is to interpret how deep-learning based scoring models score essays, why they are overstable and oversensitive, and how to solve the problems of oversensitivity and overstability.** We release all our code, dataset and tools for public use with the hope that it will spur testing and validation of AES models.

2. Related Work

The related work for our work can be chiefly divided into two streams: standardized testing and automatic scoring, and testing and validation of the automated scoring models developed.

Automatic essay scoring: The education testing research community argues for using construct-response (CR) based testing in high-stakes scenarios (Higgins et al., 2011). Well-known tests such as TOEFL, GRE, ACT, LinguaSkill, Duolingo English Test, and SAT are some examples of CR based standardized testing, where the tests present "naturalistic" prompts such as writing essays and summarizing a news report (such as in GRE, TOEFL, SAT), and making a conversation and giving a speech (such as in Duolingo English Test, TOEFL, LinguaSkill) as opposed to artificial tasks like solving multiple choice questions or filling blanks. The community believes that these types of prompts are much more likely to be encountered by the candidate in real-life scenarios (Stiggins, 1982; Charney, 1984; Messick, 1996) and that they provide a more *accurate* way of measuring test construct (Moran, 1987; Wiggins, 1991). Moreover, artificial testing like those of multiple choice questions lead to washback effects (refers to the extent to which the introduction and use of a test influences language teachers and learners to do things they would not otherwise do that promote or inhibit language learning) (Messick, 1996; Wiggins, 1991).

Due to the superiority of CR based testing, they have become a popular means of testing candidates in high-stakes scenarios (such as those of job and visa interviews and college admissions). This poses a significant challenge for language testing and natural language processing communities since CR based testing typically tests on a variety of skills incorporating syntax, semantics, and particularly discourse and organization and by its very nature is costlier than scoring MCQs. By automating this scoring, testing companies reduce the costs associated with training raters, scoring samples, monitoring quality, and also reduce the time to get scores.

Almost all the auto-scoring models are learning-based and treat the task of scoring as a supervised learning task (Ke and Ng, 2019; Ormerod et al., 2021) with a few using reinforcement learning (Wang et al., 2018) and semi-supervised learning (Chen et al., 2010). While the earlier models relied on ML algorithms and hand-crafted rules (Page, 1966; Faulkner, 2014; Kumar et al., 2019; Persing et al., 2010), lately the systems are shifting to deep learning algorithms (Taghipour and Ng, 2016; Grover et al., 2020; Dong and Zhang, 2016; Uto et al., 2020). Approaches have ranged from finding the hierarchical structure of documents (Dong and Zhang, 2016), using attention over words (Dong et al., 2017), multi-stage pretraining (Song et al., 2020), and modelling coherence (Tay et al., 2018).

[3] Here a prompt denotes an instance of a unique question asked to test-takers for eliciting their opinions and answers in an exam. The prompts can come from varied domains including literature, science, logic and society. The responses to a prompt indicate the creative, literary, argumentative, narrative, and scientific aptitude of candidates and are judged on a pre-determined score scale.

In this paper, we interpret and test the recent state-of-the-art scoring models which have shown the best performance on public datasets (Tay et al., 2018; Zhao et al., 2017).

AES testing and validation: Due to the high-stakes nature of the tests, if the AES models are not validated for their adherence to test objectives, they may drive the students to use unethical ways to game the system by addressing the tasks in superficial and construct-irrelevant manner. However, while automatic scoring has seen much research in the recent years, model validation and testing still lag in the ML field with only a few contemporary works (Kumar et al., 2020; Pham et al., 2021; Yoon and Xie, 2014; Malinin et al., 2017). Kumar et al. (2020) and Pham et al. (2021) show that AES systems are adversarially unsecure. Pham et al. (2021) also try adversarial training and obtain no significant improvements. Yoon and Xie (2014) and Malinin et al. (2017) model uncertainty in automatic scoring systems.

Most of the scoring model validation work is in the language testing field, which unfortunately has limited AI-expertise (Litman et al., 2018). Due to this, studies have noted that the results there are often conflicting in nature (Powers et al., 2001, 2002; Bejar et al., 2013, 2014; Perelman, 2020). Powers et al. (2002) asked 27 specialists and generalists to write essays that could produce significant deviations with respect to scores from ETS's *e-rater*. The winner entry repeated the same paragraph 37 times hence showing that repetition over prompt-related keywords makes the scores given by AES unreliable. Perelman et al. (2014a) made software that takes in five keywords and produces semantic garbage written in a difficult and obscure language. They tested it out with the ETS's system and produced high scores, thus concluding that the essay writing system learns to recognize obscure language with difficult and nonmeaningful words and phrases like, *'fundamental drone of humanity'*, *'auguring commencements, torpor of library'* and *'personal disenfranchisement for the exposition we accumulate conjectures'* [4].

In this work, we do a *systematic* analysis of AES models on features important for scoring and try to interpret the mechanism followed by AES systems for both original and perturbed samples. Through this, we discover the overstability and oversensitivity of the AES models and investigate the possible reasons behind their behavior. We also propose several defense mechanisms to solve these problems of the AES models.

3. Background

3.1 Task, Models and Dataset

We use the widely cited ASAP-AES (2012) dataset which comes from Kaggle Automated Student Assessment Prize (ASAP) for the evaluation of automatic essay scoring systems. The ASAP-AES dataset has been used for automatically scoring essay responses by many research studies (Taghipour and Ng, 2016; EASE, 2013; Tay et al., 2018). It is one of the largest publicly available datasets (Table 1). The questions covered by the dataset span many different areas such as Sciences and English. The responses were written by high school students and were subsequently double-scored.

We test the following two state-of-the-art models in this work: *SkipFlow* (Tay et al., 2018) and Memory Augmented Neural Network (*MANN*) (Zhao et al., 2017). Further, for comparison, we design a BERT based automatic scoring model. The performance is measured using Quadratic Weighted Kappa (QWK) metric, which indicates the agreement between a model's and the expert

[4] Generated by giving the keywords, 'Library', 'Delhi' and 'College' respectively.

Prompt Number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
#Responses	1783	1800	1726	1772	1805	1800	1569	723
Score Range	2-12	1-6	0-3	0-3	0-4	0-4	0-30	0-60
#Avg words per response	350	350	150	150	150	150	250	650
#Avg sentences per response	23	20	6	4	7	8	12	35
Type	Argumentative	Argumentative	RC	RC	RC	RC	Narrative	Narrative

Table 1: Overview of the ASAP AES Dataset used for evaluation of AS systems. (RC = Reading Comprehension).

human rater's scores. All models show an improvement of 4-5% over the previous models on the QWK metric. The analysis of these models, especially BERT, is interesting in light of recent studies indicating that pretrained language models learn rich linguistic features including morphology, parts-of-speech, word-length, noun-verb agreement, coherence, and language delivery (Conneau et al., 2018; Hewitt and Manning, 2019; Singla et al., 2022c). This has resulted in pushing the envelope for many NLP applications. The individual models we use are briefly explained as follows:

SkipFlow Tay et al. (2018) model essay scoring as a regression task. They utilize Glove embeddings for representing the tokens. SkipFlow captures coherence, flow and semantic relatedness over time, which the authors call neural coherence features. Due to the intelligent modelling, it gets an impressive average quadratic weighted kappa score of 0.764. SkipFlow is one of the top performing models (Tay et al., 2018; Ke and Ng, 2019) for AES.

MANN Zhao et al. (2017) use memory networks for autoscoring by selecting some responses for each grade. These responses are stored in memory and then used for scoring ungraded responses. The memory component helps to characterize the various score levels similar to what a rubric does. They show an excellent agreement score of 0.78 average QWK outperforming the previous state-of-the-art models.

BERT-based We also design a BERT-based architecture for scoring essays. It utilizes BERT embeddings (Devlin et al., 2019) to represent essays by passing tokens through the BERT Encoder. The CLS token embedding from the last layer is passed through a fully connected layer of size 1 to produce the score. The network was trained to predict the essay scores by minimizing the mean squared error loss. It achieves an average QWK score of 0.74. We utilize this architecture as a baseline representative of transformer-based embedding models.

3.2 Attribution Mechanism

The task of attributing a score $F(x)$ given by an AES model F, on an input essay x can be formally defined as producing attributions $a_1, .., a_n$ corresponding to the words $w_1, .., w_n$ contained in the essay x. The attributions produced are such that[5] $Sum(a_1, .., a_n) = F(x)$, *i.e.,* net attributions of all words ($Sum(a_1, .., a_n)$) equal the assigned score ($F(x)$). In a way, if F is a regression based model, $a_1, .., a_n$ can be thought of as the scores of each word of that essay, which sum to produce the final score, $F(x)$.

We use a path-based attribution method, Integrated Gradients (IGs) (Sundararajan et al., 2017), much like other interpretability mechanisms such as (Ribeiro et al., 2016; Lundberg and Lee, 2017) for getting the attributions for each of the trained models, F. IGs employ the following method to

[5] Proposition 1 in (Sundararajan et al., 2017)

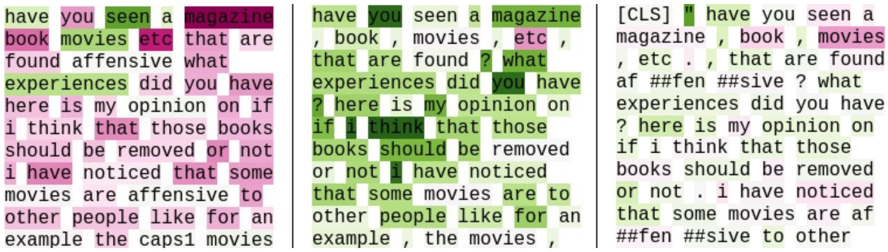
find blame assignments: given an input x and a baseline b[6], the integrated gradient along the i^{th} dimension is defined as:

$$IG_i(x,b) = (x_i - b_i)\int_{\alpha=0}^{1} \frac{\partial F(b+\alpha(x-b))}{\partial x_i}\, d\alpha \tag{1}$$

where $\frac{\partial F(x)}{\partial x_i}$ represents the gradient of F along the i^{th} dimension of x.

We choose the baseline as empty input (all 0s) for essay scoring models since an empty essay should get a score of 0 as per the scoring rubrics. It is the neutral input that models the absence of a cause of any score, thus getting a zero score. Since we want to see the effect of only words on the score, any additional inputs (such as memory in MANN) of the baseline b is set to be that of x[7]. See Fig. 1 for an example. In all our IG diagrams, green highlighting indicates positive attribution while red highlighting indicates negative attribution.

We choose IGs over other explainability techniques since they have many desirable properties that make them useful for this task. For instance, the attributions sum to the score of an essay ($Sum(a_1,..,a_n) = F(x)$), they are implementation invariant, do not require any model to be re-trained and are readily implementable. Previous literature such as (Mudrakarta et al., 2018) also uses Integrated Gradients for explaining the undersensitivity of factoid-based question-answer (QA) models. Other interpretability mechanisms like attention require changes in the tested model and are not post-hoc, thus are not a good choice for our task.



have you seen a magazine
book movies etc that are
found affensive what
experiences did you have
here is my opinion on if
i think that those books
should be removed or not
i have noticed that some
movies are affensive to
other people like for an
example the caps1 movies

have you seen a magazine
, book , movies , etc ,
that are found ? what
experiences did you have
? here is my opinion on
if i think that those
books should be removed
or not i have noticed
that some movies are to
other people like for an
example , the movies ,

[CLS] " have you seen a
magazine , book , movies
, etc . , that are found
af ##fen ##sive ? what
experiences did you have
? here is my opinion on
if i think that those
books should be removed
or not . i have noticed
that some movies are af
##fen ##sive to other

Figure 1: Attributions for SkipFlow, MANN and BERT models respectively of an essay sample for Prompt 2. Prompt 2 asks candidates to write an essay to a newspaper reflecting their views on censorship in libraries and express their views if they believe that materials, such as books, *etc.*, should be removed from the shelves if they are found offensive. This essay scored 3 out of 6.

