Using Adversarial Attacks to Reveal the Statistical Bias in Machine Reading Comprehension Models

Jieyu Lin[2], **Jiajie Zou**[2], **Nai Ding**[1,2]*
[1]Zhejiang Lab / Hangzhou, China
[2]Key Laboratory for Biomedical Engineering of Ministry of Education, College of Biomedical Engineering and Instrument Sciences, Zhejiang University / Hangzhou, China
{ljy5905,jiajiezou,ding_nai}@zju.edu.cn

Abstract

Pre-trained language models have achieved human-level performance on many Machine Reading Comprehension (MRC) tasks, but it remains unclear whether these models truly understand language or answer questions by exploiting statistical biases in datasets. Here, we demonstrate a simple yet effective method to attack MRC models and reveal the statistical biases in these models. We apply the method to the RACE dataset, for which the answer to each MRC question is selected from 4 options. It is found that several pre-trained language models, including BERT, ALBERT, and RoBERTa, show consistent preference to some options, even when these options are irrelevant to the question. When interfered by these irrelevant options, the performance of MRC models can be reduced from human-level performance to the chance-level performance. Human readers, however, are not clearly affected by these irrelevant options. Finally, we propose an augmented training method that can greatly reduce models' statistical biases.

1 Introduction

Reading comprehension tasks are useful to quantify language ability of both humans and machines (Richardson et al., 2013; Xie et al., 2018; Berzak et al., 2020). Deep neural network (DNN) models have achieved high performance on many MRC tasks, but these models are not easily explainable (Devlin et al., 2019; Brown et al., 2020). It is also shown that DNN models are often sensitive to adversarial attacks (Jia and Liang, 2017; Ribeiro et al., 2018; Si et al., 2019, 2020). Furthermore, it has been shown DNN models can solve MRC tasks with relatively high accuracy when crucial information is removed so that the tasks are no longer solvable by humans (Gururangan et al., 2018; Si et al., 2019; Berzak et al., 2020). All such evidence suggests that the high accuracy DNN models achieve on MRC tasks does not solely rely on these models' language comprehension ability. At least to some extent, the high accuracy reflects exploitation of statistical biases in the datasets (Gururangan et al., 2018; Si et al., 2019; Berzak et al., 2020).

Here, we propose a new model-independent method to evaluate to what extent models solve MRC tasks by exploiting statistical biases in the dataset. As a case study, we only focus on the classic RACE dataset (Lai et al., 2017), which requires MRC models to answer multiple-choice reading comprehension questions based on a passage. The advantage of multiple-choice questions is that its performance can be objectively evaluated. At the same time, it does not require the answer to be within the passage, allowing to test, e.g., the summarization or inference ability of models. Nevertheless, since models are trained to select the right option from 4 options, which are designed by humans and may contain statistical biases, models may learn statistical properties of the right option. Consequently, models may tend to select options with these statistical properties similar to the properties of the right option without referring to the passage and question. Our method is designed to reveal this kind of statistical bias.

The logic of our method is straightforward: For each multiple-choice question, we gather a large number of options that are irrelevant to the question and passage. We ask the model to score how likely each irrelevant option is the right option. If a model is biased, it may always assign higher scores to some irrelevant options than others, even if all the options are irrelevant. If a model is so severely biased, which turns out to be true for all models tested here, it may assign higher scores to some irrelevant options than the true answer and select the irrelevant option as the answer. Here, the irrelevant

*Corresponding author: Nai Ding

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 333–342
August 1–6, 2021. ©2021 Association for Computational Linguistics

options that are often selected as the answer are referred to as magnet options.

2 Dataset and Pre-trained Models

We used RACE dataset in our experiment (Lai et al., 2017), which is a large-scale reading comprehension data set covering more than 28,000 passages and nearly 100,000 questions. The task was to answer multi-choice questions based on a passage. Specifically, each question contained a triplet (p_i, q_i, o_i), where p_i denoted a passage, q_i denoted a question, and o_i denoted a candidate set of 4 options, i.e., $o_i = \{o_{i,1}, o_{i,2}, o_{i,3}, o_{i,4}\}$. Only one option was the correct answer, and the accuracy was evaluated by the percent of questions being correctly answered.

