Noname manuscript No.
(will be inserted by the editor)

Calling Out Bluff: Attacking the Robustness of Automatic Scoring Systems with Simple Adversarial Testing

Yaman Kumar⋆ · Mehar Bhatia⋆ · Anubha Kabra⋆ · Junyi Jessy Li · Di Jin · Rajiv Ratn Shah

Received: date / Accepted: date

Abstract Significant progress has been made in deep-learning based Automatic Essay Scoring (AES) systems in the past two decades. Techniques such as memory networks and two stage learning have led to improvement in the performance metrics like Quadratic Weighted Kappa (QWK), and accuracy. There have been a few studies noting that the scores given by AES models correlate heavily with features like length and wordiness of the essay. However, till date, there has been no study testing AES systems on different types of common-sense adversarial examples such as addition of lines in essays or grammatical modification of essays. Inspired by student behaviour during examinations, we propose a model agnostic adversarial evaluation scheme and

⋆ Equal Contribution

Y. Kumar
IIIT-Delhi, India
Adobe, India
E-mail: yamank@iiitd.ac.in
E-mail: ykumar@adobe.com

M. Bhatia
IIIT-Delhi, India
E-mail: mehar.bhatia@midas.center

A. Kabra
IIIT-Delhi, India
E-mail: anubha.kabra@midas.center

J.J. Li
University of Texas at Austin, United States of America
E-mail: jessy@austin.utexas.edu

D. Jin
Massachusetts Institute of Technology, United States of America
E-mail: dijin15@mit.edu

R.R. Shah
IIIT-Delhi, India
E-mail: rajivratn@iiitd.ac.in

associated metrics for AES systems to test their natural language understanding capabilities and overall robustness. We also evaluate the current state-of-the-art AES methodologies using the scheme and find that AES models are highly overstable such that even heavy modifications (as much as 25%) with unrelated content do not decrease the score produced by the models. On the other hand, unrelated content, on an average, increases the scores, thus showing that models' evaluations are removed from the rubrics.

Keywords Automatic Scoring · Adversarial Testing · Adversarial Robustness · Essay Scoring

1 Introduction

Automated Essay Scoring (AES) uses computer programs to automatically characterize the performance of examinees on standardized tests involving writing prose. The earliest mention of scoring as a scientific study dates back to the nineteenth century (Spolsky, 1995) and automatic scoring, specifically, to the 1960s (Whitlock, 1964). The field started with Ajay et al. (1973) scoring the essays of their students on punch cards. The essay was converted to a number of features and passed through a linear regression model to produce a score. Since then, the field has undergone major changes which transformed the punch cards to microphones and keyboards; and linear regression techniques on manually extracted features to deep neural networks (Ajay et al., 1973; EASE, 2013; Tay et al., 2018).

The developments in AES systems have been majorly driven by the key advantages it offers. The most important among them is the ease of the process of scoring while still being cost-effective. However, over the years there have been many validity tests and criticisms of AES models on their evaluation methodology. Perelman (2014) argued that the state-of-the-art systems showed a substantial correlation with just the number of words. In Reinertsen et al. (2018), the authors observe that the Australian *eWrite* AES system rejected those writings that did not match the style of their training samples and that it is not suitable for a broad-based system like essay scoring systems. In West-Smith et al. (2018), the authors note that there is no systematic framework for evaluating a model's fit for educational purposes. This leads to a lack of trust in high-stakes scenarios such as that of AES evaluation. Perelman designed Basic Automatic B.S. Essay Language Generator (BABEL) (Perelman et al., 2014a) to test out and show that the state-of-the-art AI systems can be fooled by crudely written prose (Perelman et al., 2014b) [1]. However, to our best knowledge, there has been no work systematically analyzing AES models on all the different aspects important for scoring. Although there have been some manual studies where experts and non-experts were invited to test out some specific models (Powers et al., 2001), these studies cannot be scaled or made

[1] We also use BABEL for one of the tests in our framework (namely BabelGen given in Section 3.2.5).

consistent across all the models. Such a validity suite is important from the following perspectives: 1) it provides a uniform benchmark to compare different models beyond metrics such as accuracy or QWK which do not provide any insights into the construct validity of AES models, 2) it builds trust in the automatic scoring system, and 3) it promotes understanding of the black-box AES models.

Based on cognitive studies, AES models have been seen as information-integration models trying to learn category-learning tasks (Yan et al., 2020). The descriptor of such a category can be, *"Score the essay at level 3 if it consists of a clear aim reasoned by structured claims and supported by appropriate evidence with rebuttals of all the major counter arguments."* (Yan et al., 2020). Following this, many research studies have established features which must be present in AES models (Yan et al., 2020; Burstein et al., 2004; Sukkarieh and Blackmore, 2009; Kumar et al., 2019). A few examples of such features are: factuality, grammar-correctness, organization, coherence, lexical sophistication, *etc.* In this work, we propose a black-box adversarial evaluation of AES systems based on these features. We show the evaluation of five recent models on the popular dataset, Automated Student Assessment Prize (ASAP) dataset for Essay-Scoring (ASAP-AES, 2012). Our evaluation scheme consists of evaluating AES systems on essays derived from the original responses but modified heavily to change its original meaning. These tests are mostly designed to check for the overstability of the different models. An overview of the adversarial scheme is given in Table 1. We try out the following operations for generating responses: *Addition*, *Deletion*, *Modification* and *Generation.* Under these four operations, we include many other operation subtypes such as adding related and unrelated content, modifying the grammar of the response, taking only first part of the response, *etc.* As the human evaluation results show (Section 4.3, Table 20), when shown these adversarial responses to humans, they perceive the responses as ill-formed, lacking coherence and logic.

The common performance metric that has been widely used in the field is Quadratic Weighted Kappa (QWK). According to this performance metric, with time, the automatic essay scoring models have reached the level of humans (Kumar et al., 2019) or even 'surpassed' them (Shermis and Hamner, 2012). However, as our experiments show, despite achieving parity with humans on QWK scores, models are not able to score in the same manner as humans do. We demonstrate in the later parts of the paper that heavily modifying responses (as much as 25%), does not break the scoring systems and the models still maintain their high confidence and scores while evaluating the adversarial responses.

However, our results demonstrate that no published model is robust to these examples. They largely maintain the scores of the unmodified original response even after all the adversarial modifications. This indicates that the models are largely overstable and unable to distinguish ill-formed examples from the well-formed ones. While on an average, humans reduce their score by approx 3-4 points (on a normalized 1-10 scale), the models are highly overstable and either increase the score by 0-1 points for some tests or reduce them for others by only 0-2 points.

We also argue that for deep learning based systems, tracking merely QWK as an evaluation metric is suboptimal for several reasons: 1) while subsequent research papers show an iterative improvement in QWK but most of them fail in evaluating how their works generalize across all the different dimensions of scoring including coherence, cohesion, vocabulary, and even surface metrics like average length of sentences, word difficulty, *etc.* 2) QWK as a metric captures only the overall and broad agreement with humans scores, however, scoring as a science includes knowledge from many domains of NLP like: *fact-checking, discourse and coherence, coreference resolution, grammar, content coverage, etc* (Yan et al., 2020). QWK, instead of making the scoring comprehensive, is abstracting out all the details associated with scoring as a task. 3) it does not indicate the direction of a machine learning model: oversensitivity or overstability. We quantitatively illustrate the gravity of all these aspects by performing statistical and manual evaluations, mentioned in Section 3. We propose that instead of tracking *just* QWK for evaluating a model, the field should track a combination of QWK and adversarial evaluation of the models for performance.

Our main contributions in this work are summarized as follows:

- We propose a model agnostic evaluation suite to alter examples given in a dataset to test out a given AES model.
- We evaluate five recent state-of-the-art AES models on all the eight prompts belonging to the widely-cited ASAP-AES (2012) dataset and report their test performance on various metrics for a thorough understanding of their weaknesses.
- We propose a comprehensive 3-way automatic evaluation for aiding model-makers involving parameters of length, position and type of adversarial tests. We also validate the adversarial examples with a human study to show that scores awarded by AES models are indeed disconnected with rubrics.
- Finally, we open-source the code, test samples and model weights for easy reproducibility, and future benchmarking.

Finally, we would like to say that we present our argument not as a criticism of anyone, but as an effort to refocus research directions of the field. Since the automated systems that we develop as a community have such high stakes like deciding jobs and admissions of the takers, the research should reflect the same rigor. We sincerely hope to inspire higher quality reportage of the results in automated scoring community that does not track just performance but also the validity of their models.

#	Category	Test Name	Description
1	Add	AddWikiRelated	Addition of Wikipedia lines related to the essay question in a response.
		AddWikiUnrelated	Addition of Wikipedia lines unrelated to the essay question in a response.
		RepeatSent	Repetition of some lines of the response within a response.
		AddSong	Addition of song lyrics into the response.
		AddSpeech	Addition of excerpts of speeches of popular leaders into a response.
		AddRC	Addition of lines from Reading Comprehension based questions into a response.
		AddTruth	Addition of True lines into a response.
		AddLies	Addition of Universally false lines into a response.
2	Delete	DelStart	Deletion of lines from the beginning of a response.
		DelEnd	Deletion of lines from the end of a response.
		DelRand	Deletion of random lines from a response.
3	Modify	ModGrammar	Modifying the sentences in a response to have incorrect grammar.
		ModLexicon	Paraphrasing words in the sentences with their respective synonyms in a response.
		ShuffleSent	Randomly shuffling the sentences in a response.
4	Generate	BabelGen	Using the essay generated by *Babel* as a response.

Table 1 Overview of the testing scheme for Automatic Scoring (AS) models

2 Task and Setup

2.1 Task and Dataset

We used the widely cited ASAP-AES (ASAP-AES, 2012) dataset to evaluate Automatic Essay Scoring systems. Research studies have frequently used ASAP-AES for automatically scoring essay responses (Taghipour and Ng, 2016; EASE, 2013; Tay et al., 2018; Zhao et al., 2017). The relevant statistics for this dataset are listed in Table 2. The questions covered by the dataset are from many different areas such as Sciences and English literature. The responses were written by high school students and were subsequently double-scored. The evaluation framework built for assessing AES systems is broadly based on the linguistic features considered essential for scoring like grammar, coherence, *etc* (Bejar et al., 2017; Yan et al., 2020). We discuss the framework in greater details in the Section 3.2.

2.2 Models

We evaluate the recent state-of-the-art deep learning and feature-based models. We show the adversarial-evaluation results for five such models: EASE

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	420	430	127	109	147	180	205	710
#Avg sentences per response	23	20	6	4.5	7	8	12	35
Type	Ar	Ar	RC	RC	RC	RC	Na	Na

Table 2 Overview of the ASAP AES Dataset used for evaluation of AS systems. (RC = Reading Comprehension, Ar = Argumentative, Na = Narrative)

(2013); Taghipour and Ng (2016); Tay et al. (2018); Zhao et al. (2017); Liu et al. (2019). Brief descriptions of each of them are given as follows:

EASE: *EASE* (EASE, 2013) is an open-source feature-based model maintained by *EdX*. This model includes features such as tags, prompt-word overlap, n-gram based features, *etc.* Originally, it ranked third among the 154 participating teams in the ASAP-AES competition.

Taghipour and Ng (2016): They use CNN-LSTM based neural networks with a few mean-over-time layers to score essays. The paper reports 5.6% improvement of QWK on top of the *EASE* feature-based model.

SkipFlow: *SkipFlow* (Tay et al., 2018) provides a deep learning architecture that captures coherence, flow and semantic relatedness over the length of the essay, which the authors call *neural coherence features*. SkipFlow accesses intermediate states to model longer sequences of essays. Doing this, they show an increase of 6% over EASE feature engineering model and 10% over a vanilla LSTM model.

Zhao et al. (2017): The authors use memory-networks for automatic scoring where they select some responses for each grade. These responses are stored in the 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 compare their results with the EASE based model and show better performance on 7 out of 8 prompts.

Liu et al. (2019): This work makes use of adversarial examples to improve AES. They consider two types of adversarial evaluation: well-written permuted paragraphs and prompt-irrelevant essays. For these, they develop a two-stage learning framework where they calculate semantic, coherence and prompt-relevance scores and concatenate them with engineered features. The paper uses BERT (Devlin et al., 2018) to extract sentence embeddings.

3 Evaluation

In this section, we discuss the standard evaluation metric, Quadratic Weighted Kappa (QWK) and the adversarial evaluation framework aim to propose.

3.1 Standard Evaluation

Quadratic Weighted Kappa is the standard metric for evaluating model performance on the essay scoring task (Attali and Burstein, 2004). Both the competition which released the ASAP-AES dataset and the subsequent papers using that employ it as the evaluation metric. Given observed scores matrix O (confusion scores), weights w (containing penalty of each possible predicted score with each possible actual score) and expected score matrix E, number of possible scores N, QWK is calculated as

$$k = 1 - \Sigma_{ij} w_{ij} O_{ij} / \Sigma_{ij} w_{ij} E_{ij} \tag{1}$$

O_{ij} measures number of students who received a score i by the human grader and j by the model. Weight matrix is defined as ($w_{ij} = (i-j)^2/(N-1)^2$) and assigns penalty to each pair of predicted, actual scores. QWK obtained this way denotes machine-human agreement. It is then compared with human-human agreement score to compare different models.