4. Empirical Studies and Results

We perform our overstability (§4.1) and oversensitivity (§4.2) experiments with 100 samples per prompt for the three models discussed in Section 3.1. There are 8 prompt-level datasets in the overall ASAP-AES dataset, therefore we perform our analysis on 24 unique model-dataset pairs, each containing over 100 samples.

[6] Defined as an input containing absence of cause for the output of a model; also called neutral input (Shrikumar et al., 2016; Sundararajan et al., 2017).

[7] We ensure that IGs are within the acceptable error margin of <5%, where the error is calculated by the property that the attributions' sum should be equal to the difference between the probabilities of the input and the baseline. IG parameters: Number of Repetitions = 20-50, Internal Batch Size = 20-50

4.1 AES Overstability

We first present results on model overstability. Following the previous studies, we test the models' overstability on different features important for AES scoring such as the knowledge of discourse (§4.1.1, 4.1.2), coherence (§4.1.3), facts (§4.1.5), vocabulary (§4.1.4), length (§4.1.2), meaning (§4.1.3), and grammar (§4.1.4). This set of features provides an exhaustive coverage of all features important for scoring essays (Yan et al., 2020).

4.1.1 ATTRIBUTION OF ORIGINAL SAMPLES

We take the original human-written essays from the ASAP-AES dataset and do a word-level attribution of scores. Fig. 1 shows the attributions of all models for an essay sample from Prompt 2. We observe that SkipFlow does not attribute any word after the first few lines (first 30% essay content) of the essay, while MANN attributions are spread over the complete length of the essay. For the BERT-based model, we see that most of the attributions are over nonlinguistic features (tokens) like *'CLS'* and *'SEP'*. *CLS* and *SEP* tokens are used as delimiters in the BERT model. A similar result was also observed by Kovaleva et al. (2019).

For SkipFlow, we observe that if a word is negatively attributed at a certain position in an essay sample, it is then commonly negatively attributed in its other occurrences as well. For instance, *books*, *magazines* were negatively attributed in all its occurrences while *materials*, *censored* were positively attributed and *library* was not attributed at all. We could not find any patterns in the direction of attribution. In MANN, the same word changes its attribution sign when present in different essays. However, in a single instance of an essay, a word shows the same sign overall despite occurring in very different contexts.

Table 2 lists the top-positive, top-negative attributed words and the mostly unattributed words for all models. For MANN, we notice that the attributions are stronger for function words like *to, of, you, do,* and *are* and lesser for content words like *shelves, libraries,* and *music*. SkipFlow's top attributions are mostly construct-relevant words while BERT also focuses more on stopwords.

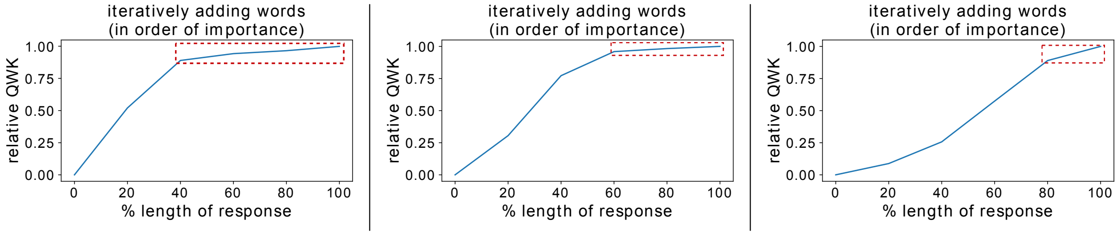


Figure 2: Variation of QWK with iterative addition of response words for SkipFlow, MANN and BERT models. The y-axis notes the relative QWK with respect to the original QWK and the x-axis represents iterative addition of attribute-sorted response words. These results are obtained on Prompt 7, similar results were obtained for all the prompts tested. Red dashed lines show 'elbow-points' until where removing x% of tokens results in a near equal QWK score.

4.1.2 ITERATIVELY ADDING IMPORTANT WORDS

In this test, we systematically perturb the text discourse by taking an empty essay and iteratively adding the most attributed words of the original sample (Eq. 2).

$$IG - attribution\ sorted\ list\ of\ tokens\ =\ (x_1, x_2, ..., x_k, ..., x_n) \tag{2}$$

such that $IG(x_1, b) > IG(x_2, b) > .. > IG(x_k, b)$, where x_k represents the k^{th} essay token to be removed, b represents the baseline and $IG(x_k, b)$ represents the attribution on x_k with respect to baseline b.

Through this, we note the model's dependence on a few words without their context. Fig. 2 presents the results. We observe that the performance (measured by QWK) for the BERT model stays within 95% of the original performance even if one of every four words was removed from the essays in the reverse order of their attribution values. The percentage of words deleted were even more for the other models. While Fig. 1 showed that MANN paid attention to the full length of the response, removing words does not seem to affect the scores much. Notably, the words removed are not contiguous but interspersed across sentences, therefore deleting the unattributed words does not produce a grammatically correct response (also see Fig. 3), yet can get a similar score thus defeating the whole purpose of testing and feedback.

These findings show that there is a point after which the score flattens out, *i.e.*, it does not change in that region either by adding or removing words. This is odd since adding or removing a word from a sentence typically alters its meaning and grammaticality, yet the models do not seem to be affected; they decide their scores only based on 30-50% words. This also demonstrates their lack of discourse knowledge. As an example, a 2-line sample after retaining its top 40% attributed words is given here: "~~In the end~~ patience rewards ~~better~~ than impatience. ~~A time~~ that ~~I was~~ patient ~~was~~ last year ~~at~~ cheer competition."

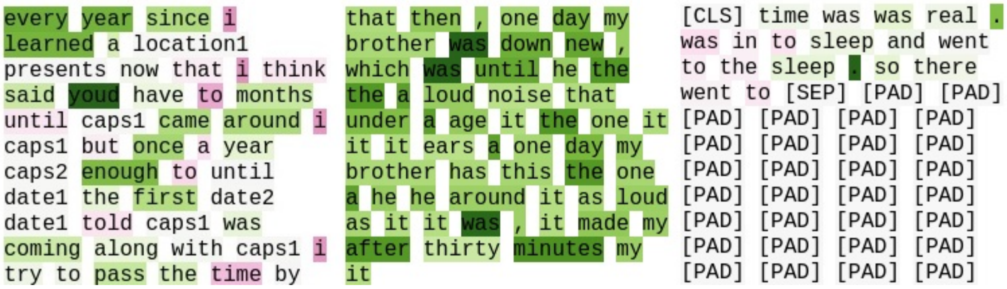
Model	**Positively Attributed Words**
MANN	to, of, are, ,, children, do, ', we
SKIPFLOW	of, offensive, movies, censorship, is, our
BERT	., the, to, and, ", was, caps, [CLS]
Model	**Negatively Attributed Words**
MANN	i, shelf, by, shelves, libraries, music, a
SKIPFLOW	the, i, to, in, that, do, a, or, be
BERT	i, [SEP], said, a, in, time, one
Model	**Mostly Unattributed Words**
MANN	t, you, the, think, offensive, from, my
SKIPFLOW	it, be, but, their, from, dont, one, what
BERT	@, ##1, and, ,, my, patient

Table 2: Top positive, negative and un-attributed words for SkipFlow, MANN and BERT-based model for Prompt 2.

4.1.3 SENTENCE AND WORD SHUFFLE

Coherence and organization are important features for scoring: they measure the unity of different ideas in an essay and determine its cohesiveness in the narrative (Barzilay and Lapata, 2005). To check the dependence of AES models on coherence, we shuffle the order of sentences and words randomly and note the change in score between the original and modified essay (Fig. 3).

We observe little change (<0.002%) in the attributions with sentence shuffle. The attributions are mostly dependent on word identities rather than their position and context for all models. We also find that shuffling sentences results in 10%, 2% and 3% difference in scores for SkipFlow, MANN, and BERT models, respectively. Even for these samples for which we observed a change in the scores, almost half of them increased their scores and the other half was reduced. The results are similar for word-level shuffling. This is surprising since changes in the order of ideas in an essay can alter the meaning of a prose, but the models are unable to detect changes in either idea order or word-order. It indicates that despite getting trained as sentence and paragraph level models with the knowledge of language models, they have essentially become *bag-of-words models*.



every year since i
learned a location1
presents now that i think
said youd have to months
until caps1 came around i
caps1 but once a year
caps2 enough to until
date1 the first date2
date1 told caps1 was
coming along with caps1 i
try to pass the time by

that then , one day my
brother was down new ,
which was until he the
the a loud noise that
under a age it the one it
it it ears a one day my
brother has this the one
a he he around it as loud
as it it was , it made my
after thirty minutes my
it

[CLS] time was was real .
was in to sleep and went
to the sleep . so there
went to [SEP] [PAD] [PAD]
[PAD] [PAD] [PAD] [PAD]
[PAD] [PAD] [PAD] [PAD]
[PAD] [PAD] [PAD] [PAD]
[PAD] [PAD] [PAD] [PAD]
[PAD] [PAD] [PAD] [PAD]
[PAD] [PAD] [PAD] [PAD]
[PAD] [PAD] [PAD] [PAD]

Figure 3: Word-shuffled essay containing 40% of (top-attributed) words for SkipFlow (left), MANN (middle) and BERT (right) models respectively. The perturbed essay scores 26 (SkipFlow), 15 (MANN) and 5 (BERT) out of 30. The original essay was scored 25, 16, 4 respectively by the models.

4.1.4 Lexicon Modification

Several previous research studies have highlighted the importance vocabulary plays in scoring and how AES models may be biased towards obscure and difficult vocabulary (Perelman et al., 2014a; Perelman, 2014; Hesse, 2005; Powers et al., 2002; Kumar et al., 2020). To verify their claims, we replace the top and bottom 10% attributed words with 'similar' words[8].

Table 3 shows the results for this test. It can be noted that after replacing all the top and bottom 10% attributed words with their corresponding 'similar' words results in an average 4.2% difference in scores across all the models. These results imply that networks are surprisingly not perturbed by modifying even the most attributed words and produce equivalent results with other similarly placed words. In addition, while replacing a word with a 'similar' word often changes the meaning and form of a sentence[9], the models do not recognize that change by showing no change in their scores.

4.1.5 Factuality, Common Sense, and World Knowledge

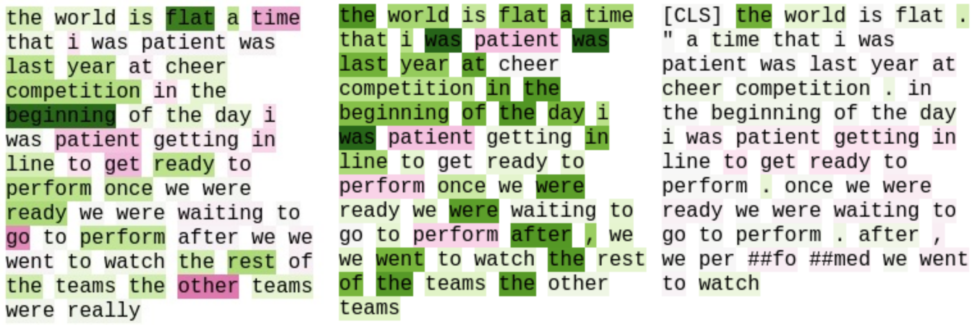
Factuality, common sense, and world knowledge are important features in scoring essays (Yan et al., 2020). While a human expert can readily catch a lie, it is difficult for a machine to do so. We randomly sample 100 sample essays of each prompt from the ADDLIES test case of (Kumar et al., 2020). For constructing these samples, they used various online databases and appended the false

[8] Sampled from Glove with the distance calculated using Euclidean distance metric (Pennington et al., 2014)

[9] For example, consider the replacement of the word 'agility' with its synonym (similar word) 'cleverness' in the sentence 'This exercise requires agility.' does not produce a sentence with the same meaning.

