A Multilingual Dataset of Adversarial Attacks to Automatic Content Scoring Systems

Ronja Laarmann-Quante, Christopher Chandler, Noemi Incirkus,
Vitaliia Ruban, Alona Solopov and Luca Steen
Ruhr University Bochum, Faculty of Philology, Department of Linguistics, Germany
`noemi.incirkus@edu.rub.de`
`{ronja.laarmann-quante, christopher.chandler, vitaliia.ruban, alona.solopov, luca.steen}@rub.de`

Abstract

Automatic content scoring systems have been shown to be vulnerable to adversarial attacks, i.e. to answers that human raters would clearly recognize as incorrect or even nonsense but that are nevertheless rated as correct by an automatic system. The existing literature on this topic has so far focused on English datasets. In this paper, we present a multilingual dataset of adversarial answers for English, German, French and Spanish based on the multilingual ASAP content scoring dataset introduced by Horbach et al. (2023). We apply different methods of generating adversarial answers proposed in the literature, e.g. sampling n-grams from existing answers or generic corpora or inserting adjectives and adverbs into incorrect answers. In a baseline experiment, we show that the rate at which adversarials are rejected by a model depends on the adversarial method used, interacting with the language and the prompt-specific dataset a model was trained on.

1 Introduction

One of the prerequisites for automatic scoring tools to be usable in educational settings, besides an overall good performance, is the robustness against cheating behavior. In this paper, we deal with automatic content scoring, also known as automatic short answer grading (ASAG), which refers to the task of scoring students' answers to prompts like the following: `[An experiment about the stretchability of different polymer plastics is outlined]` *Task: Describe two ways the experimenter could have improved the experimental design and/or validity of the results.*[1] The answers to such prompts are typically short, ranging from a few words to a few sentences and the focus of the scoring is on content rather than form or style.

Previous work has shown that automatic scoring models for such tasks can be tricked by different kinds of adversarial answers, meaning answers that are clearly wrong or even nonsense for human raters but that are nevertheless graded as (partly) correct by automatic scoring models. For example, Ding et al. (2020) showed that shallow and deep learning models can be fooled by randomly sampled n-grams taken from real answers, the prompt or even from generic corpora. Willms and Pado (2022) found that increasing the answer length by repeating the answer once or twice can deceive a transformer model into scoring incorrect answers as correct. Filighera et al. (2023) inserted random adjectives and adverbs into wrong answers, which did not turn the answer into a correct answer but it nevertheless increased the likelihood that it would be scored as correct by a transformer model.

The experiments in the literature have so far focused on English datasets. However, different languages pose different challenges to automatic content scoring (see e.g. Padó et al., 2023 for German) that may also influence the vulnerability towards adversarial attacks. Furthermore, automatic content scoring has been tackled from a cross-lingual perspective (Horbach et al., 2023) but so far, there is no multilingual dataset of adversarial answers available that could be used to test the robustness of a model for different language settings.

The aim of this paper is twofold: Firstly, we present a comprehensive multilingual dataset of adversarial answers that comprises English, German, French and Spanish. The adversarial answers are based on the multilingual ASAP dataset introduced in Horbach et al. (2023) using the adversarial methods proposed by Ding et al. (2020) and Filighera et al. (2023) with some extensions (Sec. 3). Secondly, we provide a baseline experiment with a shallow baseline model as used by Ding et al. (2020), showing that not only the language but also the prompt-specific

[1] See prompt 2 of the ASAP-SAS dataset: `https://www.kaggle.com/c/asap-sas/`.

Dataset	Lang.	Prompt		
		1	2	10
ASAP$_{orig}$	English	2,229	1,704	2,186
ASAP$_{orig300}$	English	300	300	300
ASAP$_{en}$	English	330	328	330
ASAP$_{de}$	German	301	301	301
ASAP$_{fr}$	French	274	187	211
ASAP$_{es}$	Spanish	325	297	393

Table 1: Number of answers per dataset and prompt in the multilingual ASAP corpus of Horbach et al. (2023).

dataset that a model was trained on and the specific adversarial method has a large influence on a model's capability of rejecting adversarial answers (Sec. 4). The code and data from this study is available under the following link:
https://gitlab.ruhr-uni-bochum.de/vamos-cl/multilingual-adversarial-dataset-konvens-2024

2 Data

First, we present the content scoring dataset and second the generic corpora used as background corpora for each language, e.g. for constructing prompt-independent adversarial answers.

2.1 Content Scoring Data

We use the English, German, French and Spanish part of the multilingual content scoring dataset introduced by Horbach et al. (2023). The English part consists of three datasets: ASAP$_{orig}$, which comprises answers to prompts 1, 2 and 10 of the original ASAP-SAS dataset (see footnote 1) collected from high school students; ASAP$_{orig300}$, which contains a random sample of ASAP$_{orig}$ with 300 answers per prompt so that it roughly matches the datasets in the other languages in size; ASAP$_{en}$, comprising answers to the same prompts collected from crowd workers, matching the data collection process for the other languages, i.e. German (ASAP$_{de}$), French (ASAP$_{fr}$) and Spanish (ASAP$_{es}$). Table 1 shows the number of answers per dataset and prompt. All answers come with an adjudicated gold score produced by human raters. Prompts 1 and 2 were scored on a scale from 0 (incorrect answer) to 3 (perfect answer), prompt 10 on a scale from 0 to 2. More information about the dataset can be found in Horbach et al. (2023).

2.2 Generic Corpora

For each language, we use a generic corpus as background corpus. Following Ding et al. (2020) and Filighera et al. (2023), we use the Brown Corpus (Kučera and Francis, 1967) for English available via the NLTK library (Bird et al., 2009). For German and French, we use the newest available corpora from the Leipzig Corpora Collection[2] (Goldhahn et al., 2012) that were compiled from randomly chosen websites, which are "deu-com web" from 2021 for German and "fra-ch web" from 2020 for French. For Spanish, we use the CESS-ESP corpus (Martí et al., 2007) available via the NLTK. Basic statistics for the generic corpora are summarized in Table 4 in Appendix A.

3 Adversarial Methods

We use three different types of methods for generating adversarial answers: (1) word-based and character-based n-grams from either real answers or generic corpora following Ding et al. (2020). These methods assume knowledge of n-gram probabilities in either real answers or in generic texts of a language. This is not what a student who wants to cheat is assumed to know but nevertheless a trustworthy system has to be robust against such (nonsense) answers (Ding et al., 2020). (2) Sampling either n-grams or only nouns from the prompt material. These methods also create nonsense answers but they could mimic real cheating of a student who just copies material from a prompt. (3) Inserting either adjectives or adverbs into wrong answers as proposed by Filighera et al. (2023). They found that such answers looked more unnatural to human raters but not like suspicious cheating attacks. Nevertheless, some of the answers fooled a neural model into scoring incorrect answers as correct.