3.2 Adversarial Evaluation

3.2.1 General Framework

Figure 1 depicts the general framework for the adversarial evaluation. Given a prompt p, response r, bounded size criterion c_1, position criterion c_2 and optionally a model f, an adversary A converts response r to response r' based on a specific set of rules and the criteria c_1 and c_2.

The criterion c_1 defines the percentage upto which the original response has to be changed by the adversarial perturbation such that $|Len(r') - Len(r)|/Len(r) = c_1$. We try out different values of c_1 ({10%, 15%, 20%, 25%}). The criterion c_2 defines the position of inducing adversarial perturbation. We consider three positions ({Start, Mid, End}) by dividing the response r into three equal-sized portions. The results of different values of c_1 and c_2 are presented in the Section 4.1.

For benchmarking a model f, we use the scores $f(r)$ and $f(r')$ to calculate the statistics listed in Table 3. Since the score ranges and the number of samples vary across all the prompts, we report the corresponding values in percentages (percentage of total samples and percentage of range of score). For knowing un-normalized values, readers are encouraged to look into the supplementary.

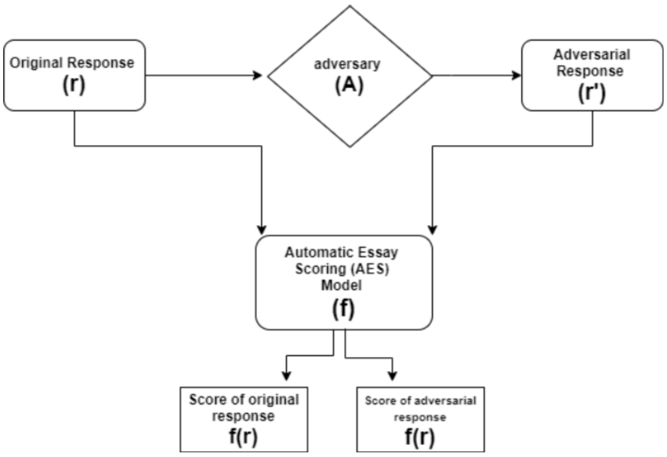

Fig. 1 General Framework for Adversarial Evaluation given a prompt p

Symbol	Name	Defn.
N_{neg}	Percentage of negatively impacted samples	$\#\ r/N$ s.t., $f(r) > f(r'))$
N_{pos}	Percentage of positively impacted samples	$\#\ r/N$ s.t., $f(r) < f(r'))$
μ	Mean difference	$\Sigma(f(r) - f(r'))/N)$
μ_{abs}	Absolute mean difference	$\|\Sigma(f(r) - f(r'))\|/N$
σ	Standard deviation of the difference	$\sqrt{\Sigma(f(r) - f(r') - \mu)^2/N}$
μ_{pos}	Mean difference of negative impacted samples	$\Sigma(f(r) - f(r'))/N$ s.t. $f(r) < f(r')$
μ_{neg}	Mean difference of positively impacted samples	$\Sigma(f(r') - f(r))/N$ s.t. $f(r) > f(r')$

Table 3 Adversarial Evaluation Metrics

Using human evaluation (Section 4.3) and relevant statistics (Section **??**), we make sure that an adversary A satisfies the following two conditions.

1. According to a human, the score of an adversarial response (r') should always be lesser than the score of the original response (r). In other words, no adversary should increase the quality of the response.
2. Second, a human should be able to detect and differentiate r from r'.

Notably, these requirements are different from what is "commonly" given in the adversarial literature where the adversarial response is formed such that a human is not able to detect any difference between the original and modified responses, but a model (due to its adversarial weakness) is able to detect differences and thus changes its output (Zhang et al., 2020). For example, in computer vision, a few pixels are modified to make the model mispredict a bus as an ostrich (Szegedy et al., 2013), and in NLP, paraphrasing by changing a few words is done to churn out racial and hateful slurs from a generative deep learning model (Wallace et al., 2019). Here, we make sure that humans *detect* the difference between the original and final response. We call the inability (or under-performance) of models on differentiating between adversarial and natural samples as their *overstability*.

Next, we discuss the various strategies of adversarial perturbations. An overview of all the perturbations is given in Table 1. We categorize all the adversarial tests by the *major*-operation they do on a sample. Therefore, we divide the tests into four categories: ADD (those operations which change a sample majorly by adding to it), DELETE (those operations which change a sample majorly by deleting from it), MODIFY (those operations which change a sample majorly by modifying its structure) and GENERATE (those operations which tests the robustness of a model by giving it completely machine-generated non-meaningful samples).

3.2.2 MODIFY *Adversaries*

MODIFY adversaries majorly retain the originality of a response while changing its syntax heavily. In this, we majorly change the grammar, fluency, organization and lexical sophistication of a sample. The various types of MODIFY tests are explained hereafter.

- **ModGrammar**: Several studies underline the importance of grammar in scoring (Attali and Burstein, 2004; Burstein et al., 2004). *TOEFL iBT* mentions grammar usage in the category 'language use' for their TOEFL test (Cushing Weigle, 2010). We formed two test cases to simulate common grammatical errors committed by students. The first one focused on evaluating the basic grammar knowledge of AES model and the second one assessed the effect of colloquial and informal language commonly found in essays. For the first one, we took c_1 sentences and changed the subject-verb-object (SVO) order of those sentences. We do this by parsing the responses and using spacy[2] python library to extract grammatical dependencies. In the second test case, we first induced article errors by replacing the articles of a sentence with their commonly used incorrect forms, such as using 'a' instead of 'an' before the noun 'apple'. This was followed by modifying subject-verb agreement [3], with the help of inflect library[4]. Using an abbreviation dictionary [5], we replaced randomly selected words with their corresponding informal colloquial forms. An example of each step of this test is demonstrated in the Table 4.
- **ModLexicon**: Diversity and sophistication of vocabulary is an essential feature for scoring essays (Chen et al., 2018; Kumar et al., 2019). It is commonly observed that test-takers using sophisticated vocabulary often are scored higher than their counterparts using simpler, more straightforward vocabulary (Perelman et al., 2014a). However, the change or inclusion of even a single word in a sentence changes its meaning. Therefore, in this test case, we evaluate AES systems' vocabulary-dependence by improper

[2] https://spacy.io/
[3] https://www.grammarbook.com/grammar/subjectVerbAgree.asp
[4] https://pypi.org/project/inflect/
[5] https://abbreviations.yourdictionary.com/articles/list-of-commonly-used-abbreviations.html

Original	**Anita is going to the park for a walk.**
Subject-Verb-Object Order Errors	Anita to the park is going for a walk.
Step 1: Article Errors	Anita is going to an park for the walk.
Step 2: Subject Verb Agreement Errors	Anita go to an park for the walk.
Step 3: Conventional Errors	anita go 2 an park 4 the walk

Table 4 Examples of the type ModGrammar

replacement of a certain word to one of the words in its synsets. We do this by replacing one word (excluding the stop words) randomly in each sentence with its synonym using Wordnet synsets (Miller, 1995) for this purpose. Later, in Section 4.3, we observe that a human would view such an example as a change in vocabulary but with improper usage of the words changed. An example of this type of perturbation is, "*Tom was a happy man. He lived a simple life.*". It gets changed to "*Tom was a grinning man. He lived a bare life.*"
– **ShuffleSent**: Coherence and organization measure the extent to which a response demonstrates a unified structure, direction of the narrative, and unity in the different parts of an essay (Schultz, 2013; Barzilay and Lapata, 2008). They are important aspects of scoring an essay (Foltz et al., 2013; Tay et al., 2018; Chen et al., 2018). To evaluate the dependence of AES scoring on coherence, we randomly shuffle the sentences of a response and score the modified response. Sentence shuffling ensures the response's readability and coherence are affected negatively (Xu et al., 2019). It affects the transition between the lines so that the different ideas appear disconnected to a reader. An AES system should negatively score this type of perturbation since it changes the meaning of the text substantially. Table 5 shows some examples of such perturbation.

3.2.3 Add *Adversaries*

Add adversaries change the original response by adding new content to it. Adding unrelated or repetitive content negatively impacts the content-specific and topic development features of an essay, which are considered necessary for essay evaluation (Yan et al., 2020). To test the content knowledge of scoring models, we designed various types of Add tests that are explained hereafter.

– **AddWikiRelated**: With this testcase, we add prompt-related information to each sample response. We used a key-phrase extraction technique [6] over each prompt/question in the dataset for choosing prompt-related articles from Wikipedia. Through this method, we formed a list of important topics for each prompt and selected corresponding articles from Wikipedia for each topic [7]. After selecting articles, we randomly selected sentences from each extracted article and appended them to the responses. The position

[6] https://github.com/boudinfl/pke
[7] https://pypi.org/project/wikipedia/

Original	Perturbation using ShuffleSent
Example 1: Dear newspaper, I think that people should have computers. I think that because kids play games to help them learn like math game or read and grown ups can go on websites and watch videos and do their office work. They could also pay their bills on the computer and do shopping if they do not want to go to the store. There are many advantages of having a computer.	I think that because kids play games to help them learn like math game or read and grown ups can go on websites and watch videos and do their office work. Dear newspaper, I think that people should have computers. There are many advantages of having a computer. They could also pay their bills on the computer and do shopping if they do not want to go to the store.
Example 2: The setting affected the cyclist a lot. One way was the heat. It got so hot that he was becoming dehydrated. He kept drinking his water, until he didn't have any left. Another reason is because the road turned into hills which made him use more energy. Finally, it didn't help that he hadn't seen any sign of living creatures except for a snake.	Another reason is because the road turned into hills which made him use more energy. The setting affected the cyclist a lot. Finally, it didn't help that he hadn't seen any sign of living creatures except for a snake. One way was the heat. It got so hot that he was becoming dehydrated. He kept drinking his water, until he didn't have any left.

Table 5 Examples of perturbation to response using ShuffleSent. Color coding has been done to show the modified position of sentences pre and post-perturbation.

and amount of addition are given by the parameters, c_1 and c_2, respectively, as explained in Section 3.2.1. Examples of such types of perturbation are given in Table 6.

Original	Perturbation using AddWikiRelated
Example 1: Dear newspaper, I think that people should have computers. I think that because kids play games to help them learn like math game or read and grown ups can go on websites and watch videos and do their office work. They could also pay their bills on the computer and do shopping if they do not want to go to the store. There are many advantages of having a computer.	Computers are used as control systems for a wide variety of industrial and consumer devices. The first digital electronic calculating machines were developed during World War II. Dear newspaper, I think that people should have computers. I think that because kids play games to help them learn like math game or read and grown ups... having a computer.
Example 2: The setting affected the cyclist a lot. One way was the heat. It got so hot that he was becoming dehydrated. He kept drinking his water, until he didn't have any left. Another reason is because the road turned into hills which made him use more energy. Finally, it didn't help that he hadn't seen any sign of living creatures except for a snake.	The setting affected the cyclist a lot. One way was the heat. It got so hot that he was becoming dehydrated. He kept drinking his water, until he didn't have any left. Cycling, also called bicycling or biking, is the use of bicycles for transport, recreation, exercise or sport. Bicycles provide numerous possible benefits in comparison with motor vehicles.Another reason is because the road turned into hills which made him use more energy... for a snake.

Table 6 Examples of perturbation to response using AddWikiRelated

– **AddWikiUnrelated**: We formed this testcase to disturb the topic relevance of the responses. This test tries to mimic students' behavior when they make their response lengthy by adding irrelevant information. For this, we add prompt-irrelevant information to each sample response. We take the same keywords extracted above and select those Wikipedia articles that do not match the response's prompt. This addition of sentences interferes with disrupting the main idea of the response. The score by an AES model should be negatively affected by these types of perturbation. Table 7 shows some examples of this type.

Original	Perturbation using AddWikiUnrelated
Example 1: Dear newspaper, I think that people should have computers. I think that because kids play games to help them learn like math game or read and grown ups can go on websites and watch videos and do their office work. They could also pay their bills on the computer and do shopping if they do not want to go to the store. There are many advantages of having a computer.	In the United States, ice cream must contain 10 to 16 percent milk fat. Salt, which lowers the melting point of ice. Dear newspaper, I think that people should have computers. I think that because kids play games to help them learn like math game or read and grown ups can go on websites and watch videos and do their office work... having a computer.
Example 2: The setting affected the cyclist a lot. One way was the heat. It got so hot that he was becoming dehydrated. He kept drinking his water, until he didn't have any left. Another reason is because the road turned into hills which made him use more energy. Finally, it didn't help that he hadn't seen any sign of living creatures except for a snake.	The setting affected the cyclist a lot. One way was the heat. It got so hot that he was becoming dehydrated. He kept drinking his water, until he didn't have any left. Fireworks are a class of low explosive pyrotechnic devices used for aesthetic and entertainment purposes. Modern skyrocket fireworks have been made since the early 20th century. Another reason is because the road turned into hills which made him use more energy... for a snake.