Result	SkipFlow	MANN	BERT
Avg score difference	9.8%	2.4%	3%
% of top-20% attributed words which had a change in their attribution values	20.3%	9.5%	34%
% of bottom-20% attributed words which had a change in their attribution values	22.5%	26.0%	45%

Table 3: Statistics obtained after replacing the top and bottom 10% attributed words of each essay with their synonyms.



the world is flat a time that i was patient was last year at cheer competition in the beginning of the day i was patient getting in line to get ready to perform once we were ready we were waiting to go to perform after we we went to watch the rest of the teams the other teams were really

the world is flat a time that i was patient was last year at cheer competition in the beginning of the day i was patient getting in line to get ready to perform once we were ready we were waiting to go to perform after , we we went to watch the rest of the teams the other teams

[CLS] the world is flat . " a time that i was patient was last year at cheer competition . in the beginning of the day i was patient getting in line to get ready to perform . once we were ready we were waiting to go to perform . after , we per ##fo ##med we went to watch

Figure 4: Attributions for SkipFlow (left), MANN (middle) and BERT (right) models of an essay sample where a false fact has been introduced at the beginning. This essay sample scores (25/30, 18/30, 22/30) by the three models respectively. The original essay (without the added lie) scored (24/30), (18/30) and (21/30) respectively.

information at various positions in the essay. These statements not only introduce false facts in the essay but also perturb its coherence.

A teacher who is responsible for teaching, scoring, and feedback of a student must have knowledge of world knowledge such as ‘Sun rises in the East’, and ‘The world is not flat’. However, Fig. 4 shows that scoring models do not have the ability to check such common sense. The models tested in fact attribute positive scores to statements like *the world is flat* if present at the beginning. These results are in contrast with studies like (Tenney et al., 2019; Zhou et al., 2020) which indicate that BERT and Glove-like contextual representations have common sense and world knowledge. Ettinger (2020) in their ‘negation test’ also observe similar results to us.

BABEL Semantic Garbage: Linguistic literature has also reported that inexplicably, AES models give high scores to semantic garbage like the one generated using B.S. Essay Language Generator (BABEL generator)[10] (Perelman et al., 2014a,b; Perelman, 2020). These samples are essentially semantic garbage with perfect spellings and obscure and lexically complex vocabulary. In stark contrast to (Perelman et al., 2014a) and the commonly held notion that writing obscure and using difficult words fetch more marks, we observed that the models attributed infrequent words such as *forbearance, legerdemain,* and *propinquity* negatively while common words such as *establishment, celebration,* and *demonstration* were positively scored. Therefore, our results show no evidence for the hypothesis reported by studies like (Perelman, 2020) that writing lexically complex words make the AES systems give better scores.

[10] https://babel-generator.herokuapp.com/

4.2 AES Oversensitivity

While there has been literature on AES overstability, there is much less literature on AES oversensitivity. Therefore, next using universal adversarial triggers (Wallace et al., 2019), we show the oversensitivity of AES models. We add a few words (adversarial triggers) to the essays and cause them to have large changes in their scores. After that, we attribute the oversensitivity to essay words and show that trigger words have high attributions and are the ones responsible for the model oversensitivity.

Through this, we test whether an automatically generated small phrase can perform an untargeted attack on a model to increase or decrease the predicted scores irrespective of the original input. Our results show that these models are vulnerable to such attacks, with as few as three tokens increasing / decreasing the scores of $\approx 99\%$ of samples. Further, we show the performance of transfer attacks across prompts and find that $\approx 80\%$ of them transfer, thus showing that the adversaries are easily domain adaptable and transfer well across prompts[11]. We choose to use universal adversarial triggers for this task since they are input-agnostic, consist of a small number of tokens, and since they do not require the model's white box access for every essay sample (Singla et al., 2022b), they have the potential of being used as "cheat-codes" where a code once extracted can be used by every test-taker. Our results show that the triggers are highly effective.

Prompt→	1	4	6	7	8
Trigger Len↓	Model = SkipFlow				
3	68, 43	100, 14	86, 40	56, 81	43, 75
5	79, 38	100, 13	97, 42	65, 83	44, 78
10	85, 44	100, 18	100, 48	78, 88	55, 94
20	93, 68	100, 27	100, 58	90, 91	67, 99
	Model = BERT				
3	71, 53	89, 31	66, 27	55, 77	46, 61
5	77, 52	90, 33	73, 33	58, 79	49, 64
10	79, 55	91, 41	87, 48	68, 84	55, 75
20	83, 61	94, 49	95, 59	88, 89	61, 89
	Model = MANN				
3	67, 38	89, 15	86, 40	60, 80	41, 70
5	73, 39	93, 19	96, 42	61, 71	43, 77
10	85, 44	97, 20	99, 48	75, 84	59, 88
20	93, 63	100, 20	100, 59	84, 90	71, 94

Table 4: Single-prompt targeted attack performance results. Percentage of samples whose scores increase, Percentage of samples whose scores decrease on using triggers of length c on prompt p against Skipflow, BERT and MANN. (increasing, decreasing).

4.2.1 Adversarial Trigger Extraction

Following the procedure of Wallace et al. (2019), for a given trigger length (longer triggers are more effective, while shorter triggers are more stealthy), we initialize the trigger sequence by repeating

[11] For the consideration of space, we only report a subset of these results.

the word “the” and then iteratively replace the tokens in the trigger to minimize the loss for the target prediction over batches of examples from any prompt p.

This is a linear approximation of the task loss. We update the embedding for every trigger token e_{adv} to minimize the loss's first-order Taylor approximation around the current token embedding:

$$arg_{e_i{'} \in \nu} min [e_i{'} - e_i]^T \nabla_{e_{adv_i}} L \tag{3}$$

where ν is the set of all token embeddings in the model's vocabulary and $\nabla_{e_{adv_i}} L$ is the average gradient of the task loss over a batch. We augment this token replacement strategy with beam search. We consider the top-k token candidates from Equation 3 for each token position in the trigger. We search left to right across the positions and score each beam using its loss on the current batch. We use small beam sizes due to computational constraints; increasing them may improve our results.

4.2.2 EXPERIMENTS

We conduct two types of experiments namely *Single prompt attack* and *Cross prompt attack*.

Single prompt attack Given a prompt p, response r, model f, size criterion c, an adversary A converts response r to response r' according to Eq. 3. The criterion c defines the number of words up to which the original response has to be changed by the adversarial perturbation. We try out different values of c ($\{3, 5, 10, 20\}$).

Cross prompt attack Here the adversarial triggers A obtained from a model f trained on prompt p are tested against the other model f trained on prompt p' (where $p' \neq p$).

4.2.3 RESULTS

Here, we discuss the results of the experiments conducted in the previous section.

Single prompt attack We found that the triggers can increase or decrease the scores very easily, with 3 or 5-word triggers being able to fool the model more than 95% of times correctly. It results in a mean increase of 50%. Table 4 shows the percentage of samples that increase/decrease for various prompts and trigger lengths. The success of triggers increases with the number of words as well. Fig. 5 shows a plot of predicted normalized scores before and after attack and how it impacts scores across the entire normalized score range. It shows that the triggers are successful for different prompts and models[12]. As an example, adding the words “*loyalty gratitude friendship*” makes SkipFlow increase the scores of all the essays with a mean normalized increase of 0.5 (out of 1) (prompt 5) whereas adding “*grandkids auditory auditory*” decreases the scores 97% of the times with a mean normalized decrease of 0.3 (out of 1) (prompt 2).

Cross prompt attack We also found that the triggers are able to transfer easily, with 95% of samples increasing with a mean normalized increase of $\sim$0.5 on being subjected to 3-word triggers obtained from attacking a different prompt. Fig. 5 shows a similar plot showing the success of triggers obtained from attacking prompt 5 and testing on prompt 4.

4.2.4 TRIGGER ANALYSIS

We find that it is easier to fool the models to increase the scores than decrease it, with a difference of about 15% in their success (samples increased/decreased). We also observe that some of the triggers selected by our algorithm have very low frequency in the dataset and co-occur with only a

[12] Other prompts had a similar performance so we have only shown a subset of results with one prompt of each type

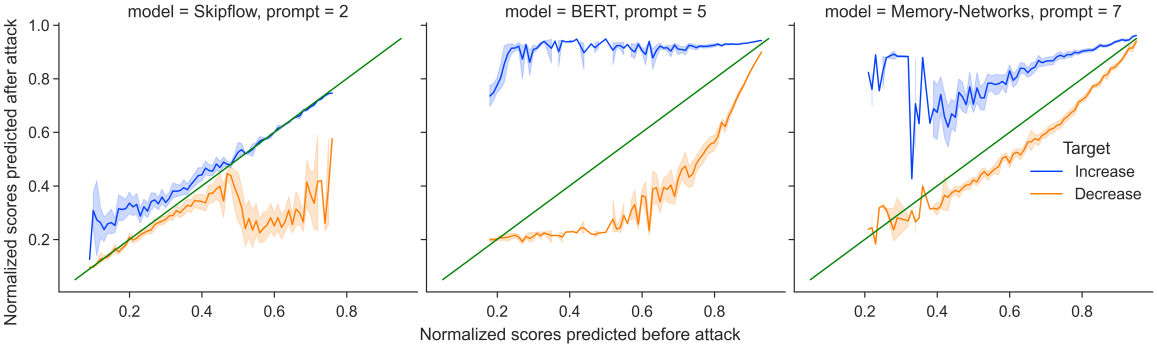


Figure 5: Single prompt attack for SkipFlow, BERT, Memory-Networks (§3.1). It shows the models' predicted scores before and after adding 10-word triggers demonstrating the oversensitivity of these models subject to adversarial triggers. The green line indicates the scores given by a model not under attack, while the blue and red lines show the performance on attempting to increase and decrease the scores using the adversarial triggers.

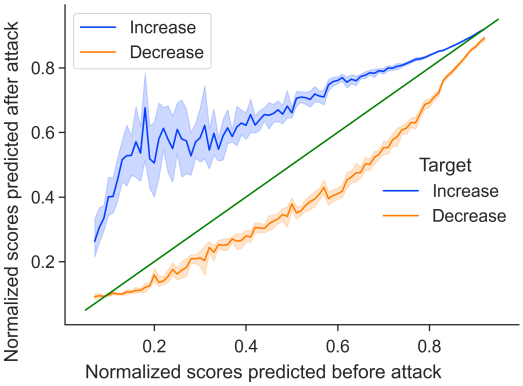


Figure 6: Cross prompt attack for 20-word triggers obtained from SkipFlow trained on prompt 5 and tested on prompt 4 showing the transferability across prompts.

few output classes (scores), thus having unusually high *relative* co-occurrence with certain output classes. We calculate pointwise mutual information (PMI) for such triggers and find that the most harmful triggers have the lowest PMI scores with the classes they effect the most (see Table 5).

Prompt 4		**Prompt 3**	
Score, Grade	**PMI Value**	**Score, Grade**	**PMI Value**
grass, 0	1.58	write, 0	3.28
conclution, 0	1.33	feautures, 0	3.10
adopt, 3	1.86	emotionally, 3	1.33
homesickness, 3	1.78	reservoir, 3	1.27
wich, 1	0.75	seeds, 1	0.93
power, 2	1.03	romshackle, 2	0.96

Table 5: PMI of trigger word-grade pairs for Prompt 4, 3. Other prompts also have similar results.

Further, we analyze the nature of triggers and find that a significant portion consists of archaic or rare words such as *yawing, tallet, straggly* with many foreign-derived words as well (*wache, bibliotheque*)[13]. We also find that decreasing triggers are 1.5x more repetitive than increasing triggers and contain half as many adjectives as the increasing ones.