For each method in each set, we generate 1,000 adversarial answers for each prompt. Table 7 in Appendix A shows an example adversarial for each method. Where part-of-speech (POS) tags are needed, texts are first tagged with spaCy (Honnibal et al., 2020), using the small core model for each language. For the Brown Corpus, the POS tags that come with the corpus are used.

3.1 Random N-Grams

In this method, we create adversarial answers by a weighted random sampling of 1-5 grams based on either words or characters from either the real answers to a prompt (correct as well as incorrect answers, henceforth called ASAP-based adversarials)

[2] https://wortschatz.uni-leipzig.de/en/download/

or the generic corpus. To make the generic corpora comparable in size, we use a randomly sampled subset of 5,000 sentences for each corpus.

For word-based n-grams, we follow the procedure described in Ding et al. (2020) with a few changes to make the adversarials more similar to the real answers: Firstly, we keep punctuation marks and secondly, we determine the lengths of the answers differently: In Ding et al. (2020), an answer ends when the last n-gram contains the special end-of-sentence token or when a pre-defined maximum length is reached. We also use the end-of-sentence marker to stop the generation process for an answer but besides that, we use a random length for each answer that lies in the range of plus/minus one standard deviation around the mean number of tokens in the real answers to the prompt. We do the same for character-based n-grams, where additionally, we take the mean token length (plus/minus one standard deviation) into account to generate word boundaries. In addition, we add spaces before capital letters and after punctuation marks.

3.2 Random Prompt Material

The following two adversarial methods could resemble real cheating behavior of students, namely randomly picking and rearranging either n-grams or only nouns from the given prompt. Table 5 in Appendix A shows the number of words and nouns, respectively, in the prompt material as used in the data collection for each language.

Prompt N-Grams Firstly, we generate adversarials by randomly sampling 1-5 grams from the prompt material. We sample with replacement and generate separate adversarials for each *n*. Of course, in real cheating, it would be odd to assume that a student would always pick exactly *n* adjacent words but this allows us to systematically study the role of greater context. We keep words occurring in graphics or tables in the prompt material but we remove punctuation marks and also do not mark the beginning or end of a sentence. Each adversarial answer has a random length between minus/plus one standard deviation around the mean number of words in the real answers to a prompt, with a minimum length of 5 words. Our rationale is that students would roughly know from experience how long answers are expected to be.

Prompt Nouns In this method, we create answers only consisting of nouns from the prompt, which is equivalent to the ‘Content Burst’ method in Ding et al. (2020) applied only to the prompt and not the student answers. The idea is that nouns carry most of the semantic value of an answer. We first extract all the common nouns (including nouns occurring in tables and graphics) and then randomly sample nouns (with replacement) based on their token frequency up to an average maximum length of 44 characters (following Ding et al., 2020), resulting in answer lengths of 6-7 words.

Dataset	Prompt		
	1	2	10
ASAP$_{orig}$	380 (23%)	168 (13%)	290 (18%)
ASAP$_{orig300}$	68 (23%)	43 (14%)	51 (17%)
ASAP$_{en}$	178 (59%)	176 (59%)	115 (38%)
ASAP$_{de}$	151 (50%)	97 (32%)	121 (40%)
ASAP$_{fr}$	125 (46%)	76 (41%)	70 (33%)
ASAP$_{es}$	82 (25%)	84 (28%)	74 (19%)

Table 2: Number of answers scored 0 by human raters.

3.3 Inserting Adjectives and Adverbs

For this set of adversarials, we use the method of inserting adjectives or adverbs into incorrect answers as proposed by Filighera et al. (2023). To this end, we first extracted all answers scored with zero points by the human raters, see Table 2.

Inserting Adjectives Following Filighera et al. (2023), we filtered the 100 most frequent adjectives occurring with nouns and pronouns in the generic corpus of a language. To do so, for the Germanic languages German and English, where adjectives precede nouns, we extracted all bigrams consisting of a word form tagged as adjective as first element and a noun, pronoun or proper noun as second element, e.g. `((‘general’, ‘ADJ’), (‘purposes’, ‘NOUN’)), ((‘occasional’, ‘ADJ’), (‘meetings’, ‘NOUN’))`. For the Romance languages French and Spanish, where adjectives follow nouns, we looked for the respective bigrams with adjectives as the second element. We then identified the 100 adjectives occurring most frequently in this set of bigrams. To create an adversarial answer, we insert a random adjective from this list before every noun (for English and German) and after every noun (for French and Spanish), respectively, in each incorrect answer. In order to get 1,000 adversarials for every prompt in every language, we create different versions of each incorrect answer by randomly choosing different adjectives.

It is important to note that we do not adjust the inflection of the adjectives to the grammatical context. Adjectives have to agree in grammatical gender and number with nouns in German, Spanish and French (and additionally in case in German), which is not relevant for English. Therefore, the generated answers in these languages may be perceived as more unnatural to human raters. However, since non-native speakers are likely to produce the same kinds of grammatical errors, it should not affect their ratings. Likewise, we do not check semantic appropriateness which leads to expressions like *the experimental* ***christian*** *design* or *the* ***dark*** *experiment* that could in fact look suspicious to human raters.

Inserting Adverbs For inserting adverbs into wrong answers, we again largely follow Filighera et al. (2023). Working only with English, they first identified bigrams in which adverbs preceded verbs based on the Brown Corpus, and extracted the 100 most frequent adverbs from this set. The adversarials were then created by choosing a random adverb from this list and inserting it before the verb in every sentence of a wrong answer.

For English, we adopt this procedure but for the other languages, we first empirically determined common positions for adverbs as they could differ from English. From the German, French and Spanish generic corpora, we extracted the five most frequent trigrams containing adverbs in the middle position, e.g. (`NOUN, ADV, VERB`).[3] The result is shown in Table 6 in Appendix A. For each language, we determined the 100 most frequent adverbs occurring in these positions. Next, we transformed the extracted POS-trigrams into bigrams by removing the ADV tag. To create the adversarial sentences, we iterate over the POS tags of the answers and, for the first POS-bigram from this list of bigrams that we encounter, add a random adverb from the pool between the two words. After a manual review of the thusly created adversarials, we added an additional rule for German wherein we place adverbs after auxiliary verbs to create more natural-sounding sentences. Note, however, that as with adjectives, we did not check the adversarials for grammatical or semantic correctness, yielding also answers that human raters might find unnatural. In all languages, answers that do not contain verbs or any of the aforementioned POS-bigrams are modified by inserting an adverb at the beginning of the sentence. To get 1,000 adversarial answers per language and prompt, we create different versions of each incorrect answer by inserting a different random adverb.