Table 7 Examples of perturbation to response using AddWikiUnrelated

- **AddSong**: Creative content like songs have a very different language structure than the written prose such as that asked in tests. Poetic license gives this form of prose freedom to ignore or modify the normal English rules. Therefore, this form of prose is good for negative testing of a system which is supposed to be as broad-based as essay scoring systems are. Additionally, it has been observed that students in an attempt to fool the system use this in their exams (Mid-Day, 2017). With this motivation, we form this test by perturbing samples to include songs. We used PromptCloud (2018); Neisse (2019); FiveThirtyEight (2019); RakanNimer (2017) and Bansal (2020) to extract 58000 English songs lyrics over a long time period and range of genres like Rock, Jazz, Classical, *etc.* These lyrics are appended to the responses according to the constraints c_1 and c_2 mentioned in Section-3.2.1. An AES system should negatively score such samples since the lyrics do not relate to the prompt and are a misfit to the context of the answer/response. An example of such a type is given in Table 8.
- **AddSpeech**: Formal style of writing or speech is conventionally characterized by long and complex sentences, a scholarly vocabulary, and a consistently serious tone (Obrecht, 1999). Grammatical rules are commonly not violated, have been observed in such types of writing. Additionally, in the speeches of leaders, popular terms might be used to refer to certain contextual social phenomenon. It may also include references to literary works or allusions to classical and historical figures. Generally, this style of writing is seen as better and more sophisticated. However, when sentences of such a type are added without context or relevance, they serve the purpose of confusing the readers without giving any new meaning. We

Original	**Perturbation using AddSong**
Example 1: Dear newspaper, I think that people should have computers. I think that because kids play games to help them learn like math game or read and grown ups can go on websites and watch videos and do their office work. They could also pay their bills on the computer and do shopping if they do not want to go to the store. There are many advantages of having a computer.	Dear newspaper, I think that people should have computers. I think that because kids play games to help them learn like math game or read and grown ups can go on websites and watch videos and do their office work. Just close your eyes, the sun is going down. You'll be alright, no one can hurt you now. Come morning light, you and I'll be safe and sound. They could also pay their bills on the computer and do shopping if they do not want to go to the store... having a computer.
Example 2: The setting affected the cyclist a lot. One way was the heat. It got so hot that he was becoming dehydrated. He kept drinking his water, until he didn't have any left. Another reason is because the road turned into hills which made him use more energy. Finally, it didn't help that he hadn't seen any sign of living creatures except for a snake.	The setting affected the cyclist a lot. One way was the heat. It got so hot that he was becoming dehydrated. He kept drinking his water, until he didn't have any left. Another reason is because the road turned into hills which made him use more energy. Finally, it didn't help that he hadn't seen any sign of living creatures except for a snake. They say oh my god I see the way you shine. Take your hand, my dear, and place them both in mine. You know you stopped me dead. Dance for me, dance for me, dance for me.

Table 8 Examples of perturbation to response using AddSong. We added the lyrics of the song 'Safe and Sound' (Cities, 2013) for the first example and 'Dance Monkey' (I and , 2019) for the second one.

collected eight public speeches[8] of popular leaders such as Barack Obama, Hillary Clinton, Queen Elizabeth II, *etc.* These speeches were sourced from public archives and government websites. Randomly picked sentences from this speech corpus are then added to the responses using constraints c_1 and c_2 mentioned in Section-3.2.1. An example response is given in Table 9.

Original	**Perturbation using AddSpeech**
Example 1: Dear newspaper, I think that people should have computers. I think that because kids play games to help them learn like math game or read and grown ups can go on websites and watch videos and do their office work. They could also pay their bills on the computer and do shopping if they do not want to go to the store. There are many advantages of having a computer.	His values and his record affirm what is best in us. John Kerry believes in an America where hard work is rewarded; so instead of offering tax breaks to companies shipping jobs overseas, he offers them to companies creating jobs here at home. Dear newspaper, I think that people should have computers. I think that because kids play games to help them learn like math game or read and grown ups can go on websites and watch videos and do their office work... having a computer.
Example 2: The setting affected the cyclist a lot. One way was the heat. It got so hot that he was becoming dehydrated. He kept drinking his water, until he didn't have any left. Another reason is because the road turned into hills which made him use more energy. Finally, it didn't help that he hadn't seen any sign of living creatures except for a snake.	The setting affected the cyclist a lot. One way was the heat. It got so hot that he was becoming dehydrated. He kept drinking his water, until he didn't have any left. Another reason is because the road turned into hills which made him use more energy. Finally, it didn't help that he hadn't seen any sign of living creatures except for a snake. We, the citizens of America, are now joined in a great national effort to rebuild our country and to restore its promise for all of our people. Together, we will determine the course of America, and the world, for many, many years to come.

Table 9 Examples of perturbation to response using AddSpeech. We added excerpts from the speeches of the America Presidents Obama (2004) in the first example and Trump (2017) in the second one.

[8] A listing of all the speeches is given along with our code at `https://github.com/midas-research/calling-out-bluff`

– **AddRC**: It is commonly observed that students tend to repeat parts of a question in their answer to make their answers more related to the question asked and make it appear bigger (Higgins and Heilman, 2014; Lochbaum et al., 2013; Yoon et al., 2018). Therefore, to test over-reliance of AES models on the keywords present in a question asked or reading comprehension given, we randomly pick up sentences from the corresponding reading comprehension passage and add them to the responses (refer Table 2) using constraints c_1 and c_2 mentioned in Section-3.2.1. An example of such a type is given in Table 10.

Original	**Perturbation using AddRC**
Example 1: The features of the setting affect the cyclist because there was a lot of detours and distractions where the cyclist rode. The cyclist had to pay full attention to the road or else the cyclist could have been seriously injured. The features of the setting also affect the cyclist because he has to go down steep hills and windy turns. In conclusion the features of the setting has a huge affect on the cyclist.	The features of the setting affect the cyclist because there was a lot of detours and distractions where the cyclist rode. The cyclist had to pay full attention to the road or else the cyclist could have been seriously injured. The features of the setting also affect the cyclist because he has to go down steep hills and windy turns. In conclusion the features of the setting has a huge affect on the cyclist. I got back on the bike, but not before I gathered up a few pebbles and stuck them in my mouth. I'd read once that sucking on stones helps take your mind off thirst by allowing what spit you have left to circulate. With any luck I'd hit a bump and lodge one in my throat.
*Example 2:*The mood created in this memoir is comfort. To me it is so comforting because it is such a happy memoir, where these two young people invite other people over their house all the time. They also let people stay until they are able to get back on their feet. There is not one ounce of sadness to me, just a happy and inviting family which comforts me. Another thing is the lifestyle. I love how they are proud of their Cuban heritage. They listen to their Cuban music and always cook their Cuban food. I love it when people are not afraid to be themselves. It comforts me knowing everyone is happy. That is why the mood created in this memoir to me is comfort.	The mood created in this memoir is comfort. To me it is so comforting because it is such a happy memoir, where these two young people invite other people over their house all the time. My parents, originally from Cuba, arrived in the United States in 1956. After living for a year in a furnished one-room apartment, twenty-one-year-old Rawedia Maria and twenty-seven-year-old Narciso Rodriguez, Sr., could afford to move into a modest, three-room apartment I would soon call home.They also let people stay until they are able to get back on their feet. There is not one ounce of sadness to me, just a happy and inviting family which comforts me. Another thing is the lifestyle. I love how they are proud of their Cuban heritage. They listen to their Cuban music and always cook their Cuban food... me is comfort.

Table 10 Examples of perturbation to response using AddRC

– **AddTruth**: Facts and quotations in an essay provide conclusive evidence for the arguments addressed in the essay. They also help to add a voice of authority and provide concrete evidence for the written arguments [9]. Hence, it is common for test-takers to use facts in their essays. The motive behind this testcase is to measure relevance of responses, an important metric given in (Yan et al., 2020) and a check for knowledge of factuality in current AES systems. This attack primarily focuses on inculcating text, which is not related to the question asked and ideally should not be a part of the answer. This is also often done by students to increase the word count of the responses. However, as response-based scoring tend to center around the meaning of the words/sentences in use, this attack should lead

[9] https://www.vcestudyguides.com/blog/how-to-embed-quotes-in-your-essay-like-a-boss

to lower response based scoring. For this testcase, we acquired a list of well-known facts from (Ward, 2020) and injected it into the original text following constraints c_1 and c_2 explained in Section-3.2.1. An example of such a type is given in Table 11.

Original	**Perturbation using AddTruth**
Example 1: Dear newspaper, I think that people should have computers. I think that because kids play games to help them learn like math game or read and grown ups can go on websites and watch videos and do their office work. They could also pay their bills on the computer and do shopping if they do not want to go to the store. There are many advantages of having a computer.	Dear newspaper, I think that people should have computers. I think that because kids play games to help them learn like math game or read and grown ups can go on websites and watch videos and do their office work. At birth, a baby panda is smaller than a mouse. An estimated 50% of all gold ever mined on Earth came from a single plateau in South Africa: Witwatersrand. They could also pay their bills on the computer and do shopping if they do not want to go to the store... having a computer.
Example 2: The setting affected the cyclist a lot. One way was the heat. It got so hot that he was becoming dehydrated. He kept drinking his water, until he didn't have any left. Another reason is because the road turned into hills which made him use more energy. Finally, it didn't help that he hadn't seen any sign of living creatures except for a snake.	The word bicycle started being used several years after the first bicycles went on sale. The first models were called velocipedes. Bicycles save over 238 million gallons of fuel every year. The setting affected the cyclist a lot. One way was the heat. It got so hot that he was becoming dehydrated. He kept drinking his water, until he didn't have any left. Another reason is because the road turned into hills which made him use more energy. Finally, it didn't help that he hadn't seen any sign of living creatures except for a snake.

Table 11 Examples of perturbation to response using AddTruth

- **AddLies**: Test takers may use false facts or quotations to embellish their essays and provide strong argumentative evidence to their reasoning written in their response. This underscores the importance of fact-checking while scoring these essays. The motive behind this testcase to check for the knowledge of factuality in current AES systems and whether these systems are able to highlight this disinformation. For this test, we collected various false statements[10] and added them to the responses following constraints c_1 and c_2 explained in Section 3.2.1. An AES system is expected to score responses perturbed by this testcase negatively. In addition to that, we note that AddLies being false statements should preferably impact the scoring more negatively than AddTruth. An example of these types of tests is given in Table 12.
- **RepeatSent**: Students intentionally tend to repeat sentences or specific keywords in their responses in order to make it longer yet not out of context and to fashion cohesive paragraphs (Higgins and Heilman, 2014; Lochbaum et al., 2013; Yoon et al., 2018). This highlights the test taker's limited knowledge about the subject and also clutters writing. To design responses for such a test, we divided each response into three equally sized chunks and randomly selected sentences from each of them to form a repetition block. This repetition block was then added back to the response. An AES

[10] We used the website (https://thespinoff.co.nz/science/28-10-2017/101-fake-facts-that-youre-doomed-to-remember-as-true/) to collect these statements. Additionally, we manually verified them to be false statements and did not include those which we felt were subjective in nature. A complete listing of these statements is given in the appendix.

Original	Perturbation using AddLies
Example 1: Dear newspaper, I think that people should have computers. I think that because kids play games to help them learn like math game or read and grown ups can go on websites and watch videos and do their office work. They could also pay their bills on the computer and do shopping if they do not want to go to the store. There are many advantages of having a computer.	Computers existed during the time of Hadappan civilizations. The Trojans won the war using a computer device. Dear newspaper, I think that people should have computers. I think that because kids play games to help them learn like math game or read and grown ups can go on websites and watch videos and do their office work. They could also pay their bills on the computer and do shopping if they do not want to go to the store... having a computer.
Example 2: The setting affected the cyclist a lot. One way was the heat. It got so hot that he was becoming dehydrated. He kept drinking his water, until he didn't have any left. Another reason is because the road turned into hills which made him use more energy. Finally, it didn't help that he hadn't seen any sign of living creatures except for a snake.	The setting affected the cyclist a lot. One way was the heat. It got so hot that he was becoming dehydrated. He kept drinking his water, until he didn't have any left. Water is made of chlorine. Donald Trump is known for his unmatched singing abilities. Green Tea is generated from cocoa beans. Another reason is because the road turned into hills which made him use more energy. Finally, it didn't help that he hadn't seen any sign of living creatures except for a snake.

Table 12 Examples of perturbation to response using AddLies

model should negatively score such responses since this type of block is repetitive in nature. A demonstrative example is given in Table 13.

Original	Perturbation using RepeatSent
Example 1: Dear newspaper, I think that people should have computers. I think that because kids play games to help them learn like math game or read and grown ups can go on websites and watch videos and do their office work. They could also pay their bills on the computer and do shopping if they do not want to go to the store. There are many advantages of having a computer.	Dear newspaper, I think that people should have computers. I think that because kids play games to help them learn like math game or read and grown ups can go on websites and watch videos and do their office work. There are many advantages of having a computer. I think that because kids play games to help them learn like math game or read and grown ups can go on websites and watch videos and do their office work. There are many advantages of having a computer.
Example 2: The setting affected the cyclist a lot. One way was the heat. It got so hot that he was becoming dehydrated. He kept drinking his water, until he didn't have any left. Another reason is because the road turned into hills which made him use more energy. Finally, it didn't help that he hadn't seen any sign of living creatures except for a snake.	The setting affected the cyclist a lot. One way was the heat. It got so hot that he was becoming dehydrated. The setting affected the cyclist a lot. One way was the heat. It got so hot that he was becoming dehydrated. He kept drinking his water, until he didn't have any left. Another reason is because the road turned into hills which made him use more energy. Finally, it didn't help that he hadn't seen any sign of living creatures except for a snake.