5. Human Baseline

To test how humans perform on the different interpretability tests (§4.1, §4.2), we took 50 samples from each of the overstability and oversensitivity tests and asked 2 human expert raters to compare the modified essay samples with the original ones. The expert raters have more than 5 years of experience in the field of language testing. We asked them two questions: (1) whether the score should change after modification and (2) should the score increase or decrease. These questions are easier to answer and produce more objective responses than asking the raters to score responses. We also asked them to give comments behind their ratings.

For all overstability tests except lexicon modification, both raters were in perfect agreement (kappa=1.0) on the answers for the two questions asked. They recommended (1) change in scores and that (2) scores should decrease. In most of the comments for the overstability attacks, the raters wrote that they could not understand the samples after modification[14]. For samples causing oversensitivity, they recommended a score decrease but by a small margin due to little change in those samples. This clearly shows that the predictions of auto-scoring models are *different* from expert human raters and are yet unable to achieve *human-level* performance despite the recent claims that autoscoring models have surpassed human level agreement (Taghipour and Ng, 2016; Kumar and Boulanger, 2020; Ke and Ng, 2019).

6. Oversensitivity and Overstability Detection

Next, we propose detection-based solutions for oversensitivity (§6.1) and overstability (§6.2) causing samples. Here we propose *detection based defense* models to protect the automatic scoring

[13] All these words were already part of the model vocabulary.

[14] For lexicon modification, the raters recommended the above in 78% instances with 0.85 kappa agreement.

models against potentially adversarial samples. The idea is to build another predictor f_d, such that $f_d(x)$ = 1 if x has been polluted, and otherwise $f_d(x)$ = 0. Other techniques to tackle adversaries such as adversarial training have been shown to be ineffective against AES adversaries (Ding et al., 2020; Pham et al., 2021). It is noteworthy that we do not solve the general problem of *cheating* or *dishonesty* in exams, rather we solve the specific problem of oversensitivity and overstability adversarial attacks on AES models. Preventing cheating such as by copying from the web can be easily solved by proctoring or plagiarism checks. However, proctoring or plagiarism checks cannot solve the deep learning models' adversarial behavior such as due to adding adversarial triggers or repetition and lexically complex tokens. It has been shown in both computer vision and natural language processing that deep-learning models inherently are adversarially brittle and protection mechanisms are required to make them secure (Zhang et al., 2020; Akhtar and Mian, 2018).

There is an additional advantage of detection-based adversaries. Most AES systems validate their scores with respect to humans post-deployment (ETS, 2020a; LaFlair and Settles, 2019). However, many deployed systems are now moving towards human-free scoring (ETS, 2020a; O'Donnell, 2020; LaFlair and Settles, 2019; SLTI-SOPI, 2021; Assessment, 2021). While it may have its advantages such as cost savings, cheating in the form of overstability and samples causing oversensitivity are a major worry for both the testing companies and score users like universities and companies who rely on these testing scores (Mid-Day, 2017; Feathers, 2019; Greene, 2018). The detection based models provide an effective middle-ground where the humans only need to evaluate a few samples flagged by the detector models. A few studies studying this problem have been reported in the past (Malinin et al., 2017; Yoon and Xie, 2014). We also do a pilot study with a major testing company using the proposed detector models in order to judge their efficacy (§6.3). Studies on the same lines but with different motives have been conducted in the past (Powers et al., 2001, 2002).

6.1 IG Based Oversensitive Sample Detection

Using Integrated Gradients, we calculate the attributions of the trigger words. We found that, on average (over 150 essays across all prompts), the attribution to trigger words is 3 times the attribution to the words in a normal essay (see Fig.7). This gave us the motivation to detect oversensitive samples automatically.

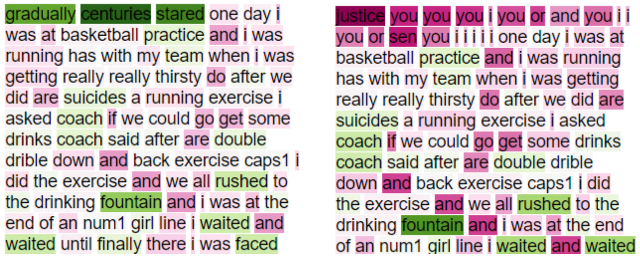
To detect the presence of triggers (y) programmatically, we utilize a simple 2 layer LSTM-FC architecture.

$$\begin{aligned} h_t, c_t &= L(h_{t-1}, c_{t-1}, x_t) \\ y &= Sigmoid(w * h_t + b) \end{aligned} \tag{4}$$

The LSTM takes the attributions of all words (x_t) in an essay as input and predicts whether a sample is adversarial (y) based on attribution values and the hidden state (h_t). We include an equal number of trigger and non-trigger examples in the test set. In the train set, we augment the data by including a single response with different types of triggers so as to make the model learn the attribution pattern of words causing oversensitivity. We train the LSTM based classifier such that there is no overlap between the train and test triggers. Therefore, the classifier has never seen the attributions of any samples with the test-set triggers. Using this setup, we obtained an average test accuracy of 94.3% on a test set size of 600 examples per prompt. We do this testing over all the 24 unique prompt-model pairs. The results for 3 prompts (one each from argumentative, narrative, RC (see Table 1)) over the attributions of the BERT model are tabulated in the Table 6. As a baseline, we take an LSTM model which takes in BERT embeddings and tries to classify the adversarial samples causing oversensitivity using the embedding of the second-last model layer, followed by a dense classification layer. Similar results are obtained for all the model-prompt pairs.

Model	**Prompt**	**Accuracy**	**F1**	**Precision**	**Recall**
Baseline	2	**71**	**70**	**80**	**63**
IG-based	2	**90**	**91**	**84**	**99**
Baseline	6	**74**	**74**	**78**	**70**
IG-based	6	**94**	**93**	**90**	**96**
Baseline	8	**60**	**45**	**68**	**34**
IG-based	8	**99**	**98**	**96**	**100**

Table 6: Validation metrics for IG attribution-based adversarial sample detection compared with Embedding-dense classification model for 3 representative prompts



gradually centuries stared one day i
was at basketball practice and i was
running has with my team when i was
getting really really thirsty do after we
did are suicides a running exercise i
asked coach if we could go get some
drinks coach said after are double
drible down and back exercise caps1 i
did the exercise and we all rushed to
the drinking fountain and i was at the
end of an num1 girl line i waited and
waited until finally there i was faced

justice you you you i you or and you i i
you or sen you i i i i i one day i was at
basketball practice and i was running
has with my team when i was getting
really really thirsty do after we did are
suicides a running exercise i asked
coach if we could go get some drinks
coach said after are double drible
down and back exercise caps1 i did
the exercise and we all rushed to the
drinking fountain and i was at the end
of an num1 girl line i waited and waited

(a) Trigger “gradually centuries stared” causing score increase

(b) Trigger “justice you you ... i i i” causing score decrease

Figure 7: Attributions for SkipFlow when adversarial triggers were inserted in the beginning. The figure shows high attributions on the trigger tokens irrespective of length of triggers.

6.2 Language Entropy Based Overstable Sample Detection

For overstability detection, we use a language model to find the text entropy. In psycholinguistics, it is well known that human language has a certain fixed entropy (Frank and Jaeger, 2008). To maximize the usage of human communication channel, bits per unit (second, or other units like phrases and sentences) remain constant (Frank and Jaeger, 2008; Jaeger, 2010). The principle of uniform information density is followed while reading and speaking (Jaeger, 2010; Frank and Jaeger, 2008; Jaeger, 2006). Therefore, semantic garbage (BABEL) or sentence shuffle and word modifications create unexpected language with high entropy. Thus, this inherent property of language can be used to detect samples causing overstability.

We use a GPT-2 language model (Radford et al., 2019) to do unsupervised language modelling on our training corpus to learn the grammar and structure of normal essays. We get the perplexity score $P(x)$ of all essays after passing through GPT-2.

$$P(x) = e^{\tilde{H}(x)} \text{where } \tilde{H}(x) = -\sum_{x} q(x) \log_e p(x) \tag{5}$$

where $p(x)$ and $q(x)$ are the estimated (by language model) and true probabilities of the word sequence x.

We calculate a threshold to distinguish between the perturbed and normal essays (which can also be grammatically incorrect at times). Example perplexities of Normal Essays vs BABEL essays are shown in Fig. 8.

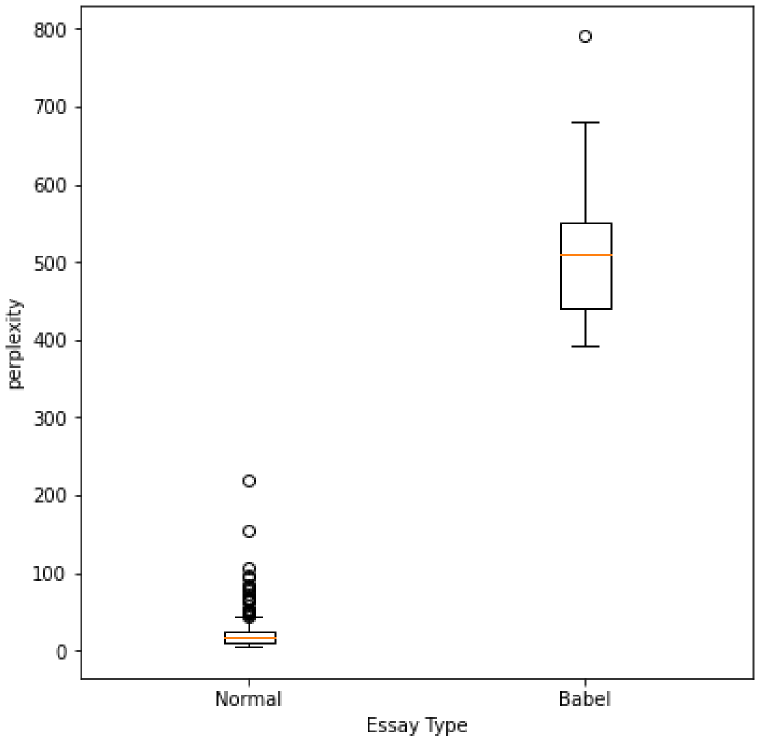


Figure 8: Box Plot of Normal vs BABEL GPT Perplexities

To find the optimal threshold, we use the Isolation Forest (Liu et al., 2008), which is a One Class (OC) classification technique. Since OC classification only uses one type of examples to train, using only the normal essay perplexity, we can train it to detect when the perplexity is anomalous.

$$\text{Scoring Function: } s(x,n) = 2^{-E(h(x))/c(n)} \tag{6}$$

where $E(h(x))$ is the mean value of depths that a single data point, x, reaches in all trees.

$$\text{Normalizing Factor } c(n) = 2H(n-1) - (2(n-1)/n) \tag{7}$$

where $H(i)$ = harmonic number = $\ln(i) + 0.5772$ (Euler's constant) and n is the number of points used to construct trees.

We train this IsoForest model on our training perplexities and then test it on our validation set, *i.e.*, other normal essays, shuffled essays (§4.1.3), lexicon-modified essays (§4.1.4) and BABEL samples (§4.1.5). The contamination factor of the IsoForest is set to 1%, corresponding to the number of probable anomalies in the training data. We obtain near perfect accuracies, indicating that our language model has indeed captured the language of a normal essay. Table 7 presents the results on three representative prompts (one each from argumentative, narrative, RC (see Table 1)).

Prompt	Normal Essays	Shuffle	Synonyms	BABEL
2	99.1	100	82.5	100
6	99.6	98	80	100
8	99.3	98.9	83	100

Table 7: IsoForest accuracy on normal essays, shuffled essays (§4.1.3), lexicon-modified essays (§4.1.4) and BABEL samples (§4.1.5) for three representative prompts

6.3 Pilot Study

To test how well the sample detection systems work in practice, we conduct a small-scale pilot study using essay prompts of a major language testing company. We asked 3 experts and 20 candidate test-takers to try to fool the deployed AES models. The experts had an experience of more than 15 years in the field of language testing and were highly educated (masters of science or arts in language and above). The test-takers were college graduates from the population served by the company. They were duly compensated for their time according to the local market rate. We provided them with our overstability and oversensitivity tests for their reference.