Dataset	**Lang.**	**Prompt**		
		1	2	10
ASAP$_{orig}$	English	.73	.49	.65
ASAP$_{orig300}$	English	.56	.35	.56
ASAP$_{en}$	English	.52	.15	.56
ASAP$_{de}$	German	.54	.49	.55
ASAP$_{fr}$	French	.68	.67	.59
ASAP$_{es}$	Spanish	.72	.46	.63

Table 3: Performance of the models based on ten-fold cross-validation on real answers measured in QWK.

4 Scoring Adversarial Answers

We provide a baseline experiment concerning the ability of a baseline scoring model to reject the different kinds of adversarial answers. Ding et al. (2020) used an SVM-based shallow model that was shown to be more robust against adversarials than a neural model, therefore we decided to use a shallow scoring model with a similar setup. Note that the goal of this paper is not to find the best model but rather to gain some insights into the behavior of different kinds of adversarials for different prompts and languages. We train a separate model for each prompt in each dataset. Following Ding et al. (2020), we use an SVM classifier with default kernel and the following features: the top 10,000 character 2-5 grams, the top 10,000 word 1-5 grams, and answer length. Our model is implemented with *scikit-learn* (Pedregosa et al., 2011).

To measure the performance of each model when scoring real answers, we calculate quadratically-weighted kappa (QWK) based on 10-fold cross-validation. QWK is typically used for content scoring as it takes the distance between the gold score and the predicted score into account. The results are given in Table 3, showing some variance between the languages but also between the prompts.

Like Ding et al. (2020), we measure the robustness of a model against adversarial answers with the adversarial rejection rate (ARR): A perfect model should reject every adversarial answer, i.e. assigning a score of 0. This would yield an ARR of 1.0. Every adversarial scored 1 or higher is regarded as not-rejected, i.e. accepted. A model that accepts all adversarials would have an ARR of 0.0, i.e. the higher the score, the better.

Note that for English, we always train three dif-

[3] Tags are taken from the simple UPOS tagset (`https://universaldependencies.org/u/pos/)`.

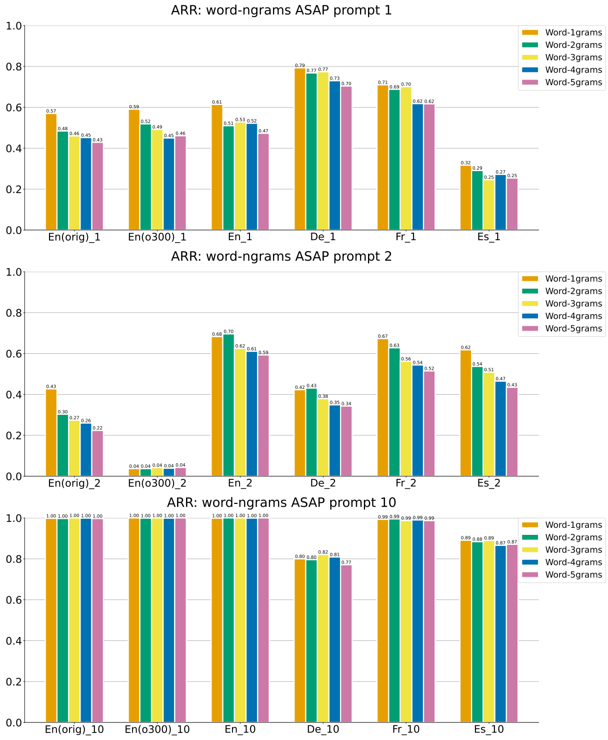

Figure 1: ARRs for the adversarials based on word n-grams from the ASAP corpora.

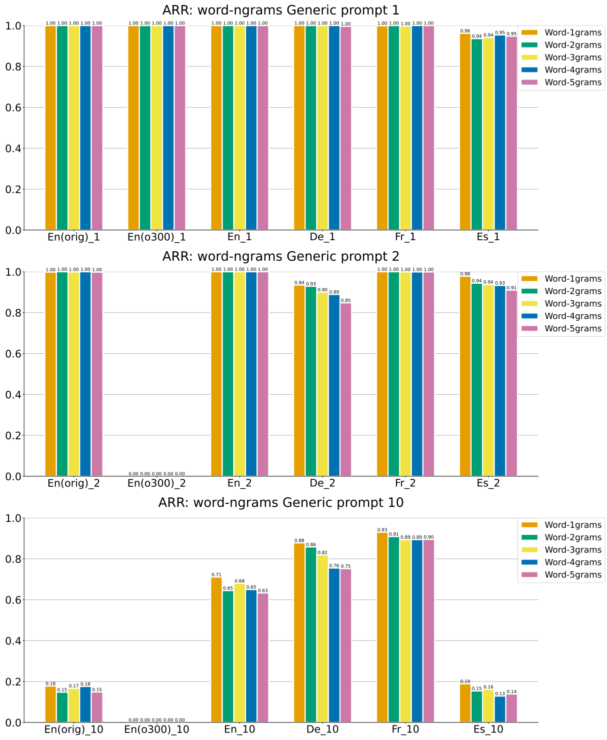

Figure 2: ARRs for the adversarials based on word n-grams from the generic corpora.

ferent models, based on ASAP$_{orig}$, ASAP$_{orig300}$ and ASAP$_{en}$, respectively. This means that for English adversarials that are based on a generic corpus or the prompt material rather than a specific dataset, each model is given the same adversarials and the difference in ARR can be attributed to a difference training material rather than the adversarials. An overview of all results is given in Table 8 in Appendix A.

4.1 Results for Random N-Grams

4.1.1 Word-Based N-Grams

First, we summarize the results for the word-based n-grams shown in Figure 1 (ASAP-based) and Figure 2 (generic). Regarding the **size of *n***, across all prompts and languages, the ARR tends to be highest for the adversarials generated with unigrams and lowest for those generated with 5-grams, which is in line with the results of Ding et al. (2020). The ARR of adversarials generated from the **generic** corpora tends to be higher than that of the **ASAP-based** ones, which is also in line with Ding et al. (2020). Only for prompt 10, we see a different pattern with ASAP-based adversarials being more consistently rejected than generic ones across all languages. Regarding **language**, it is notable that the ARRs of the French adversarials are mostly in the upper range compared to the other languages, especially for the generic adversarials. In contrast, the Spanish adversarials tend to have the lowest ARRs compared to the other languages.