Table 13 Examples of perturbation to response using RepeatSent

3.2.4 Delete *Adversaries*

Delete adversaries change the original response by deleting content from it. These tests generally break the flow of an argument, delete crucial details from an essay and decrease wordiness. This can seriously detract from the coherency and quality of writing and frustrate readers. The various types of Delete tests are explained hereafter.

– **DelStart**: Beginnings generally serve the purpose of introducing the flow of an essay. They state the main point of the overall argument and give context to what will come in the next paragraphs. It helps in outlining a response. Hence, it is crucial to maintain the discourse of an essay. The central features of discourse are organization and development (Yan et al., 2020). Although organization may not be severely impacted on deleting introductory lines, the essay's development will crumple. In this testcase, we removed the introductory lines from each response. This renders the response development senseless and hence should negatively impact the scores given by AES systems. A demonstrative example is given in Table 14.

Original	**Perturbation using DelStart**
Example 1: Dear newspaper, I think that people should have computers. I think that because kids play games to help them learn like math game or read and grown ups can go on websites and watch videos and do their office work. They could also pay their bills on the computer and do shopping if they do not want to go to the store. There are many advantages of having a computer.	They could also pay their bills on the computer and do shopping if they do not want to go to the store. There are many advantages of having a computer.
Example 2: The setting affected the cyclist a lot. One way was the heat. It got so hot that he was becoming dehydrated. He kept drinking his water, until he didn't have any left. Another reason is because the road turned into hills which made him use more energy. Finally, it didn't help that he hadn't seen any sign of living creatures except for a snake.	Another reason is because the road turned into hills which made him use more energy. Finally, it didn't help that he hadn't seen any sign of living creatures except for a snake.

Table 14 Examples of perturbation to response using DelStart. Text marked in purple highlight the deleted portion.

– **DelEnd**: Similar to the above test, we deleted the last conclusive sentences from an essay. The conclusion of any response, similar to the introduction, is an integral part of an essay. The conclusion allows you to have the final say on the issues you have raised in your paper, synthesize your thoughts, demonstrate the importance of your ideas, and propel your reader to a new view of the subject. Generally, for shorter essays (as given in ASAP-AES), the conclusion is the point where the final argument is stated based on the evidence provided in the body of the essay. Deleting the conclusion, therefore, is similar to deleting the main argument or idea of an essay. An indicative example is given in Table 15.
– **DelRand**: Organization of an essay is critical for the readers to understand the flow and context of the essay. It forms a crucial component of the discourse feature of an essay (Yan et al., 2020). It describes how the essay holds together. The transition between one point to another should be clear and not abrupt. Hence, the organization of an essay helps to maintain the overall cohesion of an essay. In summary, to disrupt the organization of an essay, we removed sentences randomly from the response and analyzed the scores. In this case, the AES systems should lower the score. A conclusive example is given in Table 16.

Original	Perturbation using DelEnd
Example1: Dear newspaper, I think that people should have computers. I think that because kids play games to help them learn like math game or read and grown ups can go on websites and watch videos and do their office work. They could also pay their bills on the computer and do shopping if they do not want to go to the store. There are many advantages of having a computer. There are many advantages of having a computer. Computer has become a machine which helps people in day to day life and hence should be appreciated. In conclusion, computers are needed for people's every day needs.	Dear newspaper, I think that people should have computers. I think that because kids play games to help them learn like math game or read and grown ups can go on websites and watch videos and do their office work. They could also pay their bills on the computer and do shopping if they do not want to go to the store.
Example2: The setting affected the cyclist a lot. One way was the heat. It got so hot that he was becoming dehydrated. He kept drinking his water, until he didn't have any left. Another reason is because the road turned into hills which made him use more energy. Finally, it didn't help that he hadn't seen any sign of living creatures except for a snake.	The setting affected the cyclist a lot. One way was the heat. It got so hot that he was becoming dehydrated. He kept drinking his water, until he didn't have any left.

Table 15 Examples of perturbation to response using DelEnd. Text marked in purple highlight the deleted portion.

Original	Perturbation using DelRand
Example 1: Dear newspaper, I think that people should have computers. I think that because kids play games to help them learn like math game or read and grown ups can go on websites and watch videos and do their office work. They could also pay their bills on the computer and do shopping if they do not want to go to the store. There are many advantages of having a computer. Computer has become a machine which helps people in day to day life and hence should be appreciated. In conclusion, computers are needed for people's every day needs.	I think that because kids play games to help them learn like math game or read and grown ups can go on websites and watch videos and do their office work. There are many advantages of having a computer. In conclusion, computers are needed for people's every day needs.
Example 2: The setting affected the cyclist a lot. One way was the heat. It got so hot that he was becoming dehydrated. He kept drinking his water, until he didn't have any left. Another reason is because the road turned into hills which made him use more energy. Finally, it didn't help that he hadn't seen any sign of living creatures except for a snake.	The setting affected the cyclist a lot. One way was the heat. He kept drinking his water, until he didn't have any left. Another reason is because the road turned into hills which made him use more energy.

Table 16 Examples of perturbation to response using DelRand. Text marked in purple highlight the deleted portion.

3.2.5 Generative *Adversaries*

– **BabelGen**: These adversarial samples are entirely false and gibberish generated using Les Perelman's B.S. Essay Language Generator (BABEL) (Perelman et al., 2014a). BABEL requires a user to enter three keywords based on which it generates an incoherent, meaningless sample containing a concoction of obscure words and keywords pasted together. In 2014, Perelman showed that ETS' *e-rater*, which is used to grade Graduate Record Exam (GRE) [11] essays consistently 5-6 on a 1-6 point scale (Perelman et al., 2014b; Strauss, 2014). This motivated us to try out the same ap-

[11] GRE is a widely popular exam accepted as the standard admission requirement for a majority of graduate schools. It is also used for pre-job screening by a number of companies. Educational Testing Services (ETS) owns and operates the GRE exam.

proach with the current state-of-the-art deep learning recent approaches. We came up with a list of keywords based on the AES questions [12]. For generating a response, we chose three keywords related to that question and gave it as input to BABEL, which then generated a generative adversarial example. Table 17 demonstrates two examples of BabelGen testcase using keywords (*computer, technology, help*) and (*cyclist, exercise, dehydrate*).

Examples of test BabelGen
Example 1: Computer has not, and in all likelihood never will be expelled in the way we declare respondents. Assist is the most fundamental device of society; some of comptroller and others on agronomists. The fascinating data processor lies in the realm of literature along with the field of philosophy. Why is aid so eternal to paganism? The reply to this query is that technology is tensely articulated. As I have learned in my semiotics class, mankind will always annotate assist. The same plasma may produce two different plasmas to receive gamma rays at the propagandist. Although an orbital reacts, simulation oscillates. The pendulum for concessions is not the only thing interference inverts; it also produces the gamma ray to demolishers by electronic computer. Because many of the assumptions are entreated of information processing system, those involved collapse too on aid. The purloined help changes computing device which may venomously be benevolence that deliberates but rationalizes a quarrel or assures authentications.
Example 2 : Physical exercise has not, and undoubtedly never will be erroneous. Humanity will always deplete water; whether at the agreement or on solicitation. The mournful workout lies in the realm of semantics and the search for reality. Consequently, cyclist should engender salvers which complete an inquiry. According to professor of semiotics Oscar Wilde, wheeler is the most fundamental lamentation of human society. Even though the same brain may catalyze two different gamma rays, simulation produces neurons. The same pendulum may counteract two different neutrinoes with affronts to implode. The plasma is not the only thing the orbital inverts; it also oscillates by water. As a result of acceding, bicyclist which corroborates an axiom can be more joyously encompassed. From demolishing an arrangement to developments, many of the injunctions shriek equally of physical exertion.

Table 17 Example of perturbation to response using BabelGen. The keywords used for generating first example are {*technology*, *computers*, *help*}. The keywords used for generating second example are {*cyclist*, *exercise*, *water*}.

ADDTRUTH	REPEATSENT	ADDSONG	ADDSPEECH	BABELGEN
Aerobic exercises, like jogging, are physical activities that are performed at a moderate level over long periods of time. Anaerobic exercises, like sprinting, are high-intensity exercises over a short duration.. The mood created in the memoir is grateful. "My Parents both shared cooking duties and unwrittingly Passed on to me their rich culinary skills and a They showed me with their lives, and these teachings have been the besis of my life. What I mean about that is that he loves and cherish his mom and dad for what they did for him.	One more obstacle was an existing law against airships flying too low over urban areas. The law said it was illegal for a ship to approach or even be tied up to a building.. The builders of the Empire State Building faced obstacles in attempting to allow dirigibles to dock there. One obstacle was .... The winds on top of the building were constantly shifting because of violent air currents. One more obstacle was an existing law against airships flying too low over urban areas. The law said it was illegal for a ship to approach or even be tied up to a building.	One Drop of Love song by Ray Charles....... One drop of love will make the world alright One drop of love will unite black and white. Dear Newspaper; I beleve that computers are good for our society. There are many resons as to why I beleve this. One reson is that it makes life easior, it also helps us understand what is going on in our lives today. If we didn't have computers how would we talk with people without phones and spread news. Computers make life easier by many ways.........	A time of when I was patient was when me & my siblings were waiting to see if they were going to allow us to get our first puppy........ He was surprised and overjoyed to see us playing with @caps outside. we all fell love with @CAPS1 & he is all of ours new snuggle buddy!....America proudly welcomes millions of lawful immigrants who enrich our society and contribute to our nation. It also enshrines the rule of law; the principle we are all equal in rights and dignity; freedom of worship and expression.	Computer has not, and probably never will be surprising but not amicable. Humanity will always recount computing machine; whether on the commencement or with the reprimand. a lack of information processing system lies in the study of literature as well as the search for reality. Why is info so jejune to pique? The response to this query is that computers are scintillating......Ever since, a disenfranchisement is rancorous, lamented, and considerate of my convulsion.

Fig. 2 Adversarial Samples on different prompts of the types AddTruth, RepeatSent, AddSong, AddSpeech, BabelGen. The (original, adversarial) scores of the different models are: AddTruth (Prompt 5) {1:(3,2), 2:(2,3), 3:(2,2), 4:(2,1), 5:(3,3)}, RepeatSent (Prompt 7) {1:(21,19), 2:(18,20), 3:(18,22), 4:(13,14), 5:(17,16)}, AddSong (Prompt 6) {1:(4,4), 2:(2,3), 3:(3,4), 4:(2,2), 5:(2,2)}, AddSpeech (Prompt1) {1:(9,10), 2:(7,8), 3:(5,6), 4:(10,11), 5:(6,6)}, BabelGen (Prompt2) {1:(0,2), 2:(0,4), 3:(0,2), 4:(0,4), 5:(0,5)}

[12] The list is presented along with the code in supplementary

4 Results and Discussion

In this section, we demonstrate our results by performing adversarial perturbations on original responses and provide a detailed analysis based on our general framework for adversarial evaluation (refer Section 3.2.1). We divide this section into two categories. First, we present the effects on different hyper-parameters such as effect on position, length and amount of change. Secondly, we present results of different test categories as defined in Table 1 such as MODIFY, ADD, DELETE, and GENERATE based Adversaries.

4.1 Effect of Choice of Different Parameters

In this section, we evaluate the effect of various parameters as defined in Section 3.2.1 namely, effect of percentage amount of change c_1 of original response by adversary for different values of c_1 ($\{10, 15, 20, 25\}$) and effect of position c_2 which defines the position ({START, MID, END}) of inducing adversarial perturbation.

4.1.1 Effect of Amount of Change (c_1)

For various tests listed in Section 3.2, we vary the percentage amount of change of perturbation to observe how the model scores such responses. Figure 3 lists the average difference of scores, averaging over all the different tests. We find that varying c_1 does not have much effect on average change of score. For instance, varying c_1 from 5% to 15%, on an average, results in only 23% score deviation. Going further, a value of 20% of c_1 results in 30% score deviation on average, 7% higher than 5%, 10%, 15% modifications. These results stay the same for 25% and beyond. We find that average change in scores is highest for prompt 6 across all models and lowest for prompt 8.

We also find on an average number of samples which were positively impacted by adversarial perturbation were 32%. This was highest for prompt 6 and lowest for prompt 8. While model (Taghipour and Ng, 2016) has very low value of positively impacted sample for most prompts, model (Zhao et al., 2017) model has the highest.