We tested 3 pre-trained language models, i.e., BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and ALBERT (Lan et al., 2019). For each model, we separately tested the base version and large version. We built our models based on pre-trained transformer models in the Huggingface (Wolf et al., 2020). We fine-tuned pre-trained models based on the RACE dataset and the parameters we used for fine-tuning were shown in Appendix A.1.

The passage, question, and an option were concatenated as the input to models, i.e., [CLS, p_i, SEP, q_i, $o_{i,j}$, SEP]. The 4 options were separately encoded. The concatenated sequence was encoded through the models and the output embedding of CLS was denoted as $C_{i,j}$. We used a linear transformation to convert vector $C_{i,j}$ into a scalar $S(o_{i,j})$, i.e., $S(o_{i,j}) = WC_{i,j}$. The scalar $S(o_{i,j})$ was referred to as the score of the option $o_{i,j}$. A score was calculated for each option, and the answer to a question was determined as the option with the highest score, i.e., $\text{argmax}_j S(o_{i,j})$.

3 Adversarial Method

3.1 Screen Potential Magnet Options

We evaluated potential statistical biases in a model by giving it a large number of irrelevant options. For each question, we augmented the options using a set of irrelevant options, i.e., $O_A = \{o_{a1}, o_{a2}, ..., o_{aN}\}$. O_A was randomly selected from the RACE dataset with 2 constraints. First, the options belonged to questions that were not targeted at passage p_i. Second, none of the options in O_A was identical to any of the original options in o_i. The augmented question was denoted as $(p_i, q_i, \{o_{i,1}, o_{i,2}, o_{i,3}, o_{i,4}, o_{a1}, ..., o_{aj}, ..., o_{aN}\})$. A score was independently computed for each option using the procedure mentioned above. Since the options in O_A were irrelevant, an ideal model should never select them as answers. If $\text{max}_j S(o_{i,j}) < S(o_{ak})$ for any k, however, the model would select the k^{th} irrelevant option as the answer. We define an interference score T_k using the following equation.

$$T_k = \frac{1}{N}\sum_{i=1}^{N} T_{i,k}, \qquad where$$

$$T_{i,k} = \begin{cases} 1, & if \quad \text{max}_j S(o_{i,j}) < S(o_{ak}) \\ 0, & otherwise \end{cases}$$

For an ideal model, $T_{i,k}$ should always be 0. For a model that makes mistakes but shows no consistent bias, the interference score should be comparable for all o_{ak}. If the model is biased, the interference score may be always high for some options so that the model always selects them as the answer whether they are relevant to the question or not.

3.2 Adversarial Attack

We constructed an adversary attack to the MRC models using one magnet option. For each question, we replaced a wrong option with a magnet option, i.e., o_{ak}. The replaced option set was $\{o_{i,1}, o_{i,2}, o_{i,3}, o_{ak}\}$. The passage and the question were not modified, and the answer did not change. An example was shown in Figure 1. If the model chooses the original answer even when a magnet option is introduced, it is stable, not sensitive to the attack. In contrast, if it chooses the magnet option, i.e., o_{ak}, as the answer, it is successfully attacked.

4 Results and Analyses

4.1 Experiments Setup

To screen potential magnet options, we constructed a large set of irrelevant options, i.e., O_A, by randomly selecting 300 passages from the RACE test set, which were associated with 1064 questions. Furthermore, to test whether options in the training set can cause stronger interference, we also randomly selected 300 passages from the RACE training set, which had 1029 questions. The options from the test and training set were pooled to create O_A, which had 8372 options in total.

Passage: "...Quantum computers could be able to do what modern supercomputers are unable to do by using transistors that are able to take on many states at the same time..."

Question: According to the text, quantum computing _ .