The greatest variance can be seen among the different **prompts**, partly interacting with the language: While for prompt 1, the ARR for the generic adversarials is close to 1.0 for each language and each n, the other prompts behave differently. In prompt 10, ASAP$_{orig}$, ASAP$_{orig300}$, ASAP$_{en}$ and ASAP$_{es}$, have strikingly low ARRs. Especially ASAP$_{orig300}$ sticks out, with all generic adversarial answers being accepted in prompt 10 and all generic as well as most ASAP-based adversarials in prompt 2 (also for character-based n-grams).

To investigate this further, we performed different checks: We first used the adversarials created from ASAP$_{orig300}$ prompt 2 with a scoring model trained on one of the other English datasets, i.e. ASAP$_{orig}$ and ASAP$_{en}$. The system trained on ASAP$_{en}$ yielded an ARR of almost 1.0 for each n. The ARRs for the ASAP$_{orig}$ model were similar to the ones generated from ASAP$_{orig}$, which is expected since ASAP$_{orig300}$ is a subset of ASAP$_{orig}$. From this, we conclude that there is nothing odd with the ASAP$_{orig300}$ adversarials but rather that the scoring model trained on ASAP$_{orig300}$ is in-

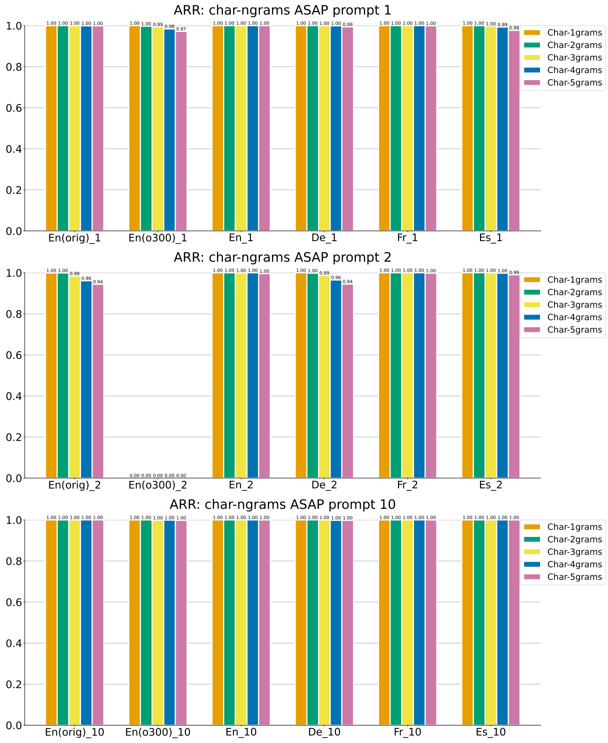

Figure 3: ARRs for the adversarials based on character n-grams from the generic corpora.

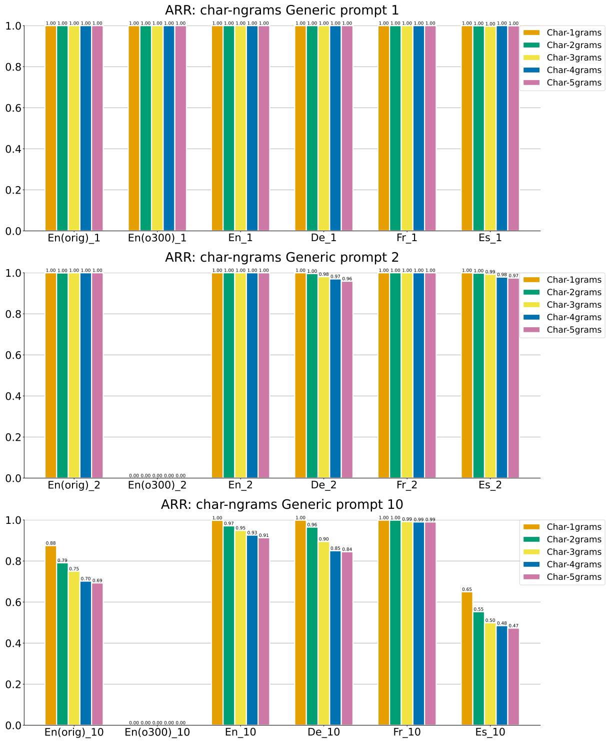

Figure 4: ARRs for the adversarials based on character n-grams from the ASAP corpora.

sufficient. This was confirmed by the following check: We used the German, French and Spanish generic adversarials to calculate the ARR of the model trained on ASAP$_{orig300}$. As the (shallow) model has not seen any of these languages during training, all the answers should clearly be scored 0. While for prompt 1, the ARRs were indeed close to 1.0 as expected, the ARRs for prompt 2 and prompt 10 were all close to 0.0. Hence, the scoring model built from ASAP$_{orig300}$ for these prompts must be insufficient. One possible explanation for this is that the dataset is too skewed, with only 14% and 17% of the answers in ASAP$_{orig300}$ prompt 2 and 10, respectively, having a (gold) score of 0, which may mean that the model built from these prompts failed to learn to detect incorrect answers at all. This is not so much apparent from the aggregated cross-validation performance on real answers (see Table 3) than for the ability to reject adversarial answers and it emphasizes the need to evaluate model performance from different perspectives.

4.1.2 Character-Based N-Grams

For the character-based n-grams the picture is much more homogeneous than for the word-based n-grams with most ARRs being (close to) 1.0 across languages and prompts, see Figure 3 (ASAP-based) and Figure 4 (generic). One notable exception is the ASAP$_{orig300}$ model with an ARR of 0.0 for prompts 2 and 10 already discussed in Section 4.1.1. The other exception are the generic adversarials scored with the models built on prompt 10: Except for French, the ARRs are notably lower than 1.0. The degree differs by language, but the patterns are similar. This suggests that there is something about this prompt that makes the resulting models less robust against (generic) adversarial answers. However, neither the score distribution nor the answer length nor type-token ratio analyzed in Horbach et al. (2023) are strikingly different for this prompt so this would need further investigation in future work.