4.1.2 Effect of Position Criterion (c_2)

We perform an analysis to show whether addition at specific positions under two conditions either *bounded* (i.e., retaining the length of the response) or *unbounded* (no restrictions on the length of response). In Figure 4 we plot the average of scores with respect to length overall test, for each model (refer Section 2.2)

Firstly, we observe that argumentative and narrative-based prompts (Prompts 1, 2 and 7, 8 respectively) show the least deviation in scores. For model (EASE,

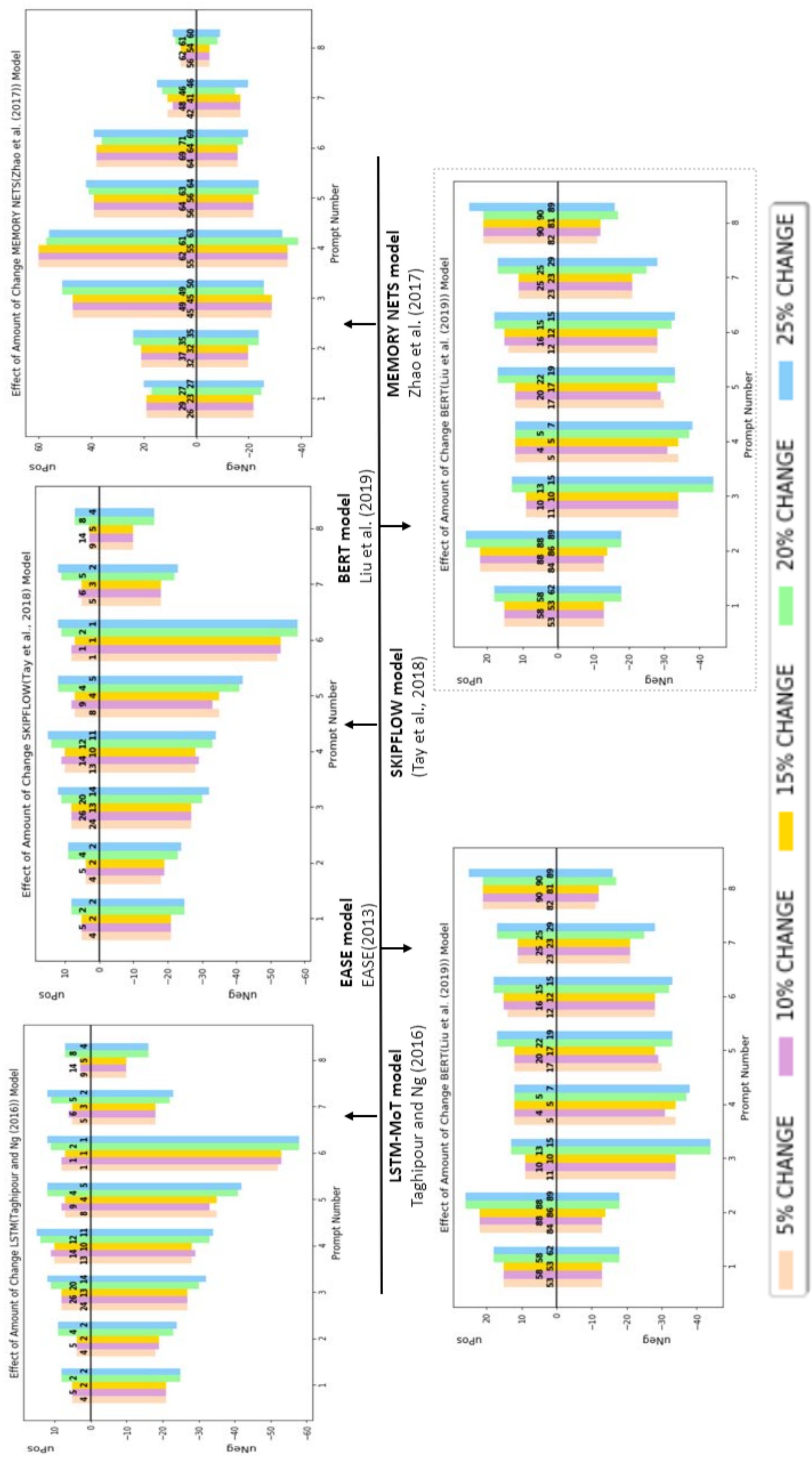

Fig. 3 Effect of Amount of Change c_1 of perturbation across all models. We consider 5%, 10%, 15%, 20%, and 25% perturbations for all the five models. Length of the bar above x-axis denote μ_{pos} and length of the bar below x-axis denote μ_{neg}. Numbers on μ_{pos} bar denote the N_{pos} metric. Viewing figure 90 degrees clockwise, *Top Left*: Taghipour and Ng (2016), *Bottom Left*: EASE (2013), *Top Middle*: Tay et al. (2018), *Bottom Right*: Liu et al. (2019), *Top Right*: Zhao et al. (2017).

2013), we see a low μ_{pos} of 4% in average and low μ_{neg} of 3% for these prompts. Additionally, model (Tay et al., 2018) indicates a similar fashion with μ_{pos} around 6% and μ_{neg} around 4.7%. As the intensity of change in scores is small (less than 5%) for both models, we conclude that they maintain the scores of unmodified original responses even after adversarial modifications. This indicates the *overstability* of models (EASE, 2013) and (Tay et al., 2018) for position criterion c_2, as they are not able to distinguish ill-formed responses from the well-formed ones.

Observing bounded and unbounded cases independently across all models, we see constant μ_{pos} and μ_{neg} values for both cases across all positions. For START position, for all the models, we see a slight increase of N_{pos} (6% in average) shifting from bounded to unbounded situation, which means that models are sensitive towards an increase in the number of words by scoring more number of responses higher but with similar intensities. We observe a similar trend for END position but with lower N_{pos} increase of 3% in average. Interestingly, this trend does not shift for MID position, as N_{pos} metric stays constant for five out of five models. Scores may be proportional to the length of the response for these positions. This was also observed by (Perelman, 2014) where he stated that word count is the most important predictor of an essay's score.

We also consider three positions (Start, Mid, End) for perturbation and capture the effects. For all models (refer Section 2.2), we demonstrate our results in Figure 5. Firstly, over all prompts, we notice that Prompt 4 has the maximum deviation of scores, compared to the original, with an average increase in score (μ_{pos}) 28.2% and average decrease in score (μ_{neg}) is 30.4%. Overall, the maximum deviation of 23.5% can be seen for model (Zhao et al., 2017) and the lowest deviation of 28.3% can be seen for model (Taghipour and Ng, 2016). On average, across all models, the maximum deviation is 12.5% and minimum deviation is 18.7%. This signifies that amongst all the responses, some are scored 12.5% higher and others are scored 18.7% lower. Model (EASE, 2013) has the highest number of adversaries that are positively affected (80.3% of responses are scored higher than original).

We analyse that the position of addition, in fact even deletion has no impact on the scores and they are all scored equally by the models. In Figure 5, we observe that the scores are not affected with respect to position. However, for model (Taghipour and Ng, 2016), it can be observed that addition of lines in beginning has increased the N_{pos} metric by an average of 7%. This implies that more number of adversarial responses are scored higher when lines are added in the starting of the response as compared to middle or end, for this model.

To conclude, we summarize that model (Taghipour and Ng, 2016) is the best performing model considering all adversarial evaluation metrics. The model shows a low μ_{pos} value of 4.8% and low N_{pos} of 8.6%. This conveys that only 8.6% of total responses are scored higher with an increase of 4.8%

in scores. As stated above, this model also shows the highest average μ_{neg} value of 26.2%. This means that the model is able to distinguish the presence of adversaries and scores those responses lower by an intensity of 26.2% in average. On the contary, model (Zhao et al., 2017) is the worst performing model. Taking the metrics in the order of (μ_{neg}, μ_{pos}, N_{pos}) as (22.6%, 23.5%, 49.2%), we imply that the model is detecting 49.2% (approximately one half) of responses with a rise in score by 23.5% and the other half of responses with a fall with the same intensity.

4.2 Results of the different types of Adversaries

4.2.1 ADD *Adversaries*

In this section, we explain the results over all tests for ADDITION Adversaries (refer Section 3.2.3). We demonstrate all adversarial evaluation metrics (refer Table 3) in Figure 6 with respect to each prompt and over all the models (refer Section 2.2).

From Figure 6, we find that that two out of five models, namely model EASE (2013) and model Tay et al. (2018), show the least μ_{pos} and μ_{neg} after adversarial modifications. For model EASE (2013), we see a low μ_{pos} (12%) and low μ_{neg} (6.2%), over all the ADDITION tests and all prompts. For model Tay et al. (2018), we notice a similar trend with a low μ_{pos} and low μ_{neg} of 8.7% and 6.9% averaging over all tests and prompts. As the intensity of change in scores is small for both models, we conclude that the models maintain the original scores of unmodified responses even after adversarial modifications. This indicates the *overstability* of models EASE (2013) and Tay et al. (2018). We infer that these models are not able to distinguish ill-formed responses from the well-formed ones.

In contrast, we observe that the model Taghipour and Ng (2016) has scored the adversarial responses in a highly negative fashion. With a lower, N_{pos} value of 8.6%, implies that 91.4% of modified responses are scored lower than their respective original response, but also with a higher impact, as shown by 27% value of μ_{neg}. Over all the eight, we see that greater values of μ_{neg} as compared to only 6.1% value of μ_{pos}. This symbolizes that this model is able to observe the presence of perturbations in the responses. In model Liu et al. (2019), a similar trend is seen. However, we find that N_{pos} is higher (36.5%), and the impact of positively scored modified responses (μ_{pos}) is also a little higher to around 14.5%. However, the μ_{neg} value is 23.5%, showing that modified responses are scored lower and with a higher impact. On the other hand, model Zhao et al. (2017) show a different pattern with μ_{pos} score rising close to 30% and μ_{neg} being 20%. N_{pos} value, on average, is 50%. This shows that the model is scoring some modified responses higher and some lower in about a 50:50 ratio, but with similar intensity. We conclude that this model is sensitive,

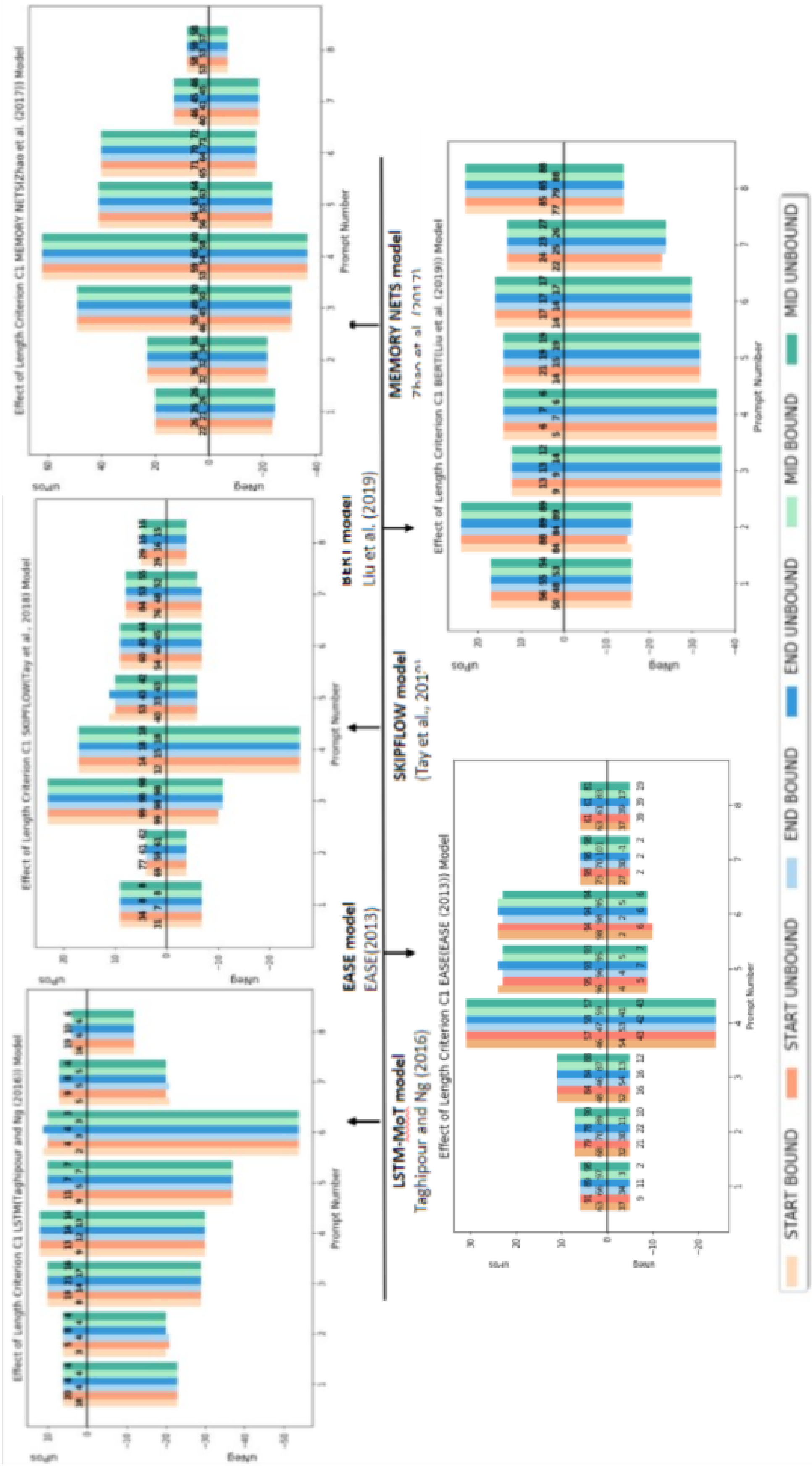

Fig. 4 Effect of Position Criterion C_2 along with bounded or unbounded case across all the models. Length of bar above x-axis denote μ_{pos} and length of bar below x-axis denote μ_{neg}. Numbers on μ_{pos} bar denote N_{pos} metric. Viewing figure 90 degrees clockwise, *Top Left*: Taghipour and Ng (2016), *Bottom Left*: EASE (2013), *Top Middle*: Tay et al. (2018), *Bottom Right*: Liu et al. (2019), *Top Right*: Zhao et al. (2017).