Original Options:	**Adversarial Options:**
A. can reduce the cost of computers	*A*. can reduce the cost of computers
B. can make computers run by themselves	*B*. misfortune may be an actual blessing
C. **will work by using transistors**	*C*. **will work by using transistors**
D. has been put in use so far	*D*. has been put in use so far
Model Choice: *C* – correct	**Model Choice:** *B* – incorrect, **successfully attacked**
A, B, or D – incorrect	*C* – correct, not attacked
	A or D – incorrect, not attacked

Figure 1: An example of the task and adversarial attack. The option in bold is the true answer, and the option in red indicates the irrelevant option that was used for attack.

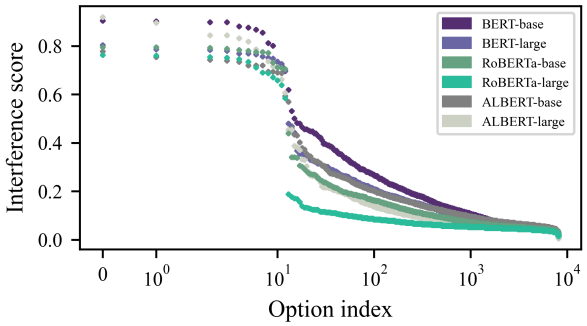

Figure 2: Interference score evaluated based on a subset of questions.

For such a large number of irrelevant options, it was computationally challenging to evaluate the interference score of each option based on each question in the RACE test set. Therefore, as a screening procedure, we first randomly selected 100 passages from the RACE test set, which have a total of 346 questions. The interference score for each of the 8372 irrelevant options was evaluated based on the 346 questions.

After potential magnet options were determined by the screening procedure, the interference score of magnet options were further evaluated using all questions in RACE test set. For RACE test set, the accuracy of the models ranged between about 0.6 and 0.85, with RoBERTa-large achieving the highest performance (Table 1).

4.2 Screening for Magnet Options

The interference score for 8372 options was independently calculated for each model. Results were shown in Figure 2, where the interference score was sorted for each model. It is found that most of the irrelevant options had a non-zero interference score, and some irrelevant options yielded high interference scores around 0.8, which meant the models would choose those irrelevant options as the answer for about 80% of the questions. Irrelevant options from the training and test sets had similar interference scores (Appendix B.1).

It was found that the options with exceptionally high interference scores around 0.8 were options that combined other options, such as "all the above", which were called the option-combination series. However, not all the magnet options were from the option-combination series. Normal statements, e.g., "The passage doesn't tell us the end of the story of the movie", could also reach an average interference score around 0.34.

The correlation between the interference score between models were shown in Appendix B.2. We separately showed the results for options from the option-combination series and the others. The correlation coefficient between models had an average value around 0.76, which proved that the interference score was correlated across models. From another perspective, it also implied that our method could work as a model-insensitive adversarial attack method.

4.3 Validate Magnet Options and Adversarial Attack

We further evaluated the interference score of potential magnet options based on all the questions in the RACE test set. To construct a set of magnet options for this analysis, we averaged the interference score across 3 models, i.e., BERT-large, RoBERTa-large, and ALBERT-large. All options in O_A were sorted based on the average score, and we selected 20 options with the highest interference scores to construct the magnet option set, with the

	BERT		ALBERT		RoBERTa	
Version	base	large	base	large	base	large
Original accuracy	0.614	0.681	0.683	0.752	0.738	0.846
Adversarial accuracy[1]	0.094	0.167	0.217	0.064	0.166	0.297
Adversarial accuracy[2]	0.381	0.524	0.334	0.506	0.656	0.798

Table 1: Model performance on the RACE test set and model performance after being attacked. The superscript 1 meant use "A, B and C" to attack, and the superscript 2 meant use "The passage doesn't tell us the end of the story of the movie" to attack.

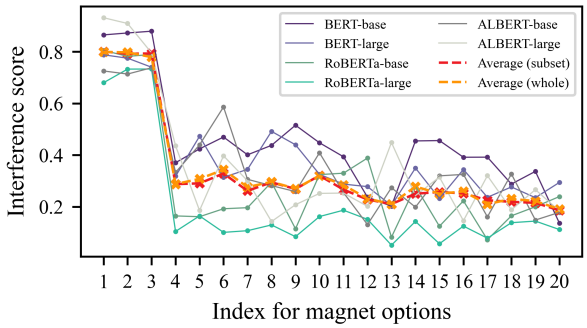

Figure 3: Interference score evaluated based on the whole RACE test set.