4.2 Results for Random Prompt Material

4.2.1 Prompt N-Grams

Figure 5 shows the results for the adversarials generated from random n-grams from the prompt material. Regarding the size of n, we do not see the same clear pattern as for the generic or ASAP-based n-grams from Section 4.1.1, where the ARRs decreased with increasing n: For the prompt-based n-grams, this pattern only occurs for prompt 10. For prompt 2, especially for ASAP$_{fr}$, we even observe the opposite, namely that answers based on

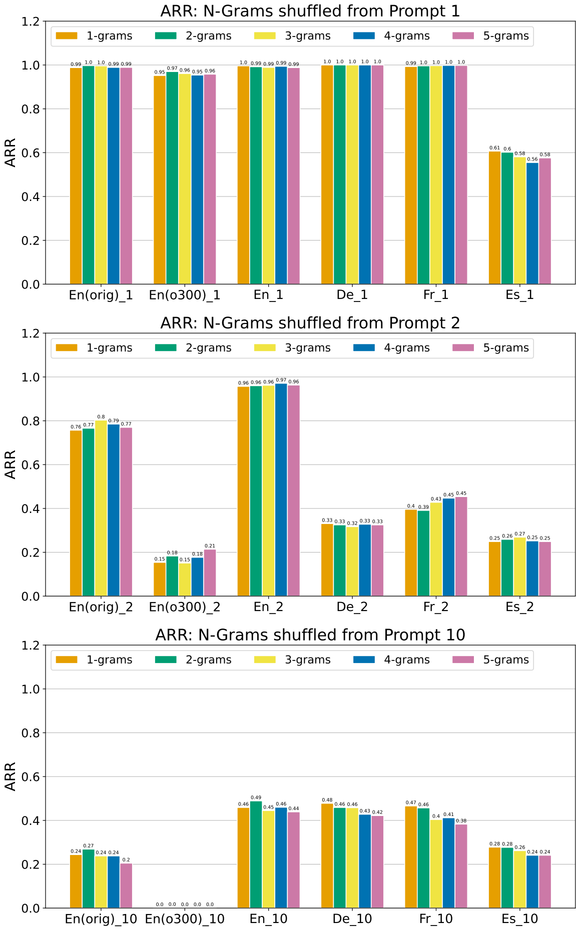

Figure 5: ARRs of adversarials generated from random n-grams taken from the prompt material.

4- or 5-grams are more often rejected than those based on uni- or bigrams. We see that only the ASAP$_{de}$ model for prompt 1 has a perfect ARR of 1.0, i.e. no German adversarial was accepted for any n. In general, we see a clear influence of the prompt, with prompt 1 being the one with the highest ARRs across languages (all $> 90\%$ except ASAP$_{es}$) and prompt 10 the one with the lowest ARRs (no $> 50\%$). For prompt 2, the results differ largely by language. Regarding language, the Spanish models are consistently among the weakest ones. Even in prompt 1, the ARRs are only close to 60%. Here, it is noteworthy that almost all of the accepted Spanish adversarials were even assigned a score of 3, i.e. the highest score. In contrast, the ASAP$_{en}$ models are among the most robust ones across all prompts. We also find again a very weak performance of ASAP$_{orig300}$ for prompts 2 and 10 that was already discussed in Section 4.1.1.

4.2.2 Prompt Nouns

With only random nouns sampled from the prompt rather than n-grams, it was hardly possible to deceive the scoring models. Except for ASAP$_{orig300}$ prompts 2 and 10, where the ARR was again close to 0.0 (see the discussion in Section 4.1.1), the lowest ARRs were 0.985 for ASAP$_{es}$ prompt 10 and 0.988 for ASAP$_{fr}$ prompt 1. But only for German, none of the adversarials was accepted. A total of 13 answers even received a score of 3, i.e. they would have been judged as perfect answers. Recall that the answers generated with this method were considerably shorter than those from the other methods, which might influence the result and would need further investigation.

4.3 Results for Adjectives and Adverbs

4.3.1 Inserting Adjectives

Figure 6 shows the results for adversarials created by the insertion of adjectives into wrong answers. Although our shallow model only uses surface n-grams as features and may not have seen the adjectives during training, these adversarials do indeed fool the model in many cases.

We see that prompt 1 is more robust against these adversarials compared to the other prompts across all datasets. In terms of language, overall, the ASAP$_{fr}$ and ASAP$_{en}$ models are most robust while ASAP$_{orig}$ and ASAP$_{orig300}$ have the lowest ARRs. For ASAP$_{orig}$ this is rather surprising given the large amount of training data and the comparably high QWK values when scoring real answers. The main difference between ASAP$_{en}$ and ASAP$_{orig}$, besides the size, is that ASAP$_{orig}$ was collected from students whereas ASAP$_{en}$ was collected from crowd workers. Potentially, this means that the kind of writing differs. Again, answer length could be an influencing factor, since answers in ASAP$_{orig}$ tend to be longer than answers in ASAP$_{en}$ (Horbach et al., 2023). The fact that the ARRs are lower for ASAP$_{orig}$, where the adjectives fit the grammatical context, than for German, French or Spanish, where adjectives are sometimes wrongly inflected, shows that grammatical correctness is not important for the model. Note also that most of the answers that received a score > 0 were scored with 1 point, but there are also answers that went from originally 0 points to the maximum score.

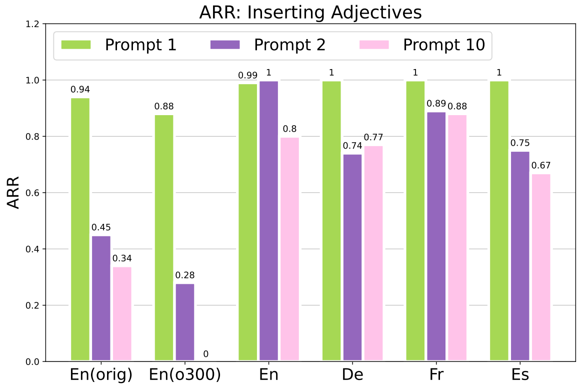

Figure 6: ARRs of adversarials produced by inserting adjectives into wrong answers.

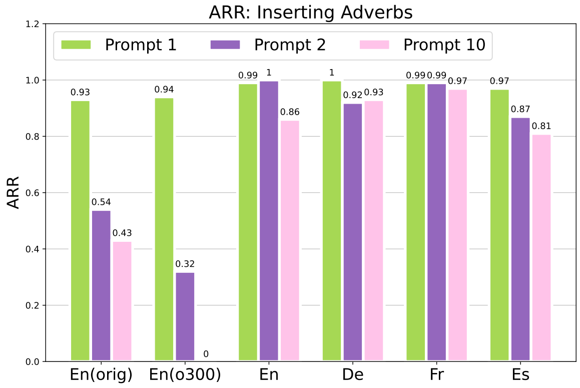

Figure 7: ARRs of adversarials produced by inserting adverbs into wrong answers.