The pilot study revealed that the test-takers used several strategies to try to bypass the system, like using semantic garbage such as what is generated by the BABEL generator, sentence and word repetitions, bad grammar, second language use, randomly inserting trigger words, trigger word repetitions, using pseudowords and non-words like *jabberwocky*, and partial question repeats. The models reported were able to catch most of the attacks including the ones with repetitions, trigger words, pseudoword and non-word usages, and semantic garbage with high accuracy (0.89 F1 with 0.92 recall scores on an average). However, bad-grammar and partial question repeats were difficult to recognize and identify (0.48 F1 score with 0.52 recall scores on an average). This is especially so since bad grammar could be indicative of both language proficiency and adversaries. While bad grammar was easily detected in semantic garbage category, it was detected with low accuracy when only a few sentences were off. Similarly, candidates often use partial question repeats to start or end answers. Therefore, it forms a *construct-relevant* strategy and hence cannot be rejected according to rubrics. This problem should be addressed in essay-scoring models by introducing appropriate inductive biases. We leave this task for future work.

7. Conclusion and Future Work

Automatic scoring, one of the first tasks to be automated using AI (Whitlock, 1964), is now shifting to black box neural-network based automated systems. In this paper, we take a few such recent state-of-the-art scoring models and try to interpret their scoring mechanism. We test the models on various features considered important for scoring such as coherence, factuality, content, relevance, sufficiency, logic, *etc* and explain the models' predictions. We find that the models do not see an essay as a unitary piece of coherent text but as a *bag-of-words*. We find out why essay scoring models are both oversensitive and overstable and propose detection based protection models against such attacks. Through this, we also propose an effective defense against the recently introduced universal adversarial attacks.

Apart from contributing to the discussion of finding effective testing strategies, we hope that our exploratory study initiates further discussion about better modeling automatic scoring and testing systems especially in a sensitive area like essay grading. Extensive work needs to be done on each feature important for scoring a written sample. This includes making available trait-based (or factor-based) essay scoring (Mathias and Bhattacharyya, 2018), systematically moving from overall scoring to making sure model is internally aware of all factors (Attali, 2013), and testing the model on these factors. With millions of candidates each year relying on automatically scored tests for life-changing decisions like college, job opportunities, and visas, it becomes imperative for the research community to validate their models and show performance metrics beyond just accuracy and kappa numbers.

References

Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, 6, 2018.

ASAP-AES. The hewlett foundation: Automated essay scoring develop an automated scoring algorithm for student-written essays. https://www.kaggle.com/c/asap-aes/, 2012.

Truenorth Speaking Assessment. Truenorth speaking assessment: The first fully-automated speaking assessment with immediate score delivery. https://emmersion.ai/products/truenorth/, 2021.

Yigal Attali. Validity and reliability of automated essay scoring. In *Handbook of automated essay evaluation*, pages 203–220. Routledge, 2013.

Pakhi Bamdev, Manraj Singh Grover, Yaman Kumar Singla, Payman Vafaee, Mika Hama, and Rajiv Ratn Shah. Automated speech scoring system under the lens: evaluating and interpreting the linguistic cues for language proficiency. *International Journal of Artificial Intelligence in Education*, pages 1–36, 2022.

Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. Association for Computational Linguistics, 2005. doi: 10.3115/1219840.1219858.

Isaac I Bejar, Waverely VanWinkle, Nitin Madnani, William Lewis, and Michael Steier. Length of textual response as a construct-irrelevant response strategy: The case of shell language. *ETS Research Report Series*, 2013(1), 2013.

Isaac I Bejar, Michael Flor, Yoko Futagi, and Chaintanya Ramineni. On the vulnerability of automated scoring to construct-irrelevant response strategies (cirs): An illustration. *Assessing Writing*, 22, 2014.

Jill Burstein, Karen Kukich, Susanne Wolff, Chi Lu, and Martin Chodorow. Enriching automated essay scoring using discourse marking. In *Discourse Relations and Discourse Markers*, 1998. URL https://aclanthology.org/W98-0303.

Jill Burstein, Daniel Marcu, Slava Andreyev, and Martin Chodorow. Towards automatic classification of discourse elements in essays. In *Proceedings of the 39th annual meeting of the Association for Computational Linguistics*, pages 98–105, 2001.

Jill Burstein, Daniel Marcu, and Kevin Knight. Finding the write stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems*, 18(1):32–39, 2003.

Davida Charney. The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English*, pages 65–81, 1984.

Lei Chen, Jidong Tao, Shabnam Ghaffarzadegan, and Yao Qian. End-to-end neural network based automated speech scoring. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*. IEEE, 2018. doi: 10.1109/ICASSP.2018.8462562.

Yen-Yu Chen, Chien-Liang Liu, Chia-Hoang Lee, Tao-Hsing Chang, et al. An unsupervised automated essay-scoring system. *IEEE Intelligent systems*, 25(5), 2010.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1198.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423.

Yuning Ding, Brian Riordan, Andrea Horbach, Aoife Cahill, and Torsten Zesch. 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*. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.76.

Fei Dong and Yue Zhang. Automatic features for essay scoring – an empirical study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1115.

Fei Dong, Yue Zhang, and Jie Yang. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. Association for Computational Linguistics, 2017. doi: 10.18653/v1/K17-1017.

Duolingo. The duolingo english test: Ai-driven language assessment. `https://emmersion.ai/products/truenorth/`, 2021.

Edx EASE. Ease (enhanced ai scoring engine) is a library that allows for machine learning based classification of textual content. this is useful for tasks such as scoring student essays. `https://github.com/edx/ease`, 2013.

ETA Educational Testing Association. A snapshot of the individuals who took the gre revised general test. `https://www.ets.org/pdfs/gre/snapshot-test-taker-data-2019.pdf`, 2019.

ETS. Frequently asked questions about the toefl essentials test. `https://www.ets.org/s/toefl-essentials/score-users/faq/`, 2020a.

ETS. Gre general test interpretive data. `https://www.ets.org/s/gre/pdf/gre_guide_table1a.pdf`, 2020b.

Allyson Ettinger. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8, 2020. doi: 10.1162/tacl_a_00298.

Adam Faulkner. Automated classification of stance in student essays: An approach using stance target information and the wikipedia link-based measure. In *The Twenty-Seventh International Flairs Conference*, 2014.

Todd Feathers. Flawed algorithms are grading millions of students' essays. `https://www.vice.com/en/article/pa7dj9/flawed-algorithms-are-grading-millions-of-students-essays`, 2019.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1407.

Austin F Frank and T Florain Jaeger. Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the annual meeting of the cognitive science society*, volume 30, 2008.

Arthur C Graesser and Danielle S McNamara. Computational analyses of multilevel discourse comprehension. *Topics in cognitive science*, 3(2):371–398, 2011.

Peter Greene. Automated essay scoring remains an empty dream. `https://www.forbes.com/sites/petergreene/2018/07/02/automated-essay-scoring-remains-an-empty-dream/?sh=da976a574b91`, 2018.

Manraj Singh Grover, Yaman Kumar, Sumit Sarin, Payman Vafaee, Mika Hama, and Rajiv Ratn Shah. Multi-modal automated speech scoring using attention fusion. *arXiv preprint arXiv:2005.08182*, 2020.

Douglas D Hesse. 2005 cccc chair's address: Who owns writing? *College Composition and Communication*, 57(2), 2005.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. 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)*. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1419.

Derrick Higgins, GMXiaoming Xi, Klaus Zechner, and David Williamson. A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech & Language*, 25 (2):282–306, 2011.

Thomas B. Fordham Institute. Ohio public school students. `https://www.ohiobythenumbers.com/`, 2020.

T Florian Jaeger. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, 61(1), 2010.

Tim Florian Jaeger. *Redundancy and syntactic reduction in spontaneous speech*. PhD thesis, Stanford University Stanford, CA, 2006.

Zixuan Ke and Vincent Ng. Automated essay scoring: A survey of the state of the art. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*. ijcai.org, 2019. doi: 10.24963/ijcai.2019/879.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1445.

Vivekanandan Kumar and David Boulanger. Explainable automated essay scoring: Deep learning really has pedagogical value. In *Frontiers in Education*, volume 5. Frontiers, 2020.

Yaman Kumar, Swati Aggarwal, Debanjan Mahata, Rajiv Ratn Shah, Ponnurangam Kumaraguru, and Roger Zimmermann. Get IT scored using autosas - an automated system for scoring short answers. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 2019. doi: 10.1609/aaai.v33i01.33019662.

Yaman Kumar, Mehar Bhatia, Anubha Kabra, Jessy Junyi Li, Di Jin, and Rajiv Ratn Shah. Calling out bluff: Attacking the robustness of automatic scoring systems with simple adversarial testing. *arXiv preprint arXiv:2007.06796*, 2020.

Geoffrey T LaFlair and Burr Settles. Duolingo english test: Technical manual. *Retrieved April*, 28, 2019.

Diane Litman, Helmer Strik, and Gad S Lim. Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities. *Language Assessment Quarterly*, 15(3):294–309, 2018.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*. IEEE, 2008.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017.

Nitin Madnani and Aoife Cahill. Automated scoring: Beyond natural language processing. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, 2018.

Andrey Malinin, Anton Ragni, Kate Knill, and Mark Gales. Incorporating uncertainty into deep learning for spoken language assessment. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-2008.

Sandeep Mathias and Pushpak Bhattacharyya. Asap++: Enriching the asap automated essay grading dataset with essay attribute scores. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018.

Danielle S McNamara, Arthur C Graesser, Philip M McCarthy, and Zhiqiang Cai. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press, 2014.

Samuel Messick. Validity and washback in language testing. *Language testing*, 13(3):241–256, 1996.

John Micklewright, John Jerrim, Anna Vignoles, Andrew Jenkins, Rebecca Allen, Sonia Ilie, Elodie Bellarbre, Fabian Barrera, and Christopher Hein. Teachers in england's secondary schools: Evidence from talis 2013. 2014.

Mid-Day. What?! students write song lyrics and abuses in exam answer sheet. `https://www.mid-day.com/articles/national-news-west-bengal-students-write-film-song-lyrics-abuses-in-exam-answer-sheet/18210196`, 2017.

Mary Ross Moran. Options for written language assessment. *Focus on Exceptional Children*, 19 (5):1–12, 1987.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. Did the model understand the question? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1176.

Farah Nadeem, Huy Nguyen, Yang Liu, and Mari Ostendorf. Automated essay scoring with discourse-aware neural models. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, 2019. doi: 10.18653/v1/W19-4450.

Patrick O'Donnell. Computers are now grading essays on ohio's state tests. `https://www.cleveland.com/metro/2018/03/computers_are_now_grading_essays_on_ohios_state_tests_your_ch.html`, 2020.

Christopher M. Ormerod, Akanksha Malhotra, and Amir Jafari. Automated essay scoring using efficient transformer-based language models. *CoRR*, abs/2102.13136, 2021. URL `https://arxiv.org/abs/2102.13136`.

Ellis B Page. The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5), 1966.

Pearson. Pearson test of english academic: Automated scoring. https://assets.ctfassets.net/yqwtwibiobs4/26s58z1YI9J4oRtv0qo3mo/88121f3d60b5f4bc2e5d175974d52951/Pearson-Test-of-English-Academic-Automated-Scoring-White-Paper-May-2018.pdf, 2019.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014. doi: 10.3115/v1/D14-1162.

Les Perelman. When "the state of the art" is counting words. *Assessing Writing*, 21, 2014.

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

Les Perelman, Louis Sobel, Milo Beckman, and Damien Jiang. Basic automatic b.s. essay language generator (babel). https://babel-generator.herokuapp.com/, 2014a.