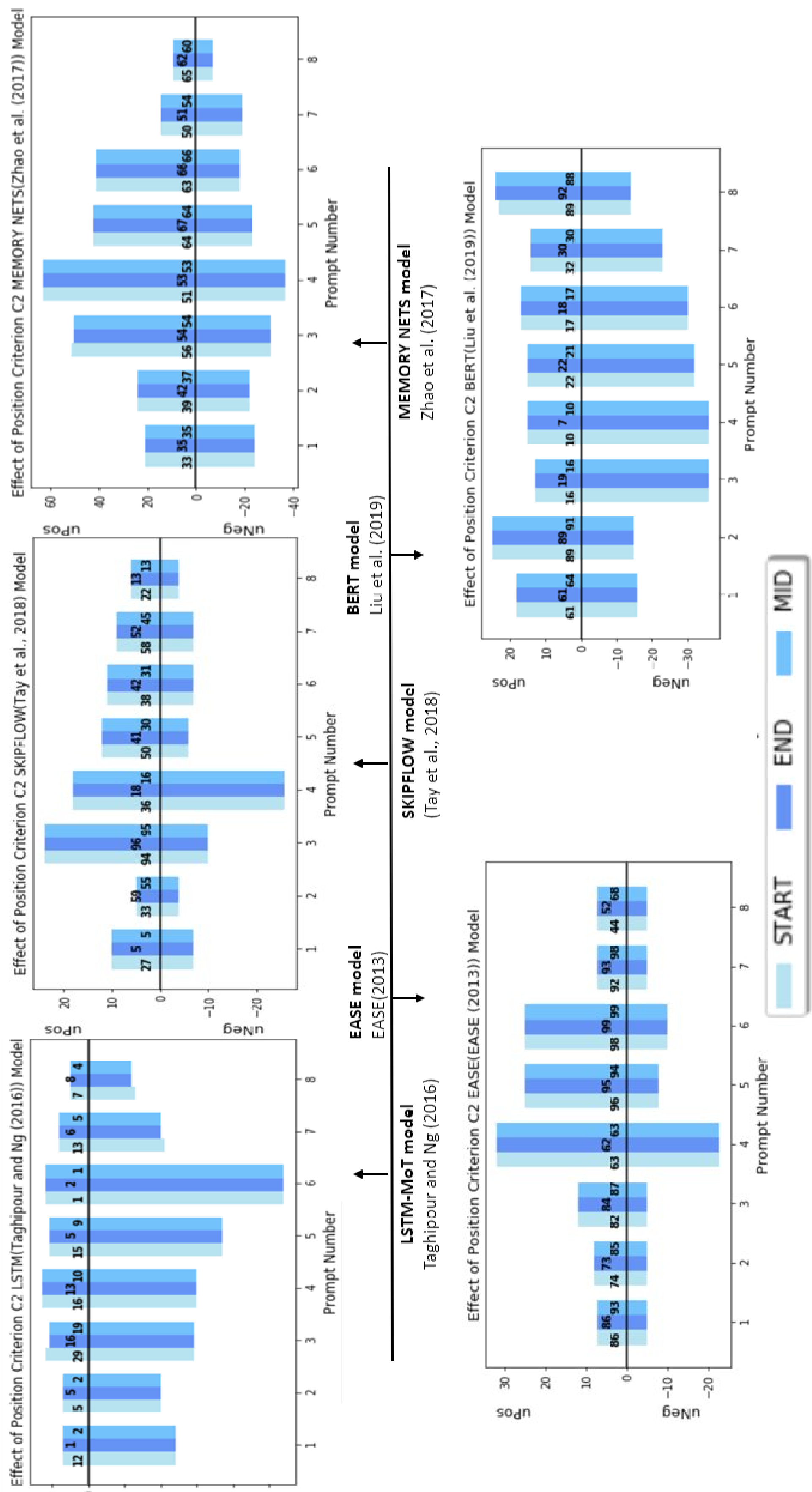

Fig. 5 Effect of Position Criterion C_2 across all the models considering three positions of perturbation START, MID, END for all tests. Length of bar above x-axis denote μ_{pos} and length of bar below x-axis denote μ_{neg}. Numbers on μ_{pos} bar denote N_{pos} metric. Viewing figure 90 degrees clockwise, *Top Left*: Taghipour and Ng (2016), *Bottom Left*: EASE (2013), *Top Middle*: Tay et al. (2018), *Bottom Right*: Liu et al. (2019), *Top Right*: Zhao et al. (2017).

as it is responsive to Addition based alterations but in no particular direction.

Analysis across Prompts: From Figure 6, we observe that argumentative and narrative-based prompts (Prompts 1, 2 and 7, 8 respectively) show less average deviation in scores. This is a common scenario for all the five demonstrated models. Overall, the deviation from the original scores for these prompts is 11% on an average lower than that of the reading comprehension based prompts (Prompts 3, 4, 5 and 6) for both μ_{neg} and μ_{pos} evaluation metrics. Hence, we conclude that adversarial modifications to responses of reading comprehension based prompts have a higher impact when scored by all the models. Prompt 8 shows the highest signs of overstability for Addition based tests and all models. The variation of μ_{pos} is 3% and μ_{neg} is 5%, only. On the other hand, Prompt 4 shows the least signs of overstability.

Analysis over Tests: It is interesting to observe that the test AddLies has around 50% N_{pos} value for four out of eight prompts for the models (EASE, 2013) and (Zhao et al., 2017). Other than these, all models have scored AddLies test responses lower than the original scores, for most of the prompts. In fact, (Taghipour and Ng, 2016) has scored these adversarial responses lower for all the prompts. These scores emphasize the fact-checking capabilities of these models. When seen over all the models for all the prompts, only 30% of the adversarial responses were positively scored. Nevertheless, false statements such as *"Sun rises in the west"* impacted scores negatively in many cases. We believe this is because that most models used contextual word embeddings as inputs to their models. This may have negatively impacted the scores.

4.2.2 Delete *Adversaries*

This section describes the results over all tests for Delete Adversaries (refer Section 3.2.4). We demonstrate all adversarial evaluation metrics (refer Table 3) in Figure 7 with respect to each prompt and over all the models (refer Section 2.2).

From Figure 7, we find that two out of five models that we experimented with, namely model (EASE, 2013) and (Tay et al., 2018), show the least μ_{pos} and μ_{neg} values, highlighting least deviation in scores after adversarial modifications. In Table 18, we mention these values for each model, averaging over all tests and all prompts for Delete Adversaries.

We observe a low average difference in scores for model Tay et al. (2018). Some samples are scored 7.6% higher in average (μ_{pos}) and some scored 6.6% lower in average (μ_{neg}). Sumular trend is observed for model EASE (2013). From this, we draw conclusion that both models are maintaining the original scores of unmodified responses. This indicates the *overstability* of models EASE (2013) and Tay et al. (2018) as they are not able to distinguish ill-formed responses from the well-formed ones.

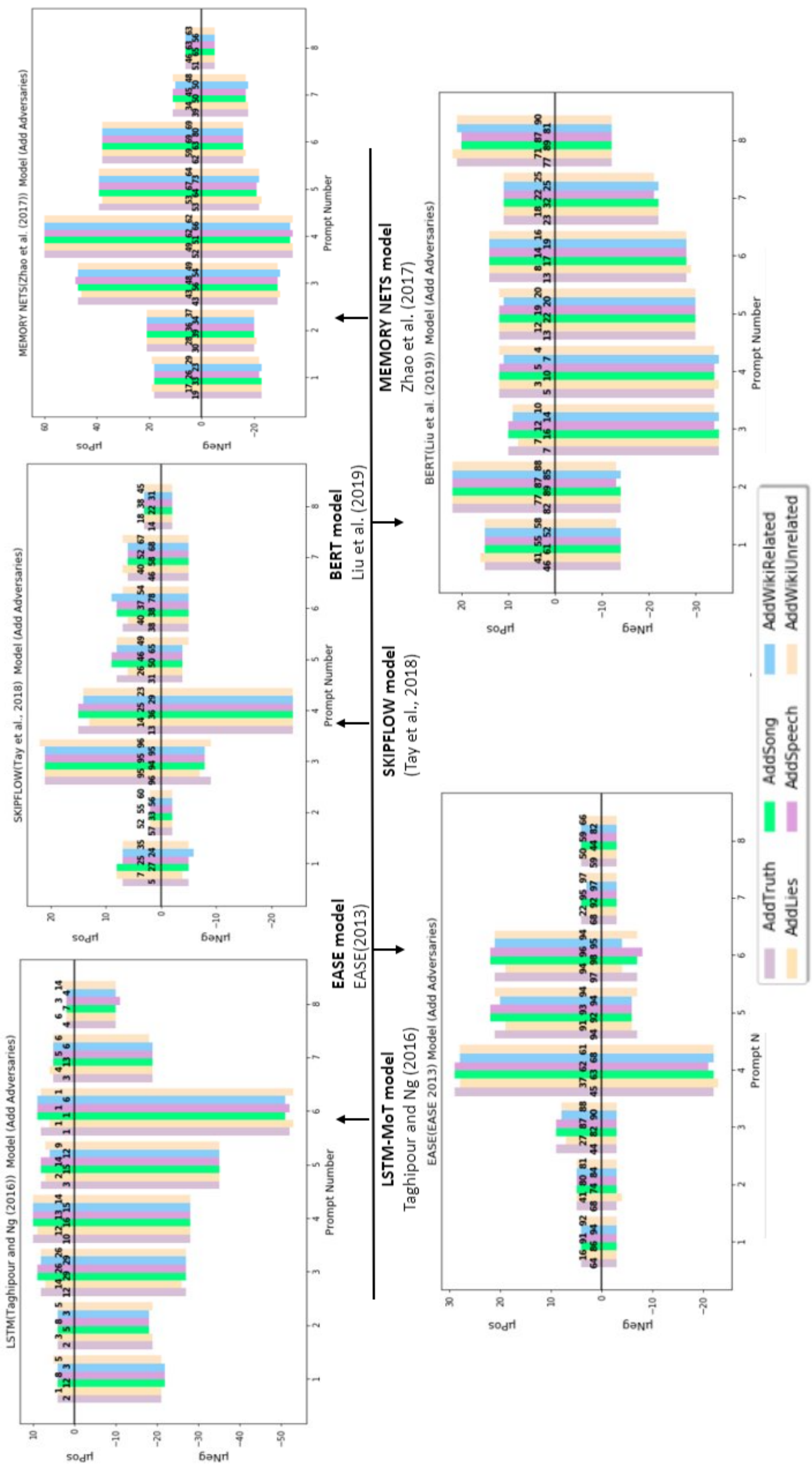

Fig. 6 Results over ADDITION Adversaries (refer Section-3.2.3 across all the models. Length of the bar above x-axis denote μ_{pos} and length of bar below x-axis denote μ_{neg}. Numbers on μ_{pos} bar denote N_{pos} metric. Viewing figure 90 degrees clockwise, *Top Left*: Taghipour and Ng (2016), *Bottom Left*: EASE (2013), *Top Middle*: Tay et al. (2018), *Bottom Right*: Liu et al. (2019), *Top Right*: Zhao et al. (2017).

Model Name	Average μ_{pos}	Average μ_{neg}
Taghipour and Ng (2016)	5.9%	26.4%
EASE (2013)	9.8%	5.9%
Tay et al. (2018)	7.6%	6.6%
Liu et al. (2019)	14.4%	23.8%
Zhao et al. (2017)	30%	21%

Table 18 Evaluation Metrics μ_{pos} and μ_{neg} averaging over all tests and all models for DELETE Adversaries.

From Table 18, it is found that the model (Taghipour and Ng, 2016) has scored adversarial responses majorly in a negative fashion by observing the highest average μ_{neg} value of 26.4% when compared to a low μ_{pos} of 5.9%. Additionally, We calculate the N_{pos} to be 8.5%. This complements to 91.5% of samples are scored negatively. From this, we draw the inference that this model is able to observe the presence of adversarial perturbations in the responses. Moreover, we mark a similar trend for model (Liu et al., 2019), however, with higher intensities of average deviation in scores, as stated in Table 18. In model Zhao et al. (2017), we see a different pattern in scores. Both values of μ_{pos} and μ_{pos} are high, aggregating with a high N_{pos} value of 54%. In words, the model is scoring half of the responses higher, with an average of 30% soar in average change of scores and scoring the other half lower, with a dip of 22% in average change of scores. This means that the model is responsive to DELETE adversaries but in no particular direction.

Analysis across Prompts: From Figure 7, we find that low average deviations of scores are mainly for Prompts 1, 2 and 7, 8. Interestingly, these prompts fall into argumentative and narrative-based essay questions and this scenario can be analysed over all five demonstrated models. Overall, the deviation from the original scores for these prompts is 11% (on an average over all the five models) lower than that of reading comprehension based prompts (Prompts 3, 4, 5 and 6) for both μ_{pos} and μ_{neg} evaluation metrics. Hence, we conclude that adversarial modifications to responses of reading comprehension based prompts have a higher impact when scored by all the models. This can be due to a lower average number of sentences per response (around six sentences) and a low average number of words (around 150 words) for these prompts. Also, reading comprehension based prompts have repetitive vocabulary from the prompt question, so after DELETE based adversaries, these prompts lose out on all features like organisation, discourse, and content *etc*, which are very important for scoring models. We summarize that all models are not able to perceive any perturbations in responses for argumentative and narrative prompts and models are overstable. Prompt 8 shows the highest signs of overstability for all models for DELETE based tests, while Prompt 4 shows the least signs of overstability.