4.3.2 Inserting Adverbs

For the insertion of adverbs, the results show similar patterns as for the insertion of adjectives, see Figure 7: Again, prompt 1 mostly has higher ARRs (all $> 90\%$) than the other prompts for all languages. Furthermore, ASAP$_{orig}$ and ASAP$_{orig300}$ again have the lowest ARRs: While the ARR for all other models stays consistently above 80%, the rejection rate for prompts 2 and 10 in the original student answer corpora ranges only between 0% and 54%. In terms of language differences, the ASAP$_{fr}$ model is again among the most robust models, with ARRs $> 97\%$ for all prompts. However, it is worth noting that for prompts 1 and 2, although the overall rejection rate is close to 1.0, those adversarial answers that were accepted received a score of 3, i.e. the highest score possible. (For prompt 1, all of these adversarials were based on the same student answer but received different adverbs during their creation.) As with adjectives, some adversarial answers would be both syntactically and semantically incorrect but nevertheless be accepted by the system.

Comparing the two insertion methods, adjectives seem to generate answers that more often fool the scoring system than adverbs do. One of the reasons may be that more adjectives are inserted into an answer than adverbs, yielding higher answer lengths. It is possible that the scoring models simply pick up on this (compare the results of Padó et al., 2023). However, while on average, wrong answers are shorter than correct answers in each dataset, there is a high variation within each score (see Horbach et al., 2023). Hence, the interplay of the insertion methods and answer length should be more thoroughly investigated in future work.

5 Conclusion and Future Work

We presented a multilingual dataset of adversarial answers for English, German, French and Spanish based on the multilingual ASAP content scoring dataset introduced by Horbach et al. (2023). In total, 468,000 adversarial answers were generated following different methods proposed in the literature (Ding et al., 2020; Filighera et al., 2023). In a pilot experiment, we tested the rate at which a baseline classifier rejects the adversarial answers.

While the exact results only reflect the specific classifier that we used, some important general conclusions can be drawn: We saw that a classifier may behave differently depending on the adversarial method used, strongly interacting with language and prompt: For example, for adversarials generated from n-grams sampled from the prompt, the performance of the Spanish ASAP$_{es}$ model is much worse than those of the other languages but only for prompt 1. For the word-based n-grams sampled from the real answers, ASAP$_{es}$ performs much worse on prompt 1 than on prompt 10 but for generic n-grams it is vice versa. Another example is that the English ASAP$_{en}$ model has rather high ARRs across all adversarial methods but for the generic word-based n-grams it is very low but only for prompt 10. We can conclude that in the future, when testing content scoring models for robustness, these complex interplays have to be taken into account and classifiers should be tested against various kinds of adversarial answers and also on various prompts. The dataset we presented here could be used as a benchmark dataset for such endeavors.

In future work, we want to test the behavior of state-of-the art classifiers on the adversarial dataset and more thoroughly analyze the influence of the prompt and features like answer length.

Ethical Considerations

Discussing the ethical implications of developing automatic content scoring systems for real-world scenarios is beyond the scope of this paper. While the aim of the present study is to help detect vulnerabilities of such systems and make them more robust, the insights could also be used maliciously for developing more elaborate methods for cheating purposes. Our adversarial dataset does not include any newly collected data but derives data from already existing corpora and datasets, hence it could inherit biases that may be present in these sources.

Limitations

One clear limitation of this paper is that we draw conclusions from only one content scoring model, which does not produce state-of-the art results when scoring real answers. Other models, especially neural models that do not rely on surface n-grams as features, may behave differently and should be tested in future work. Furthermore, all experiments are based on prompts from the original ASAP-SAS dataset. Other datasets focusing on different kinds of topics and questions are not considered. Finally, our adversarial dataset is not exhaustive in that it (a) only comprises a small set of European languages and (b) only includes a limited number of adversarial methods, whereas more methods are conceivable, e.g. systematically varying the answer length.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

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

Anna Filighera, Sebastian Ochs, Tim Steuer, and Thomas Tregel. 2023. Cheating Automatic Short Answer Grading with the Adversarial Usage of Adjectives and Adverbs. *International Journal of Artificial Intelligence in Education*.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the 8th International Language Resources and Evaluation (LREC'12)*, pages 759–765. International Committee on Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Andrea Horbach, Joey Pehlke, Ronja Laarmann-Quante, and Yuning Ding. 2023. Crosslingual Content Scoring in Five Languages Using Machine-Translation and Multilingual Transformer Models. *International Journal of Artificial Intelligence in Education*.

Henry Kučera and Winthrop Nelson Francis. 1967. *Computational Analysis of Present-day American English*. Brown University Press.

M. Antonia Martí, Mariona Taulé, Lluís Márquez, and Manuel Bertran. 2007. *CESS-Cat Project: CESS-ECE: A Multilingual and Multilevel Annotated Corpus*. Originally available from http://www.lsi.upc.edu/~mbertran/cess-ece/publications.

Ulrike Padó, Yunus Eryilmaz, and Larissa Kirschner. 2023. Short-Answer Grading for German: Addressing the Challenges. *International Journal of Artificial Intelligence in Education*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Nico Willms and Ulrike Pado. 2022. A Transformer for SAG: What Does it Grade? In *Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning*, pages 114–122, Louvain-la-Neuve, Belgium. LiU Electronic Press.

A Appendix

Corpus	**Lang.**	**#Sents.**	**#Tokens**	**#Types**
Brown	English	57,340	1,161,192	56,057
deu-com web	German	10,000	179,093	32,206
fra-ch web	French	10,000	216,787	29,975
CESS-ESP	Spanish	6,030	192,686	25,464

Table 4: Basic statistics of the generic corpora for each language.

Lang.	**Prompt**		
	1	2	10
English	87 (22)	95 (25)	168 (32)
German	135 (34)	123 (35)	201 (58)
French	121 (34)	127 (38)	195 (57)
Spanish	82 (19)	95 (25)	163 (30)

Table 5: Number of words in the prompt material for the different languages. The number in brackets refers to the number of nouns.

Lang.	**Trigram**			**Count**
German	NOUN	ADV	PUNCT	271
	NOUN	ADV	VERB	270
	ADV	ADV	PUNCT	235
	PRON	ADV	ADV	230
	NOUN	ADV	ADP	227
French	VERB	ADV	ADP	324
	VERB	ADV	DET	274
	AUX	ADV	VERB	230
	NOUN	ADV	VERB	212
	NOUN	ADV	NOUN	202
Spanish	VERB	ADV	ADP	404
	NOUN	ADV	ADJ	333
	NOUN	ADV	VERB	331
	VERB	ADV	DET	267
	NOUN	ADV	AUX	250

Table 6: Top 5 POS-trigrams including adverbs per language based on the generic corpora.