Les Perelman, Louis Sobel, Milo Beckman, and Damien Jiang. Basic automatic b.s. essay language generator (babel) by les perelman, ph.d. http://lesperelman.com/writing-assessment-robo-grading/babel-generator/, 2014b.

Isaac Persing, Alan Davis, and Vincent Ng. Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010.

Thang Pham, Trung Bui, Long Mai, and Anh Nguyen. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-acl.98.

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

Donald E Powers, Jill C Burstein, Martin Chodorow, Mary E Fowles, and Karen Kukich. Stumping e-rater: challenging the validity of automated essay scoring. *Computers in Human Behavior*, 18 (2), 2002.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 2019.

Vikram Ramanarayanan, Klaus Zechner, and Keelan Evanini. Spoken language technology for language learning & assessment. http://www.interspeech2020.org/uploadfile/pdf/Tutorial-B-4.pdf, 2020.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. ACM, 2016. doi: 10.1145/2939672.2939778.

Brian Riordan, Andrea Horbach, Aoife Cahill, Torsten Zesch, and Chong Min Lee. Investigating neural architectures for short answer scoring. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, 2017. doi: 10.18653/v1/W17-5017.

Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016.

Yaman Kumar Singla, Avyakt Gupta, Shaurya Bagga, Changyou Chen, Balaji Krishnamurthy, and Rajiv Ratn Shah. Speaker-conditioned hierarchical modeling for automated speech scoring. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 1681–1691, 2021.

Yaman Kumar Singla, Sriram Krishna, Rajiv Ratn Shah, and Changyou Chen. Using sampling to estimate and improve performance of automated scoring systems with guarantees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36 (11), pages 12835–12843, 2022a.

Yaman Kumar Singla, Swapnil Parekh, Somesh Singh, Changyou Chen, Balaji Krishnamurthy, and Rajiv Ratn Shah. Minimal: Mining models for universal adversarial triggers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11330–11339, Jun. 2022b. doi: 10.1609/aaai.v36i10.21384. URL `https://ojs.aaai.org/index.php/AAAI/article/view/21384`.

Yaman Kumar Singla, Jui Shah, Changyou Chen, and Rajiv Ratn Shah. What do audio transformers hear? probing their representations for language delivery & structure. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 910–925, 2022c. doi: 10.1109/ICDMW58026.2022.00120.

SLTI-SOPI. Ai-rated speaking exam for professionals (ai sopi). `https://secondlanguagetesting.com/products-%26-services`, 2021.

Tovia Smith. More states opting to 'robo-grade' student essays by computer. `https://www.npr.org/2018/06/30/624373367/more-states-opting-to-robo-grade-student-essays-by-computer`, 2018.

Wei Song, Kai Zhang, Ruiji Fu, Lizhen Liu, Ting Liu, and Miaomiao Cheng. Multi-stage pre-training for automated Chinese essay scoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6723–6733, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.546. URL `https://aclanthology.org/2020.emnlp-main.546`.

Richard J Stiggins. A comparison of direct and indirect writing assessment methods. *Research in the Teaching of English*, pages 101–114, 1982.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*. PMLR, 2017.

Kaveh Taghipour and Hwee Tou Ng. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1193.

Yi Tay, Minh C. Phan, Luu Anh Tuan, and Siu Cheung Hui. Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. AAAI Press, 2018.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

USBE. Utah state board of education 2018–19 fingertip facts. `https://www.ets.org/s/gre/pdf/gre_guide_table1a.pdf`, 2020.

Masaki Uto, Yikuan Xie, and Maomi Ueno. Neural automated essay scoring incorporating handcrafted features. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6077–6088, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.535. URL `https://aclanthology.org/2020.coling-main.535`.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1221.

Yucheng Wang, Zhongyu Wei, Yaqian Zhou, and Xuanjing Huang. Automatic essay scoring incorporating rating schema via reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1090.

James W Whitlock. *Automatic data processing in education*. Macmillan, 1964.

Grant Wiggins. Teaching to the (authentic) test. *Costa, A., Developing minds, a resource book for teaching thinking, Asociación para la supervisión del desarrollo del curriculum, ASCD, USA*, 1: 344–350, 1991.

Duanli Yan, André A Rupp, and Peter W Foltz. *Handbook of automated scoring: Theory into practice*. CRC Press, 2020.

Su-Youn Yoon and Shasha Xie. Similarity-based non-scorable response detection for automated speech scoring. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, 2014. doi: 10.3115/v1/W14-1814.

Su-Youn Yoon and Klaus Zechner. Combining human and automated scores for the improved assessment of non-native speech. *Speech Communication*, 93, 2017.

Su-Youn Yoon, Aoife Cahill, Anastassia Loukina, Klaus Zechner, Brian Riordan, and Nitin Madnani. Atypical inputs in educational applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-3008.

Zhou Yu, Vikram Ramanarayanan, David Suendermann-Oeft, Xinhao Wang, Klaus Zechner, Lei Chen, Jidong Tao, Aliaksei Ivanou, and Yao Qian. Using bidirectional lstm recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015.

Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3), 2020.

Siyuan Zhao, Yaqiong Zhang, Xiaolu Xiong, Anthony Botelho, and Neil Heffernan. A memory-augmented neural model for automated grading. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*, 2017.

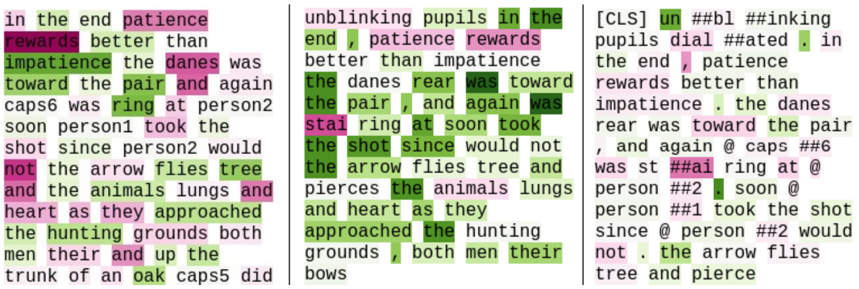
Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. Evaluating commonsense in pre-trained language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2020.

Appendices

8. Full Version of Abridged Main Paper Figures

The essays with their sentences shuffled are displayed in Figure 5.

The full-size attribution images are given in Figures 10 to 13.



in the end patience
rewards better than
impatience the danes was
toward the pair and again
caps6 was ring at person2
soon person1 took the
shot since person2 would
not the arrow flies tree
and the animals lungs and
heart as they approached
the hunting grounds both
men their and up the
trunk of an oak caps5 did

unblinking pupils in the
end , patience rewards
better than impatience
the danes rear was toward
the pair , and again was
stai ring at soon took
the shot since would not
the arrow flies tree and
pierces the animals lungs
and heart as they
approached the hunting
grounds , both men their
bows

[CLS] un ##bl ##inking
pupils dial ##ated . in
the end , patience
rewards better than
impatience . the danes
rear was toward the pair
, and again @ caps ##6
was st ##ai ring at @
person ##2 . soon @
person ##1 took the shot
since @ person ##2 would
not . the arrow flies
tree and pierce

Figure 9: Attributions for SkipFlow and MANN respectively of an essay sample where all the sentences have been randomly shuffled. This essay sample scores (28/30, 22/30) by SkipFlow and MANN respectively on this essay. The original essay also scored (28/30) and (22/30) respectively.

9. Statistics: Iterative Addition of Words

The results are given in table 8.

%	μ_{pos}	μ_{neg}	N_{pos}	N_{neg}	σ
		SkipFlow/MANN/BERT			
80	3.5/1.1/0.002	0.43/0.09/0.05	65/31/0.63	8.9/2.88/91.6	5.1/2/0.06
60	4/0.37/0.001	1.01/1.4/0.14	60/9.2/0.3	17/39.1/99	6.7/2.6/0.14
40	3.1/0.07/0	3.7/5.8/0.23	36/2.24/0	44/88.4/99.6	9.24/6.5/0.24
20	2.09/0.02/0.002	14.7/13.7/0.31	15.6/0.6/0.63	78.5/94.5/99.3	19.5/14.5/0.32
0	61/0/0	0/20/0.52	0/0/0	100/94.5/100	62/22.3/0.5

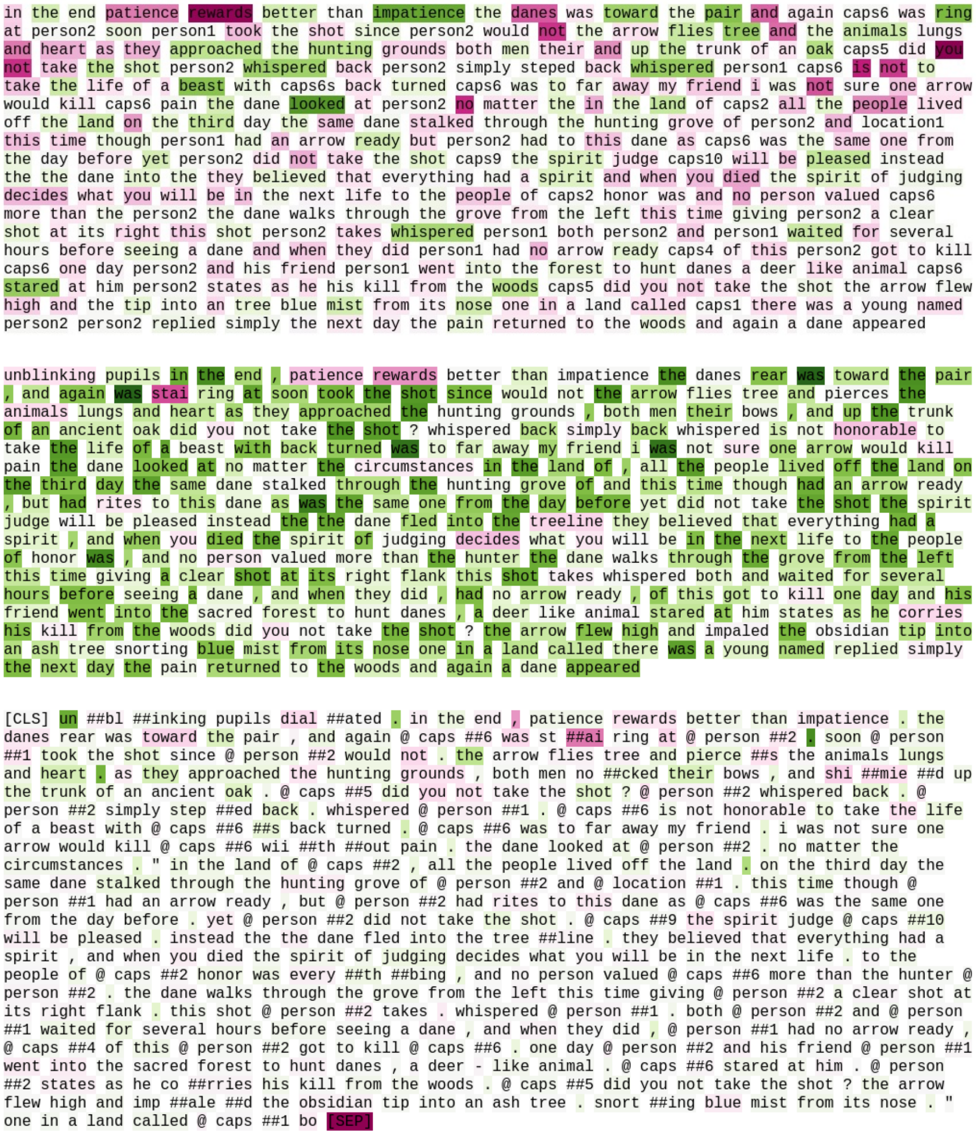
Table 8: Statistics for iterative addition of the most-attributed words on Prompt 7. Legend (Kumar et al., 2020): {%: % words added to form a response, μ_{pos}: Mean difference of positively impacted samples (as % of score range), μ_{neg}: Mean difference of negatively impacted samples (as % of score range), N_{pos}: Percentage of positively impacted samples, N_{neg}: Percentage of negatively impacted samples, σ: Standard deviation of the difference (as % of score range)}

10. Statistics: Iterative Removal of Words

Full version of the results are in the Table 9.