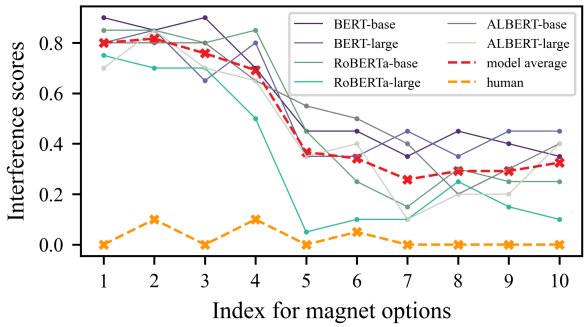

Figure 4: Interference score for the human experiment and the corresponding interference scores for the models.

following constraint: Since options with the highest interference scores were often from the option-combination series, to increase diversity, we only included 3 options from the option-combination series. We listed all the 20 magnet options in Appendix A.2. The interference score calculated based on the whole RACE test set was shown in Figure 3, which was very similar to the results based on the subset of 346 questions in Figure 2 (comparing average-whole and average-subset in Figure 3).

Table 1 showed the accuracy of models when attacked by 2 example magnet options. When attacked, the model performance could drop by as much as 0.68.

4.4 Human Evaluation

Next, we verified whether humans were also confused by the magnet options. We randomly selected 20 questions and 10 magnet options. The 10 magnet options selected were listed in Appendix A.3. Ten questions were not modified while the other 10 questions were attacked using the procedure shown in Figure 1. Twenty human evaluators answered these 20 questions online. The accuracy of humans did not reduce under attack (0.90 in the original samples and 0.94 in the adversarial samples). The interference score for humans, also the corresponding interference score for the models, was shown in Figure 4. Humans were not confused by the magnet options.

4.5 Training with Adversarial Examples

To reduce sensitivity to magnet options and to potentially reduce the statistical biases of MRC models, we proposed an augmented training method and tested the method using the base version of all models. In the augmented training method, 400 options with the highest interference scores were selected as the irrelevant option set. For each question in the RACE training set, the option set was augmented by adding an option randomly chosen from the irrelevant option set. In other words, although each original question has 4 options, during the augmented training each question has 5 options, including the 4 original options and a randomly chosen irrelevant option. We fine-tuned pre-trained models based on the training set with augmented options.

The accuracy of models fine-tuned using augmented options were shown in Table 2, comparable to the original accuracy in Table 1. When attacked, however, the accuracy of models fine-tuned using augmented options were much higher than the adversarial accuracy in Table 1.

The 1000 options with the highest interference

base version	BERT	ALBERT	RoBERTa
Original accuracy	0.601	0.689	0.723
Adversarial accuracy[1]	0.576	0.681	0.725
Adversarial accuracy[2]	0.670	0.740	0.778

Table 2: Model performance on the RACE test set based on augmented training.

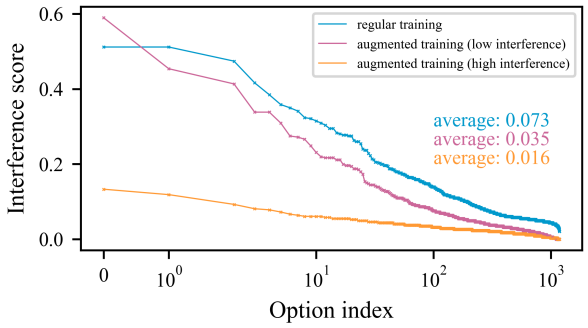

Figure 5: Interference score of 1186 randomly chosen options that are not used in augmented training.

scores were selected to evaluate the effect of augmented training, as shown in Appendix C. Result showed that the interference score dropped for both the 400 options used for augmented training and the other 600 options that were not used for training. Therefore, the effect of augmented training could generalize to samples not used for augmented training.

Another experiment was implemented to explore the impact of irrelevant option set selection. We separately used options with high and low interference scores for training and found that options with higher interference score were more effective at reducing statistical biases (Figure 5).