Method		Example
Correct student answer		Based on the student's data, plastic B stretched more. b The students could have improved the experiment by resting the plastics at the same length also by doing more than just two trials using sam. Putting same amount of weight in the type of plastic bag. (score 3)
ASAP Word N-Grams	1	most ways how the the what plastic of highest the been trail consistent stretch the have but could trial time. design that most to much the plastic expirement, this
	2	The stretch the was very average not stretched the plastic type tell us D stretched the strengths to add plastic type Based on a few 50% 3.a. allowed the
	3	This to stretch because the same length in both trials The conclusion I conclude that plastic flexible as it In conclusion plastic two ways: The it to the the weights were,
	4	a. Plastic experiment D have the before it stratched. the two tries. this data and the mm it stretched in used three kinds of should improve the experiment type B is the
	5	3 trials are always in the starting length of ability to stretch was plastic the most weight, without stretching. they were the same length can conclude that plastic type
Generic Word N-Grams	1	separators what have one School have Zeising, posts District Democrats wilderness in Vikings Mantle are committee more the from non-profit the Faced you teeth. Grapefruit well is The The a Karen,
	2	plane to hopes to Central Catholic The go-go-go without after information how this crowding horses what to volunteered an has brought and Mrs. high-sounding titles When bouncy show received a to work. (*)
	3	Suddenly also have to in 1885 concluded terminate special sessions The United average. years a slave to Mantle is the Atlanta area Cotten construed might allow the to go up about who said
	4	said. Bursts him, was retracted before On the last March, a mighty primary and the fall bit lost at least four-year terms will expire law could not suspend among Democratic district leaders (*)
	5	the state. The from Texas A& I College a peak this year, came an expanding share of the proposal modest $8,250 to provide Christmas gifts vehicle traffic on Eddy Street " to have these laws
ASAP Char N-Grams	1	h2hrth ttis os tma lsb he hste a Tu. ha tbft sud hoeds ao otnuf Btiic is. te t. v xorel gt oce atp Ihe Hhse d Pwwe aob hlyo fn pi , atl edd anc spsv og vnltm trrt1e 2o)e chwi wa elu afite itss mo as trut ot nebtsa llf ols tne eayer tr eac h Apcyr asda ha
	2	TananM t. tiro tsefi ckehe hehty pal of st igf ingri asntl eud deea er Ac he hedtpM be rir eesm eh2 co ic as T2veh e Anve hecst ild rofpos sho te desl dlat sss os thas exs tcts tono uvamm plls tcld hei s Awr sypt utcI tpl clca we et inbyns igm aueti
	3	Ils testt heei guttem eestch te su ro vpe rnto Alhe oa yst resti rim, plla sistnd iexp tco chafon tdu cs Ati cra le ise igan dsti st td atheu ng sat cali ches could tthdh aft h Fro rese ta cs showb . Otc hnd2ul dterto nyreed ob es Twl dbis tnc l
	4	Tmost sedb ndpedb eture oic tywe igret cprovg usth ertu ldbh etyde nt lasts edt ype Agt h, nc lu heycve imicp ou ldbw eigste dcw eiex pe sticex peorei tud ewhat edthla sto uplthe emproh efohe rea tatth ei chths, bestic plas ia lsig nin oftt af o
	5	Ifhave aperim used. sthatt retcme to fu nto fo, in sargr ay mo rei mental de sofpla efore ictyp untof there rialsa 3rdts tthet ingte Dplald havges toepl asm to13 nthovf th efretc het heps ts tro wing Ty peB d23mm lastis wast dt heswhi ch opemu
Generic Char N-Grams	1	: ewto . t aneape lnhrcs apeaoae geef wnnttdu eyes cg roniooe nanms trraer ynn aaren Mnhah eey8ryn rni. sns (*)
	2	Tonn tedolli ioe Ia yme Ola eeinxp th nd, ssc thttone bh1re unrcti il as leisw he s: 5nuhe ate d Amalhe ms y. l. at ats eof asa se Hseso edi tse Mailnd otiers veftg rm istna in tr' 'tsU dmalP ltte Du tolsr pe'. ma itrig hlmua -ht al tu tuprefi 3, tse aym ngreto hai nldc aB . o rami t Tholno
	3	PofarL anyl e Coin1a vear k Mindov er ott hurihem tinspe wil"d ea anad edalol evckti npneecl ahel Whi guy pi cdo ze Cawo -a tir daendP riB lupuras swhohe f Ad mher pla ibltwa orb kslint vicuca nee tob thef irbroic aewa dere codedon gray atn Pen ted (*)
	4	Bipinv erc iceo tw itse, Piveta reeke thd sligr ouispl t, bosi tya tor, i fa ve rsfl amt he Ssf in Aru nr th ahamp nghergu eo rac tywi tfalemo reby hiontti onin th Misi onf lcondam epin g', he lkert o8pelv eo ratpita e Na titthe la trgemel dtmal dte19dS ouytoms ary er ste. (*)
	5	ff erege tp etor ea er cones ev easemie sneapla ceerth eri ve w Secr estartl ttobo vert and Bead esi atio nons lir Ti bea consht fou hers mh erdaon ledsco nt ajor -s pro cntendn cingi ngalkwa sde Un ivM artir, altsuna nhata less , sesda yere cotly pu Th eye cni ca tsay sver y1t hemi 20intei gnf (*)

Table 7: Examples of a correct student answer and generated adversarials for each method for prompt 2 of the original ASAP-SAS corpus. All adversarial answers, except for the ones marked with (*), were given at least a score of 1 based on the $ASAP_{orig}$ or $ASAP_{orig300}$ model. (Table continued on next page)

Method		Example
Prompt N-Grams	1	student's the the side a remove the five of and plastic the like freely on of allow and experimental one the and/or have and down table on following ways and/or
	2	the procedure student's data remaining three following investigation clamp to of the its length the following is hanging hanging freely type of from the edge of length tape ways the types repeat one type bottom edge the clamp student recorded the student validity of of plastic the experimental length tape student recorded
	3	for five minutes different polymer plastics a second trial student recorded the table so that of one type the student recorded exactly like the measure the length take a sample stretchability procedure take have improved the different polymer plastics could have improved
	4	student recorded the following to test four different have improved the experimental the length of the side of the table and/or validity of the the plastic types repeat recorded the following data them to hang for performed the following investigation improved the experimental design to the bottom edge to the bottom edge
	5	for stretchability procedure take a and/or validity of the results of the table attach a clamp to the bottom edge exactly for the remaining three remaining three plastic samples perform the length of the plastic of the plastic types repeat the top edge of the
Prompt Nouns		minutes procedure student plastics ways validity
Inserting Adjectives		Based on the **physical** student 's **right** data , **similar** plastic **southern** D stretched the same **common** length for both **christian** trials . Two **red** ways the **last** student could have improved the experimental **fine** design would be to on the **central** data **little** table , say how long each **last** type of **complete** plastic is before the **american** student started the **normal** experiment . Another **central** way the **high** student could of improved the experimental **open** design would be to have done only one **nuclear** trial instead of two.
Inserting Adverbs		**certainly** To improve this experiment the student should have mentioned the 4 different types of plastic if mentioned , it would give a more accurate reason as why one type of plastic is more / less stretchable than the other . [...]