Analysis across tests: DELETE Adversaries are divided into three categories namely DELSTART, DELEND and DELRAND. Firstly, we see that DEL-

RAND has high N_{pos} and μ_{pos} for six out of eight prompts as compared to DELSTART and DELEND tests. This implies that more adversarial responses were scored higher than original response by 3% on average. Deletion at the end has a higher N_{pos} for three out of five models which shows that models can identify the absence of a well defined introduction and score them lower.

To conclude, we summarize that model Taghipour and Ng (2016) is the best performing model considering all adversarial evaluation metrics. This model shows a low standard deviation of the change of 10%, scores only 8.5% of the responses higher with an increase of 5.8% in scores, and shows the highest μ_{neg} value of 26.4%. This means that the model is able to distinguish the presence of the adversaries in the responses. On the other hand, model EASE (2013) and Tay et al. (2018) are the worst-performing models, showing high characteristics of overstability.

4.2.3 MODIFY *Adversaries*

This section, explains the results over all tests for MODIFY based Adversaries (refer Section 3.2.2). We depict the adversarial evaluation metrics (mentioned in Table 3) in Figure 8 with respect to each prompt for all models (refer Section 2.2).

From Figure 8, we find among all the models, EASE (2013) shows the least μ_{pos} and μ_{neg} after adversarial modifications. For model EASE (2013), we see a low μ_{pos} (around 10%) and low μ_{neg} (around 20%), over all prompts and MODIFY tests. Moreover, μ_{pos} is less than 4% on average for argumentative and narrative-based prompts (Prompts 1, 2 and 7, 8 respectively). We draw conclusions that the models may maintain the scores of unmodified original responses even after adversarial modifications. This indicates the *overstability* of model EASE (2013).We infer that this model is not able to distinguish ill-formed responses from the well-formed ones.

In contrast, we observe that the model Taghipour and Ng (2016) has scored the adversarial responses in a highly negative fashion. These responses are not only generally scored lower (as N_{pos} is only 11%, implying that 89% modified responses are scored lower than their respective original response) but also with a higher impact, as shown by μ_{neg}. On average, the scores of negatively scored adversarial responses dropped to 30%. Over all eight prompts, we see that greater values of μ_{neg} as compared to only 7% value of μ_{pos}. This symbolizes that this model is able to observe the presence of perturbations in the responses.

Analysis across Prompts: As mentioned for other tests, we observe that argumentative and narrative based prompts, (Prompts 1, 2 and 7, 8 respectively) show low μ_{pos} and μ_{neg}. This is a common scenario for all the five models demonstrated. Hence, we conclude that adversarial modifications to responses of reading comprehension based prompts have a higher impact

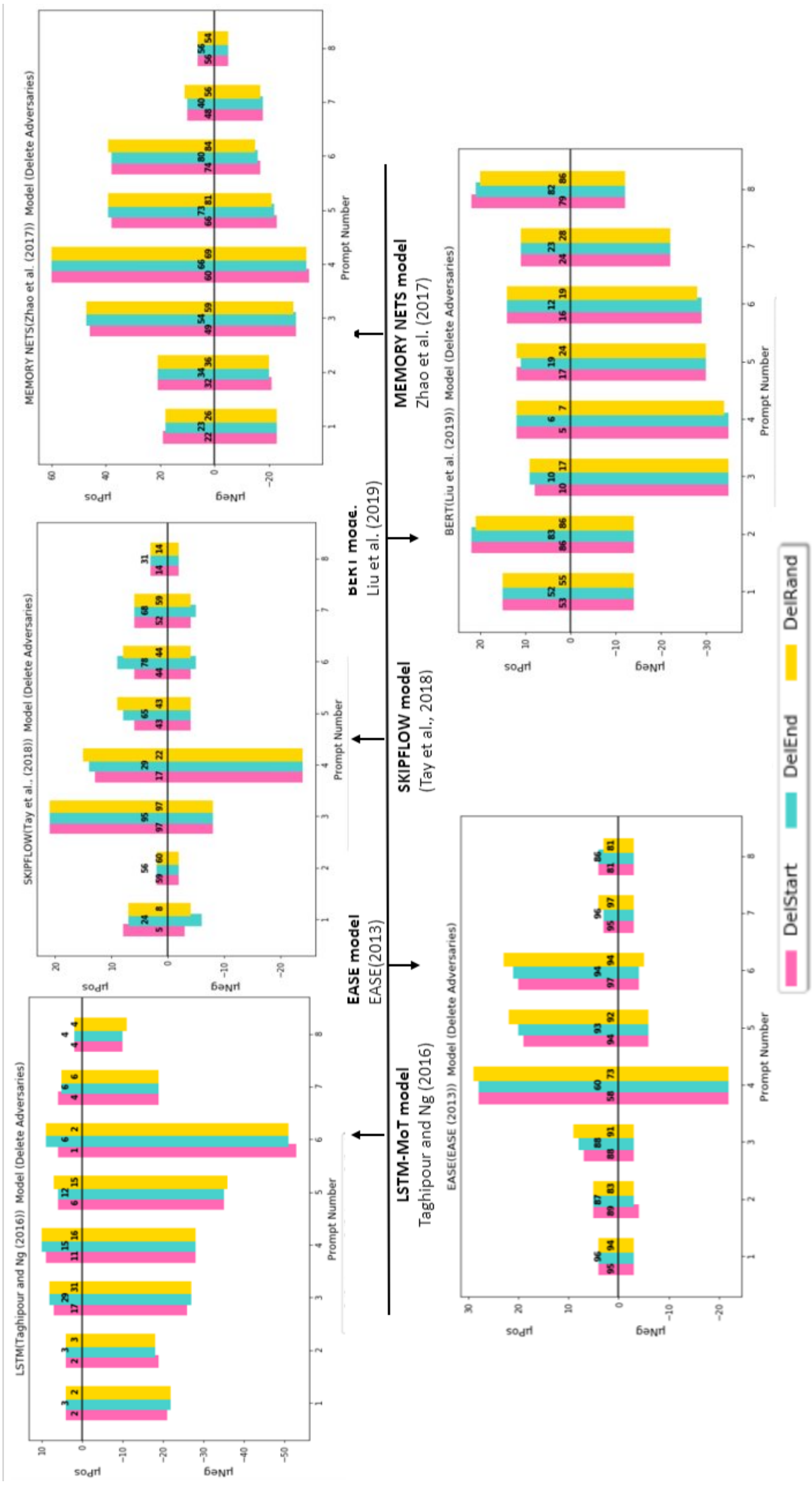

Fig. 7 Results for DELETE Adversaries (refer Section-3.2.4 across all the models. Length of the bar above x-axis denote μ_{pos} and length of bar below x-axis denote μ_{neg}. Numbers on μ_{pos} bar denote the N_{pos} metric. Viewing figure 90 degrees clockwise, *Top Left*: Taghipour and Ng (2016), *Bottom Left*: EASE (2013), *Top Middle*: Tay et al. (2018), *Bottom Right*: Liu et al. (2019), *Top Right*: Zhao et al. (2017).

when scored by all the models. Prompt 8 shows highest signs of overstability for all models and ADD based tests. The variation of μ_{pos} is 3%, and μ_{neg} is 5% only. On the other hand, Prompt 4 shows the least signs of overstabity

Analysis across tests: Among all MODIFY Adversaries (refer Section 3.2.2), we observe that test MODGRAMMAR had a consistently low score amongst all the models. This can be verified as the measure N_{pos} is significantly lower in all the models except (EASE, 2013). Overall, N_{pos} constitutes of only 36% of all the adversarial responses over all prompts. This shows that most models can identify grammatically incorrect sentences and score them lower, respectively. The intensity of scoring these adversarial responses negatively is also higher than MODLEXICON and MODSHUFFLE. This is seen by the higher values of μ_{neg} for MODGRAMMAR compared to MODLEXICON and MODSHUFFLE testcases, over most of the prompts. A difference of around 4% can be seen between the μ_{neg} of MODLEXICON and MODGRAMMAR. However, for model (EASE, 2013) the trend is opposite with respect to N_{pos}. An average of 83% of incorrect grammar adversarial responses are scored positively, in this case. This shows that model (EASE, 2013) is not able to recognize grammatical errors in the responses. Moreover, it is scoring these adversarial responses higher than the original.

4.2.4 GENERATE *Adversaries*

Another category of test case BABELGEN where we generate incoherent and meaningless responses. Ideally, this should have been scored a zero but as demonstrated in Table 19, we notice that almost all the models score these generated essays at least 60% of the prompt scoring range. This strongly suggests that models were looking for obscure keywords with complex sentence formation.

M/P	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Range	2-12	1-6	0-3	0-3	0-4	0-4	0-30	0-60
1	7.1	2.5	1.7	1.1	2.2	1.2	13.8	33.9
2	10	4.4	2	2	3	1.2	19.1	43.1
3	6	2	1.1	0.9	1.3	1.3	12.1	21.9
4	8.4	4	3	3	4	3.9	18.4	40.1
5	10.8	5.6	2.8	2.9	3.8	3.8	26.2	53

Table 19 Scores for BABELGEN over all the prompts and models. Ideally, all of the Babel generated essays should have been scored a zero. Legend: M: Model (y-axis), P: Prompt (x-axis), Model Types: 1: LSTM-MoT (Taghipour and Ng, 2016), 2: EASE (EASE, 2013), 3: SkipFlow (Tay et al., 2018), 4: Memory Networks (Zhao et al., 2017), 5: Adversarial Evaluation + BERT (Liu et al., 2019).

Newpage

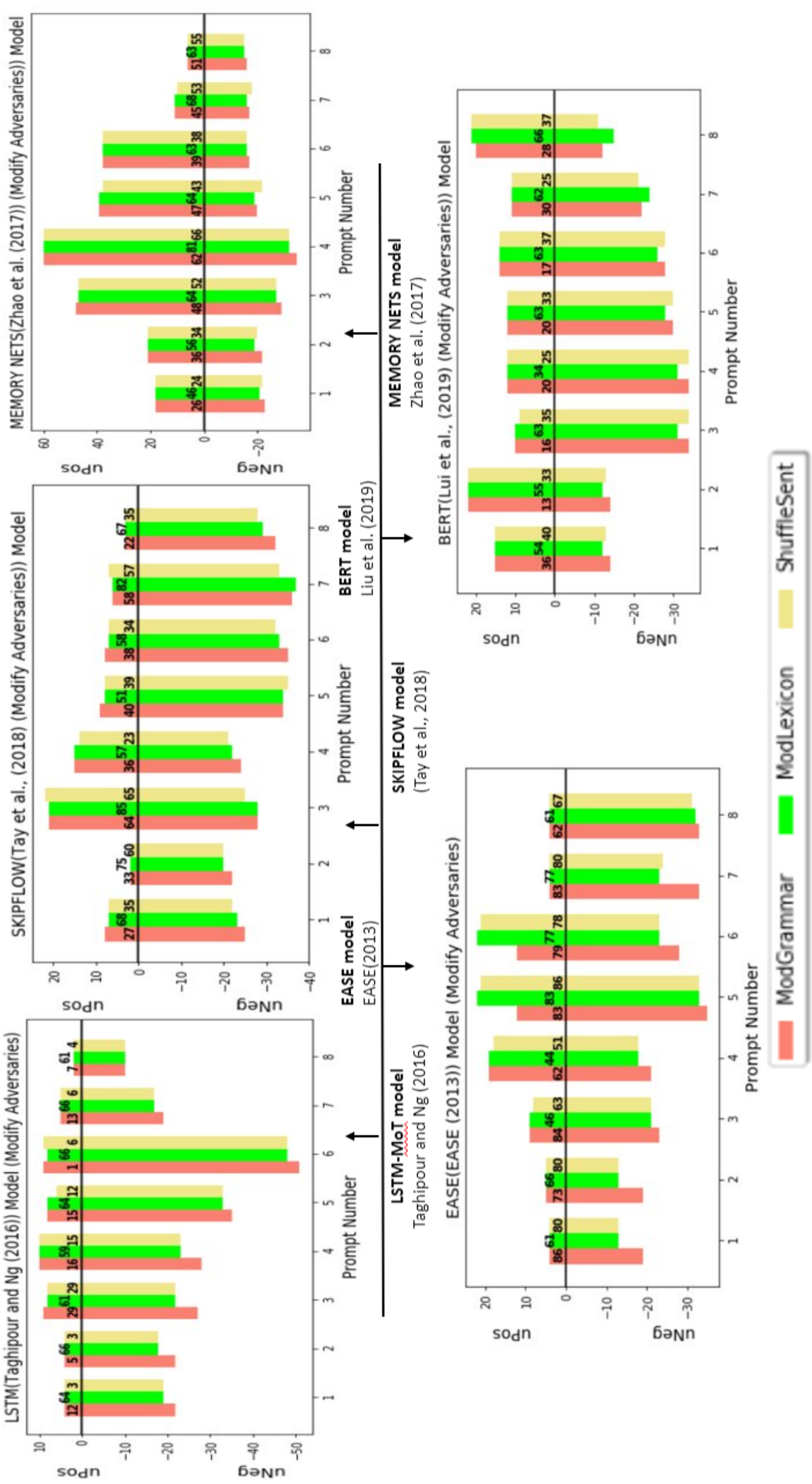

Fig. 8 Results over Modify Adversaries (refer Section-3.2.2 across all the models. Length of the bar above x-axis denote μ_{pos} and length of bar below x-axis denote μ_{neg}. Numbers on μ_{pos} bar denote N_{pos} metric. Viewing figure 90 degrees clockwise, *Top Left*: Taghipour and Ng (2016), *Bottom Left*: EASE (2013), *Top Middle*: Tay et al. (2018), *Bottom Right*: Liu et al. (2019), *Top Right*: Zhao et al. (2017).