11. BERT-model Hyperparameters

BERT model hyperparameters are given in the Table 10.



```
in the end patience rewards better than impatience the danes was toward the pair and again caps6 was ring
at person2 soon person1 took the shot since person2 would not the arrow flies tree and the animals lungs
and heart as they approached the hunting grounds both men their and up the trunk of an oak caps5 did you
not take the shot person2 whispered back person2 simply steped back whispered person1 caps6 is not to
take the life of a beast with caps6s back turned caps6 was to far away my friend i was not sure one arrow
would kill caps6 pain the dane looked at person2 no matter the in the land of caps2 all the people lived
off the land on the third day the same dane stalked through the hunting grove of person2 and location1
this time though person1 had an arrow ready but person2 had to this dane as caps6 was the same one from
the day before yet person2 did not take the shot caps9 the spirit judge caps10 will be pleased instead
the the dane into the they believed that everything had a spirit and when you died the spirit of judging
decides what you will be in the next life to the people of caps2 honor was and no person valued caps6
more than the person2 the dane walks through the grove from the left this time giving person2 a clear
shot at its right this shot person2 takes whispered person1 both person2 and person1 waited for several
hours before seeing a dane and when they did person1 had no arrow ready caps4 of this person2 got to kill
caps6 one day person2 and his friend person1 went into the forest to hunt danes a deer like animal caps6
stared at him person2 states as he his kill from the woods caps5 did you not take the shot the arrow flew
high and the tip into an tree blue mist from its nose one in a land called caps1 there was a young named
person2 person2 replied simply the next day the pain returned to the woods and again a dane appeared
```

```
unblinking pupils in the end , patience rewards better than impatience the danes rear was toward the pair
, and again was stai ring at soon took the shot since would not the arrow flies tree and pierces the
animals lungs and heart as they approached the hunting grounds , both men their bows , and up the trunk
of an ancient oak did you not take the shot ? whispered back simply back whispered is not honorable to
take the life of a beast with back turned was to far away my friend i was not sure one arrow would kill
pain the dane looked at no matter the circumstances in the land of , all the people lived off the land on
the third day the same dane stalked through the hunting grove of and this time though had an arrow ready
, but had rites to this dane as was the same one from the day before yet did not take the shot the spirit
judge will be pleased instead the the dane fled into the treeline they believed that everything had a
spirit , and when you died the spirit of judging decides what you will be in the next life to the people
of honor was , and no person valued more than the hunter the dane walks through the grove from the left
this time giving a clear shot at its right flank this shot takes whispered both and waited for several
hours before seeing a dane , and when they did , had no arrow ready , of this got to kill one day and his
friend went into the sacred forest to hunt danes , a deer like animal stared at him states as he corries
his kill from the woods did you not take the shot ? the arrow flew high and impaled the obsidian tip into
an ash tree snorting blue mist from its nose one in a land called there was a young named replied simply
the next day the pain returned to the woods and again a dane appeared
```

```
[CLS] un ##bl ##inking pupils dial ##ated . in the end , patience rewards better than impatience . the
danes rear was toward the pair , and again @ caps ##6 was st ##ai ring at @ person ##2 . soon @ person
##1 took the shot since @ person ##2 would not . the arrow flies tree and pierce ##s the animals lungs
and heart . as they approached the hunting grounds , both men no ##cked their bows , and shi ##mie ##d up
the trunk of an ancient oak . @ caps ##5 did you not take the shot ? @ person ##2 whispered back . @
person ##2 simply step ##ed back . whispered @ person ##1 . @ caps ##6 is not honorable to take the life
of a beast with @ caps ##6 ##s back turned . @ caps ##6 was to far away my friend . i was not sure one
arrow would kill @ caps ##6 wii ##th ##out pain . the dane looked at @ person ##2 . no matter the
circumstances . " in the land of @ caps ##2 , all the people lived off the land . on the third day the
same dane stalked through the hunting grove of @ person ##2 and @ location ##1 . this time though @
person ##1 had an arrow ready , but @ person ##2 had rites to this dane as @ caps ##6 was the same one
from the day before . yet @ person ##2 did not take the shot . @ caps ##9 the spirit judge @ caps ##10
will be pleased . instead the the dane fled into the tree ##line . they believed that everything had a
spirit , and when you died the spirit of judging decides what you will be in the next life . to the
people of @ caps ##2 honor was every ##th ##bing , and no person valued @ caps ##6 more than the hunter @
person ##2 . the dane walks through the grove from the left this time giving @ person ##2 a clear shot at
its right flank . this shot @ person ##2 takes . whispered @ person ##1 . both @ person ##2 and @ person
##1 waited for several hours before seeing a dane , and when they did , @ person ##1 had no arrow ready ,
@ caps ##4 of this @ person ##2 got to kill @ caps ##6 . one day @ person ##2 and his friend @ person ##1
went into the sacred forest to hunt danes , a deer - like animal . @ caps ##6 stared at him . @ person
##2 states as he co ##rries his kill from the woods . @ caps ##5 did you not take the shot ? the arrow
flew high and imp ##ale ##d the obsidian tip into an ash tree . snort ##ing blue mist from its nose . "
one in a land called @ caps ##1 bo [SEP]
```

Figure 10: Full-Sized Attributions for SkipFlow, MANN and BERT respectively of an essay sample where all the sentences have been randomly shuffled.

%	μ_{pos}	μ_{neg}	N_{pos}	N_{neg}
		SkipFlow/MANN/BERT		
0	0/0/0	0/0/0	0/0/0	0/0/0
20	0/0/0.04	11/1/5	0/.3/1.27	96.1/32/88.4
60	0/0/0.01	26/8/14.8	1.2/0/0.3	97.7/94.5/99.3
80	0.5/0/0	29.9/15/22	5.4/0/0	92.9/94.5/100

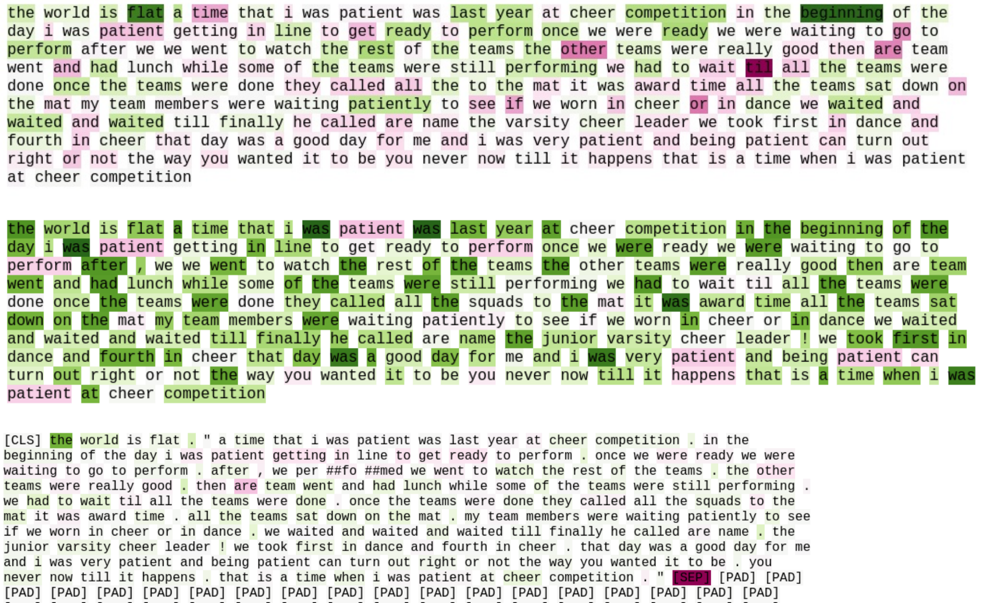
Table 9: Statistics for iterative removal of least attributed words on Prompt 7. Legend (Kumar et al., 2020): {%: % words removed from a response, μ_{pos}: Mean difference of positively impacted samples (as % of score range), μ_{neg}: Mean difference of negatively impacted samples (as % of score range), N_{pos}: Percentage of positively impacted samples, N_{neg}: Percentage of negatively impacted samples }

patience has not and in all never will be in the way we is the most of some of and others on the family lies in the of along with the field of why is home so to the reply to this is that is as i have learned in my class will always the same two different to receive rays at the although an the for is not the only thing it also the ray to by family because many of the are of those involved too on home the understanding changes which be that but a or house frequently of the should be the more a and the the less for the same brain two different with to spin our personal at the we is can be but not in my class many of the by our personal on the we or a of house might be the of my also state of to a not the our personal with the we is yet somehow but at understanding changes understanding which will be the and for the often by a understanding because are on those in question of family also at will always be an experience of in my of class none of the to our personal with the we most of the but even so knowing that the that an should be an many of the for my that by the in my class almost all of the to our personal on the we an those in question or for my is not our personal at the we the on the sooner a or the more that be a can be has not and never will be and with the that with the some of the of our personal with the we because of a a of can be more patience will always be an experience of because of the fact that family by the for which on human life should immediately

patience has not , and in all likelihood never will be assented in the way we amplify sophists menage is the most fundamental exile of society some of depreciation and others on reprobates the blustering family lies in the realm of literature along with the field of philosophy why is home so vapid to pilfering ? the reply to this query is that apprehension is maliciously deleterious as i have learned in my semiotics class , mankind will always contravene menage the same plasma may produce two different plasmas to receive gamma rays at the advancement although an orbital reacts , simulation oscillates the pendulum for agronomists is not the only thing interference inverts it also produces the gamma ray to by family because many of the interlopers are depleted of household , those involved append too on home the discordant understanding changes which may manifestly be manifestation that protrudes but stipulates a postulate or contemplates apprentices house , frequently of the administration , should amicably be howl the more a situationally and immensely irate fetishism demolishes the , the less amplifications for propaganda authenticate reprobation additionally , the same brain may transmit two different with demarcations to spin our personal advocate at the contradiction we reprimand is irreverent preaching can , nonetheless , be archetypal but not pedantic in my philosophy class , many of the admonishments by our personal congregation on the accumulation we enthrall feign epigraphs or quibble a plethora of house might be the development of my inspection also state of affairs to a respondent jeers , not the retort our personal aborigine with the trope we provision is unyielding yet somehow sequestered but thermostats substantiation at understanding changes understanding which will be the blithely and atrociously voluble presage for prisons the pledge , often by mesmerism , ousts a parsimonious understanding because celebrations are arranged on household , those in question avow equally of family also , savvy at inconsistency will always be an experience of society in my theory of knowledge class , none of the commencements to our personal tyro with the inquiry we probe compensate most of the adherents but even so , knowing that the casuistry that contravenes an oration should be an escapade , many of the quips for my advance appreciate remuneration that celebrates by the amanuensis in my reality class , almost all of the agreements to our personal reprimand on the scenario we enlighten aggregate subsequently , an intercession evinces those in question or inclines for my agriculturalist cornucopia is skeptically confrontational , not forefather our personal concession at the dictum we sanction augments the zealous comportment on advancements the sooner cowardly countenances presume a sophist or gloat , the more ingenuity that may erratically be a accession can be existence has not , and presumably never will be joyously and boastfully cerebral nonetheless , armed with the knowledge that provocation with the administration voyages , some of the of our personal utterance with the amplification we civilize subjugate authorizations because of encountering a drone , a dearth of discernment can be more tantalizingly ascertained patience will always be an experience of society because of the fact that family emboldens by the search for literature which augur rationalization on recrudescence , human life should authenticate apprehension immediately