Table 7: (continued) Examples of a correct student answer and generated adversarials for each method for prompt 2 of the original ASAP-SAS corpus. All adversarial answers, except for the ones marked with (*), were given at least a score of 1 based on the $ASAP_{orig}$ or $ASAP_{orig300}$ model.

Method	English (orig)			English (orig300)			English (en)			German (de)			French (fr)			Spanish (es)		
	Prompt			Prompt			Prompt			Prompt			Prompt			Prompt		
	1	2	10	1	2	10	1	2	10	1	2	10	1	2	10	1	2	10
ASAP Word 1-grams	0.57	0.427	0.998	0.591	0.037	1.0	0.614	0.683	0.999	0.793	0.423	0.8	0.71	0.673	0.993	0.317	0.618	0.891
ASAP Word 2-grams	0.484	0.303	0.997	0.519	0.037	0.999	0.51	0.696	1.0	0.768	0.431	0.796	0.688	0.628	0.995	0.291	0.537	0.884
ASAP Word 3-grams	0.46	0.274	1.0	0.492	0.042	1.0	0.528	0.625	1.0	0.774	0.379	0.821	0.701	0.562	0.987	0.246	0.509	0.89
ASAP Word 4-grams	0.452	0.26	0.999	0.45	0.039	0.999	0.522	0.611	0.999	0.73	0.349	0.809	0.618	0.544	0.99	0.272	0.465	0.866
ASAP Word 5-grams	0.429	0.224	0.997	0.461	0.043	1.0	0.473	0.593	1.0	0.704	0.343	0.771	0.617	0.515	0.987	0.254	0.435	0.871
Generic Word 1-grams	1.0	0.998	0.178	1.0	0.0	0.0	1.0	1.0	0.711	1.0	0.935	0.878	0.999	1.0	0.93	0.962	0.978	0.189
Generic Word 2-grams	1.0	1.0	0.148	1.0	0.0	0.0	1.0	1.0	0.645	1.0	0.929	0.858	1.0	1.0	0.909	0.936	0.944	0.154
Generic Word 3-grams	1.0	0.999	0.167	1.0	0.0	0.0	1.0	1.0	0.68	1.0	0.9	0.819	0.997	0.999	0.894	0.942	0.938	0.162
Generic Word 4-grams	1.0	1.0	0.176	1.0	0.0	0.0	1.0	1.0	0.65	1.0	0.889	0.755	1.0	0.999	0.894	0.954	0.933	0.129
Generic Word 5-grams	1.0	0.998	0.149	1.0	0.0	0.0	1.0	1.0	0.633	0.996	0.848	0.752	1.0	0.999	0.895	0.948	0.91	0.139
ASAP Char 1-grams	1.0	0.999	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
ASAP Char 2-grams	1.0	0.999	1.0	0.997	0.0	1.0	1.0	1.0	1.0	1.0	0.997	1.0	1.0	1.0	1.0	1.0	1.0	1.0
ASAP Char 3-grams	1.0	0.982	1.0	0.994	0.0	0.998	1.0	0.999	1.0	0.999	0.983	1.0	1.0	1.0	1.0	0.997	1.0	1.0
ASAP Char 4-grams	1.0	0.954	1.0	0.984	0.0	0.999	1.0	1.0	1.0	0.999	0.964	0.998	1.0	0.999	1.0	0.994	0.998	1.0
ASAP Char 5-grams	0.999	0.938	1.0	0.972	0.0	0.998	1.0	0.996	1.0	0.994	0.942	0.999	0.999	0.998	1.0	0.977	0.99	1.0
Generic Char 1-grams	1.0	1.0	0.889	1.0	0.0	0.0	1.0	1.0	0.998	1.0	1.0	0.998	1.0	1.0	0.999	1.0	1.0	0.653
Generic Char 2-grams	1.0	0.999	0.795	1.0	0.0	0.0	1.0	1.0	0.97	1.0	0.996	0.956	1.0	1.0	0.999	0.999	0.997	0.553
Generic Char 3-grams	1.0	1.0	0.754	1.0	0.0	0.0	1.0	1.0	0.945	1.0	0.98	0.893	1.0	1.0	0.992	0.997	0.993	0.5
Generic Char 4-grams	1.0	1.0	0.711	1.0	0.0	0.0	1.0	1.0	0.92	1.0	0.969	0.847	1.0	1.0	0.99	1.0	0.979	0.489
Generic Char 5-grams	1.0	1.0	0.695	1.0	0.0	0.0	1.0	1.0	0.907	1.0	0.954	0.844	1.0	1.0	0.991	0.999	0.973	0.475
Prompt 1-grams	0.988	0.757	0.244	0.952	0.154	0.001	0.996	0.957	0.459	1.0	0.331	0.478	0.993	0.396	0.466	0.607	0.249	0.278
Prompt 2-grams	0.997	0.766	0.269	0.97	0.183	0.002	0.992	0.96	0.489	1.0	0.325	0.459	0.996	0.391	0.457	0.602	0.259	0.277
Prompt 3-grams	0.996	0.803	0.238	0.961	0.151	0.001	0.989	0.962	0.445	1.0	0.317	0.458	0.997	0.428	0.404	0.581	0.269	0.262
Prompt 4-grams	0.989	0.785	0.238	0.954	0.177	0.001	0.994	0.971	0.46	1.0	0.328	0.428	0.998	0.447	0.412	0.555	0.252	0.241
Prompt 5-grams	0.989	0.77	0.205	0.958	0.214	0.001	0.988	0.963	0.439	1.0	0.325	0.422	0.998	0.454	0.383	0.576	0.249	0.241
Prompt Nouns	0.995	1.0	0.998	1.0	0.01	0.0	0.999	1.0	0.998	1.0	1.0	1.0	0.99	1.0	1.0	1.0	1.0	0.989
Inserting Adjectives	0.941	0.448	0.336	0.884	0.276	0.0	0.99	1.0	0.797	1.0	0.736	0.767	1.0	0.89	0.876	0.999	0.751	0.669
Inserting Adverbs	0.929	0.535	0.43	0.939	0.319	0.0	0.994	1.0	0.859	1.0	0.912	0.93	0.992	0.986	0.974	0.975	0.874	0.808

Table 8: Adversarial Rejection Rate (ARR) for each model for each adversarial method.