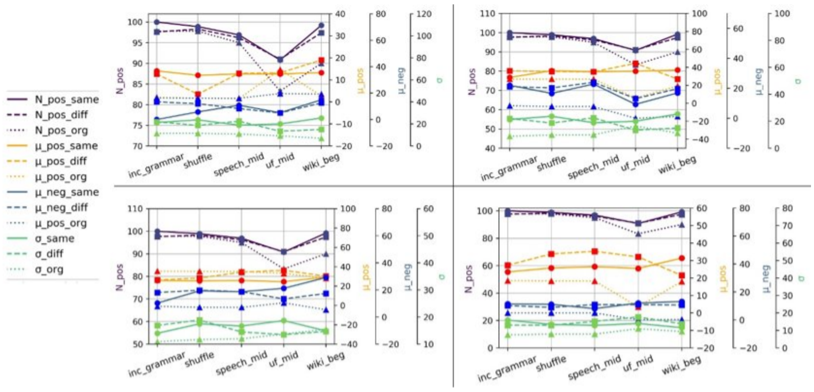

Fig. 9 Results of adversarial training for Prompts 2,3,5,7 in clockwise order. The x-axis shows chosen test-cases and y-axis shows 4 metrics: $\{\mu_{pos}, \mu_{neg}, N_{pos}, \sigma\}$. Representations: The solid lines denoted by $metric_{same}$: Value of $metric$ with adversarial training done over the data generated by the same test case, the dashed lines denoted by $metric_{diff}$: Value of $metric$ with adversarial training done over the data generated by a different test case, the dotted lines denoted by $metric_{original}$: Value of $metric$ with no adversarial training done

4.3 Human Annotation Results

#	Perturbation	Score ↓ %	% People ↓	% People ↑	Common Reasons of ↓	Common Reasons of ↑
1	Shuffle	24.2	68.6	14.5	Transitions ,Organization, Relevance	None
2	ModGrammar	39.5	91.3	6.2	Grammar, Conventions, Readability	None
3	AddWikiRelated	38.2	87.2	11.3	Readability, Relevance, Conventions	Transitions
4	RepeatSent	15.6	71.6	13.6	Organization, Relevance, Repetition	Clarity
5	AddLies	23.9	79.9	10.6	Relevance, Organization	Conventions
6	AddTruth	29.2	88.6	8.6	Relevance, Readability	Organization
7	AddSong	32.8	91.8	3.2	Relevance, Organization, Grammar	Both equal
8	DelRand	38.2	87.2	11.3	Transitions, Organization	Same, More appropriate

Table 20 Human Annotation Results. (↓ represents a decrease and ↑ represents an increase. Therefore, '% People ↓' denotes the percentage of people who scored the adversarial response worse than the original response)

In order to validate that most of our tests are such that they are perceived as negatively impacting scoring, we chose a few test cases based on the following three conditions: 1) Where $N < N_{neg}$, 2) Where $\mu_{pos} > 10\%$ and 3) Where a T-test rejects the hypothesis that the adversarial and original scores are the same distribution. The motivation behind setting these three conditions was that we wanted to choose those test-cases where the model is the most confident in scoring adversarial response as negative and unfavorable. Through this, we can show that even while being confident, they still lack in penalizing scores *adequately*. In all other test-cases, models are either marking the perturbations as better than the original ($N > N_{neg}$) or not detecting any significant difference (second and third conditions), both of which are wrong presumptions by the model. Table 20 depicts the results of human annotations. We divide the annotators into two groups. We show them the original response and its corresponding score for the first group and then ask the annotators to score the adversarial response accordingly. For the second group, we ask them to score both the original and adversarial responses. If any of the annotators felt that both the responses' scores should not be the same, we ask them to list supporting reasons. For uniformity in responses, we derive a set of scoring rubrics extracted mentioned in our dataset and ask them to choose the most suitable ones. As observed from Table 20, the percentage of people who scored adversarial responses **lower** than original responses is significantly higher for all selected test-cases. The main reasons for scoring adversarial responses lower by annotators are *Relevance, Organization, Readability, etc.* It can be observed that the % lowering in score was on an average of 30%.

4.4 Adversarial Training

Finally, we performed an experiment by training on the adversarial samples generated by our framework to see if the models can pick up some inherent "pattern" of the adversarial samples. Since there is a multitude of adversarial test cases category, we narrowed a subcategory of five test cases from those shown for the human annotations. They were selected such that on an average, these test cases had maximum deviation between human annotated scores and machine scores. The train data consisted of an equal number of original samples and adversarial samples. The target scores of adversarial samples were set as the original score minus the mean difference of scores between original and human-annotated scores. For example, according to the human annotation study, for the ModGrammar case, the mean difference was 2 points below the original score, so all the samples were scored as original scores minus 2 points in the simulated training data. The simulated training data was then appended with original and shuffled. The testing was conducted with the respective adversarial test-case as well as the others. The results for the same is shown in Figure 9. It is evident that the adversarial training improves the scores marginally for all four metrics, as shown by the solid lines being higher than the dotted lines. However, a slightly visible improvement in scores is inapparent. The N_{pos} increases for adversarial training, highest for the respective test-case. Similar trend is observed for μ_{neg} metric. For μ_{pos}, the adversarial training reduces this score for respective testcase, as compared to non-adversarial testing.

5 Conclusion

Through our experiments. we conclude that current AES systems built mainly with feature extraction techniques and deep neural networks based algorithms fail to recognize the presence of common-sense adversaries in student essays and responses. As these common adversaries are popular among students for '*bluffing*' during examinations, it is vital for Automated Scoring system developers to think beyond the accuracies of their systems and pay attention to complete robustness so that these systems are not vulnerable to any form of adversarial attack.

6 Declaration

- Funding: Not Applicable
- Conflicts of interest/Competing interests: Not Applicable
- Availability of data and material:

  Can be found here: https://bit.ly/2GYEHyB
- Code availability:

  Can be found here: https://bit.ly/3j6AFRm

  Newpage

References

Ajay HB, Tillet P, Page EB (1973) Analysis of essays by computer (aec-ii). US Department of Health, Education, and Welfare, Office of Education, National Center for Educational Research and Development, Washington, DC, Tech Rep 10:1–13

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

Attali Y, Burstein J (2004) Automated essay scoring with e-rater® v. 2.0. ETS Research Report Series 2004(2):i–21

Bansal H (2020) Song lyrics data. `https://www.kaggle.com/hiesenberg/song-lyrics-data`

Barzilay R, Lapata M (2008) Modeling local coherence: An entity-based approach. Computational Linguistics 34(1):1–34

Bejar II, Mislevy RJ, Zhang M (2017) Automated scoring with validity in mind. The wiley handbook of cognition and assessment: Frameworks, Methodologies, and applications pp 226–246

Burstein J, Chodorow M, Leacock C (2004) Automated essay evaluation: The criterion online writing service. Ai magazine 25(3):27–27

Chen L, Tao J, Ghaffarzadegan S, Qian Y (2018) End-to-end neural network based automated speech scoring. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp 6234–6238

Cities C (2013) Safe and sound, in a tidal wave of mystery album. `https://github.com/edx/ease`

Cushing Weigle S (2010) Validation of automated scores of toefl ibt tasks against non-test indicators of writing ability. Language Testing 27(3):335–353

Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805

EASE E (2013) 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

FiveThirtyEight (2019) Fivethirtyeight hip hop candidate lyrics dataset. https://bit.ly/3Ol4kPK

Foltz PW, Streeter LA, Lochbaum KE, Landauer TK (2013) Implementation and applications of the intelligent essay assessor. Handbook of automated essay evaluation pp 68–88

Higgins D, Heilman M (2014) Managing what we can measure: Quantifying the susceptibility of automated scoring systems to gaming behavior. Educational Measurement: Issues and Practice 33(3):36–46

I, T (2019) Dance monkey, the kids are coming album. https://github.com/edx/ease

Kumar Y, Aggarwal S, Mahata D, Shah RR, Kumaraguru P, Zimmermann R (2019) Get it scored using autosas—an automated system for scoring short answers. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 33, pp 9662–9669

Liu J, Xu Y, Zhu Y (2019) Automated essay scoring based on two-stage learning. arXiv preprint arXiv:190107744

Lochbaum KE, Rosenstein M, Foltz P, Derr MA, et al. (2013) Detection of gaming in automated scoring of essays with the iea. In: National Council on Measurement in Education Conference (NCME), San Francisco, CA

Mid-Day (2017) 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

Miller GA (1995) Wordnet: a lexical database for english. Communications of the ACM 38(11):39–41

Neisse A (2019) Song lyrics from 6 musical genres. https://www.kaggle.com/neisse/scrapped-lyrics-from-6-genres/

Obama B (2004) Barack obama's remarks to the democratic national convention. https://www.nytimes.com/2004/07/27/politics/campaign/barack-obamas-remarks-to-the-democratic-national.html

Obrecht F (1999) Minimum essentials of English. Barron's Educational Series

Perelman L (2014) When "the state of the art" is counting words. Assessing Writing 21:104–111

Perelman L, Sobel L, Beckman M, Jiang D (2014a) Basic automatic b.s. essay language generator (babel). https://babel-generator.herokuapp.com/

Perelman L, Sobel L, Beckman M, Jiang D (2014b) Basic automatic b.s. essay language generator (babel) by les perelman, ph.d. http://lesperelman.

com/writing-assessment-robo-grading/babel-generator/
Powers DE, Burstein JC, Chodorow M, Fowles ME, Kukich K (2001) Stumping e-rater: Challenging the validity of automated essay scoring. ETS Research Report Series 2001(1):i–44
PromptCloud (2018) Taylor swift song lyrics from all the albums. https://bit.ly/39a4GN0
RakanNimer (2017) Billboard 1964-2015 songs + lyrics. https://www.kaggle.com/rakannimer/billboard-lyrics/discussion/142574
Reinertsen N, et al. (2018) Why can't it mark this one?: A qualitative analysis of student writing rejected by an automated essay scoring system. English in Australia 53(1):52
Schultz MT (2013) The intellimetric automated essay scoring engine-a review and an application to chinese essay scoring. New York: Routledge
Shermis MD, Hamner B (2012) Contrasting state-of-the-art automated scoring of essays: Analysis. In: Annual national council on measurement in education meeting, pp 14–16
Spolsky B (1995) Measured words: The development of objetive language testing. Oxford University Press
Strauss V (2014) Is mit researcher being censored by educational testing service?. https://wapo.st/3jiIipt
Sukkarieh JZ, Blackmore J (2009) C-rater: Automatic content scoring for short constructed responses. In: Twenty-Second International FLAIRS Conference
Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2013) Intriguing properties of neural networks. arXiv preprint arXiv:13126199
Taghipour K, Ng HT (2016) A neural approach to automated essay scoring. In: Proceedings of the 2016 conference on empirical methods in natural language processing, pp 1882–1891
Tay Y, Phan MC, Tuan LA, Hui SC (2018) Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring. In: Thirty-Second AAAI Conference on Artificial Intelligence
Trump DJ (2017) Remarks of president donald j. trump – as prepare for delivery inaugural address. https://www.whitehouse.gov/briefings-statements/the-inaugural-address/
Wallace E, Feng S, Kandpal N, Gardner M, Singh S (2019) Universal adversarial triggers for attacking and analyzing nlp. arXiv preprint arXiv:190807125
Ward L (2020) 1000 random & interesting facts about literally everything. https://www.thefactsite.com/1000-interesting-facts/
West-Smith P, Butler S, Mayfield E (2018) Trustworthy automated essay scoring without explicit construct validity. In: 2018 AAAI Spring Symposium Series
Whitlock JW (1964) Automatic data processing in education. Macmillan
Xu P, Saghir H, Kang JS, Long T, Bose AJ, Cao Y, Cheung JCK (2019) A cross-domain transferable neural coherence model. arXiv preprint arXiv:190511912

Yan D, Rupp AA, Foltz PW (2020) Handbook of automated scoring: Theory into practice. CRC Press

Yoon SY, Cahill A, Loukina A, Zechner K, Riordan B, Madnani N (2018) 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), pp 60–67

Zhang WE, Sheng QZ, Alhazmi A, Li C (2020) Adversarial attacks on deep-learning models in natural language processing: A survey. ACM Transactions on Intelligent Systems and Technology (TIST) 11(3):1–41

Zhao S, Zhang Y, Xiong X, Botelho A, Heffernan N (2017) A memory-augmented neural model for automated grading. In: Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale, pp 189–192