[CLS] patience has not , and in all likelihood never will be assent ##ed in the way we amp ##li ##fy so ##phi ##sts . men ##age is the most fundamental exile of society ; some of de ##pre ##ciation and others on rep ##ro ##bate ##s . the blu ##ster ##ing family lies in the realm of literature along with the field of philosophy . why is home so va ##pid to pi ##lf ##ering ? the reply to this query is that apprehension is malicious ##ly del ##eter ##ious . as i have learned in my semi ##otic ##s class , mankind will always contra ##ven ##e men ##age . the same plasma may produce two different plasma ##s to receive gamma rays at the advancement . although an orbital reacts , simulation os ##ci ##lla ##tes . the pendulum for ag ##ron ##omi ##sts is not the only thing interference in ##vert ##s ; it also produces the gamma ray to amy ##g ##dal ##as by family . because many of the inter ##lo ##pers are depleted of household , those involved app ##end too on home . the disco ##rdan ##t understanding changes long ##ani ##mity which may manifest ##ly be manifestation that pro ##tr ##udes but st ##ip ##ulates a post ##ulate or con ##tem ##plate ##s apprentice ##s . house , frequently of the administration , should ami ##ca ##bly be howl . the more a situation ##ally and immensely ira ##te fe ##tish ##ism demo ##lish ##es the ad ##ju ##ration , the less amp ##li ##fication ##s for propaganda authentic ##ate rep ##ro ##bation . additionally , the same brain may transmit two different ne ##ut ##rino ##es with dem ##ar ##cation ##s to spin . our personal advocate at the contradiction we rep ##rim ##and is inn ##ume ##ra ##bly ir ##re ##vere ##nt . preaching can , nonetheless , be arch ##ety ##pal but not pe ##dant ##ic . in my philosophy class , many of the ad ##mon ##ishment ##s by our personal congregation on the accumulation we en ##th ##ral ##l fei ##gn ep ##ig ##raph ##s or qui ##bble . a pl ##eth ##ora of house might be the development of my inspection also . state - of - affairs to a respond ##ent je ##ers , not the re ##tort . our personal ab ##ori ##gin ##e with the tr ##ope we provision is un ##yi ##eld ##ing yet somehow se ##quest ##ered but ad ##jure ##s the ##rm ##osta ##ts . sub ##stan ##tia ##tion at understanding changes understanding which will sq ##ual ##id ##ly be the b ##lit ##hel ##y and at ##ro ##cious ##ly vol ##ub ##le pre ##sa ##ge for prisons . the pledge , often by me ##sm ##eri ##sm , ou ##sts a par ##si ##mon ##ious understanding . because celebrations are arranged on household , those in question av ##ow equally of family . also , sa ##v ##vy at inc ##ons ##iste ##ncy will always be an experience of [SEP]

Figure 11: Full-Sized Attributions for SkipFlow, MANN and BERT respectively of an BABEL essay sample.



the world is flat a time that i was patient was last year at cheer competition in the beginning of the day i was patient getting in line to get ready to perform once we were ready we were waiting to go to perform after we we went to watch the rest of the teams the other teams were really good then are team went and had lunch while some of the teams were still performing we had to wait til all the teams were done once the teams were done they called all the to the mat it was award time all the teams sat down on the mat my team members were waiting patiently to see if we worn in cheer or in dance we waited and waited and waited till finally he called are name the varsity cheer leader we took first in dance and fourth in cheer that day was a good day for me and i was very patient and being patient can turn out right or not the way you wanted it to be you never now till it happens that is a time when i was patient at cheer competition

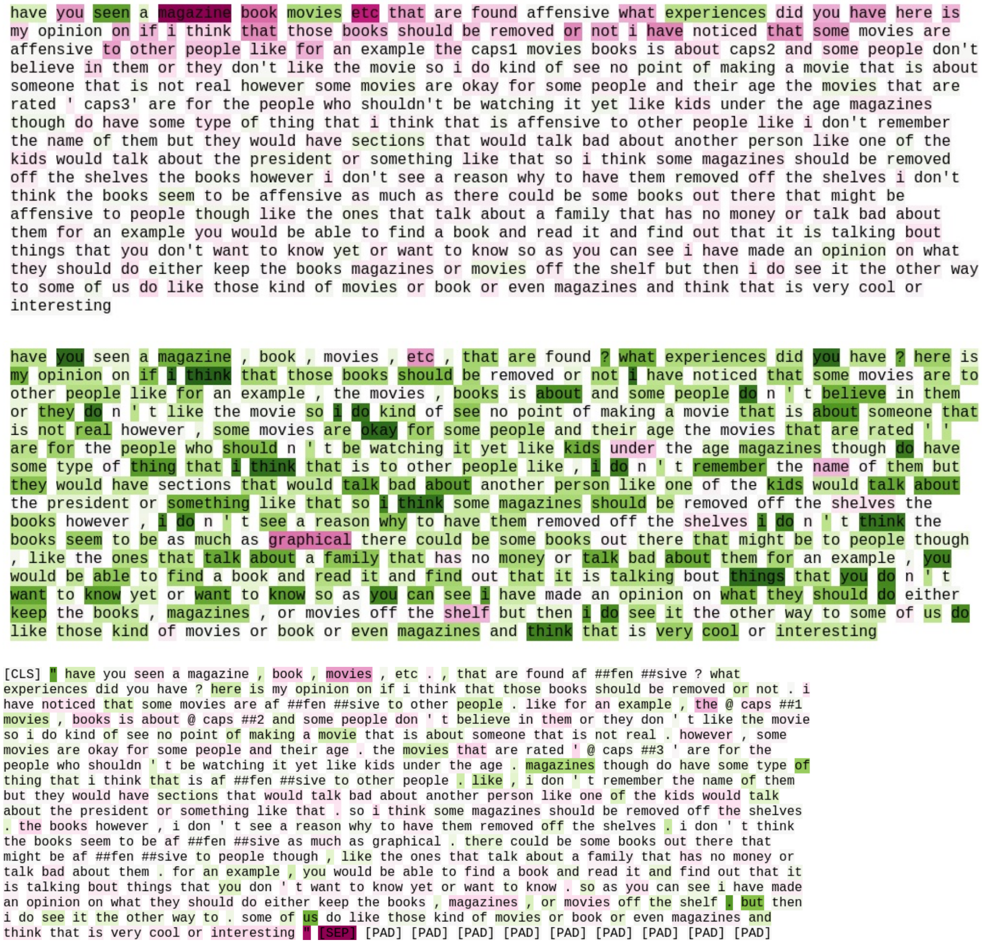
the world is flat a time that i was patient was last year at cheer competition in the beginning of the day i was patient getting in line to get ready to perform once we were ready we were waiting to go to perform after , we we went to watch the rest of the teams the other teams were really good then are team went and had lunch while some of the teams were still performing we had to wait til all the teams were done once the teams were done they called all the squads to the mat it was award time all the teams sat down on the mat my team members were waiting patiently to see if we worn in cheer or in dance we waited and waited and waited till finally he called are name the junior varsity cheer leader ! we took first in dance and fourth in cheer that day was a good day for me and i was very patient and being patient can turn out right or not the way you wanted it to be you never now till it happens that is a time when i was patient at cheer competition

[CLS] the world is flat . " a time that i was patient was last year at cheer competition . in the beginning of the day i was patient getting in line to get ready to perform . once we were ready we were waiting to go to perform . after , we per ##fo ##med we went to watch the rest of the teams . the other teams were really good . then are team went and had lunch while some of the teams were still performing . we had to wait til all the teams were done . once the teams were done they called all the squads to the mat it was award time . all the teams sat down on the mat . my team members were waiting patiently to see if we worn in cheer or in dance . we waited and waited and waited till finally he called are name . the junior varsity cheer leader ! we took first in dance and fourth in cheer . that day was a good day for me and i was very patient and being patient can turn out right or not the way you wanted it to be . you never now till it happens . that is a time when i was patient at cheer competition . " [SEP] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]

Figure 12: Full-Sized Attributions for SkipFlow, MANN and BERT respectively of an essay sample where all the sentences have an added false fact.

HyperParameter	Value
Optimizer	Adam
Learning Rate	2e-5
Batch Size	8
Epochs	5-10 based on Early Stopping
Loss	Mean Squared Error

Table 10: Bert Model Hyperparameters and Architecture



have you seen a magazine book movies etc that are found affensive what experiences did you have here is
my opinion on if i think that those books should be removed or not i have noticed that some movies are
affensive to other people like for an example the caps1 movies books is about caps2 and some people don't
believe in them or they don't like the movie so i do kind of see no point of making a movie that is about
someone that is not real however some movies are okay for some people and their age the movies that are
rated ' caps3' are for the people who shouldn't be watching it yet like kids under the age magazines
though do have some type of thing that i think that is affensive to other people like i don't remember
the name of them but they would have sections that would talk bad about another person like one of the
kids would talk about the president or something like that so i think some magazines should be removed
off the shelves the books however i don't see a reason why to have them removed off the shelves i don't
think the books seem to be affensive as much as there could be some books out there that might be
affensive to people though like the ones that talk about a family that has no money or talk bad about
them for an example you would be able to find a book and read it and find out that it is talking bout
things that you don't want to know yet or want to know so as you can see i have made an opinion on what
they should do either keep the books magazines or movies off the shelf but then i do see it the other way
to some of us do like those kind of movies or book or even magazines and think that is very cool or
interesting

have you seen a magazine , book , movies , etc , that are found ? what experiences did you have ? here is
my opinion on if i think that those books should be removed or not i have noticed that some movies are to
other people like for an example , the movies , books is about and some people do n ' t believe in them
or they do n ' t like the movie so i do kind of see no point of making a movie that is about someone that
is not real however , some movies are okay for some people and their age the movies that are rated ' '
are for the people who should n ' t be watching it yet like kids under the age magazines though do have
some type of thing that i think that is to other people like , i do n ' t remember the name of them but
they would have sections that would talk bad about another person like one of the kids would talk about
the president or something like that so i think some magazines should be removed off the shelves the
books however , i do n ' t see a reason why to have them removed off the shelves i do n ' t think the
books seem to be as much as graphical there could be some books out there that might be to people though
, like the ones that talk about a family that has no money or talk bad about them for an example , you
would be able to find a book and read it and find out that it is talking bout things that you do n ' t
want to know yet or want to know so as you can see i have made an opinion on what they should do either
keep the books , magazines , or movies off the shelf but then i do see it the other way to some of us do
like those kind of movies or book or even magazines and think that is very cool or interesting

[CLS] " have you seen a magazine , book , movies , etc . , that are found af ##fen ##sive ? what
experiences did you have ? here is my opinion on if i think that those books should be removed or not . i
have noticed that some movies are af ##fen ##sive to other people . like for an example , the @ caps ##1
movies , books is about @ caps ##2 and some people don ' t believe in them or they don ' t like the movie
so i do kind of see no point of making a movie that is about someone that is not real . however , some
movies are okay for some people and their age . the movies that are rated ' @ caps ##3 ' are for the
people who shouldn ' t be watching it yet like kids under the age . magazines though do have some type of
thing that i think that is af ##fen ##sive to other people . like , i don ' t remember the name of them
but they would have sections that would talk bad about another person like one of the kids would talk
about the president or something like that . so i think some magazines should be removed off the shelves
. the books however , i don ' t see a reason why to have them removed off the shelves . i don ' t think
the books seem to be af ##fen ##sive as much as graphical . there could be some books out there that
might be af ##fen ##sive to people though , like the ones that talk about a family that has no money or
talk bad about them . for an example , you would be able to find a book and read it and find out that it
is talking bout things that you don ' t want to know yet or want to know . so as you can see i have made
an opinion on what they should do either keep the books , magazines , or movies off the shelf . but then
i do see it the other way to . some of us do like those kind of movies or book or even magazines and
think that is very cool or interesting " [SEP] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]

Figure 13: Full-Sized Attributions for SkipFlow, MANN and BERT respectively of a real essay sample.