4.6 Interference Score Analysis

Did the statistical biases revealed in previous analyses originate from the pre-training process or the fine-tuning process? Without fine-tuning, the pre-trained models performed poorly on RACE. However, results showed that such an imprecise model could show strong biases (Appendix B.3). Interestingly, the interference score was not correlated between the pre-trained model and the fine-tuned model, suggesting that fine-tuning overrode the biases caused by pre-training and introduced new forms of biases.

5 Related Work

Our attack strategy distinguishes from previous work in two ways. First, unlike, e.g., gradient-based methods (Ebrahimi et al., 2018; Cheng et al., 2020), our method does not require any knowledge about the structure of DNN models. Second, some methods manipulate the passage in a passage-dependent way (Jia and Liang, 2017; Si et al., 2020; Zhao et al., 2018), while our method manipulate the options in a passage-independent way. Furthermore, we proposed a strategy to train more robust models that are insensitive to our attack.

Here, we restricted our discussion to RACE, but our method is applicable to other tasks in which the answer is selected from a limited set of options. For example, for span extraction tasks, such as SQuAD, the method will insert a large number of irrelevant phrases into the passage and analyze which phrases are often selected as the answer. In this way, our method is similar to the trigger-based attack methods (Wallace et al., 2019), but the difference is that our method test whether the inserted irrelevant phrase is selected as the answer while the trigger-based methods test whether the content following the trigger phrase is selected.

6 Conclusion

In summary, we propose a new method to evaluate the statistical biases in MRC models. It is found that current MRC models have strong statistical biases, and are therefore sensitive to adversarial attack. When attacked using the method proposed here, model performance can drop from human-level performance to chance-level performance. To alleviate sensitivity to such attacks, we provided an augmented training procedure that effectively enhances the robustness of models.

Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful suggestions and comments. Work supported by Major Scientific Research Project of Zhejiang Lab 2019KB0AC02 and National Natural Science Foundation of China 31771248.

References

Yevgeni Berzak, Jonathan Malmaud, and Roger Levy. 2020. STARC: Structured annotations for reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5726–5735, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Minhao Cheng, Jinfeng Yi, Pin-Yu Chen, Huan Zhang, and Cho-Jui Hsieh. 2020. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 3601–3608. AAAI Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, Melbourne, Australia. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Qiu Ran, Peng Li, Weiwei Hu, and Jie Zhou. 2019. Option comparison network for multiple-choice reading comprehension. *arXiv preprint arXiv:1903.03033*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865.

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 193–203.

Chenglei Si, Shuohang Wang, Min-Yen Kan, and Jing Jiang. 2019. What does BERT learn from multiple-choice reading comprehension datasets? *CoRR*, abs/1910.12391.

Chenglei Si, Ziqing Yang, Yiming Cui, Wentao Ma, Ting Liu, and Shijin Wang. 2020. Benchmarking robustness of machine reading comprehension models. *CoRR*, abs/2004.14004.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen,

Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Qizhe Xie, Guokun Lai, Zihang Dai, and Eduard Hovy. 2018. Large-scale cloze test dataset created by teachers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2344–2356, Brussels, Belgium. Association for Computational Linguistics.

Shuailiang Zhang, Hai Zhao, Yuwei Wu, Zhuosheng Zhang, Xi Zhou, and Xiang Zhou. 2020. DCMN+: dual co-matching network for multi-choice reading comprehension. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9563–9570. AAAI Press.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2018. Generating natural adversarial examples. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

	BERT		ALBERT		RoBERTa	
version	base	large	base	large	base	large
learning rate	1.00E-05	1.00E-05	2.00E-05	1.00E-05	1.00E-05	1.00E-05
train epochs	5	5	/	/	4	4
train steps	/	/	12000	12000	/	/
train batch size	16	24	32	32	16	16
warmup steps	0	0	1000	1000	1200	1200
weight decay	0	0	0	0	0.1	0.1

Table 3: Hyperparameters for fine-tuning on RACE. We adapted these hyperparamemers from Lan et al. (2019); Liu et al. (2019); Ran et al. (2019); Zhang et al. (2020).

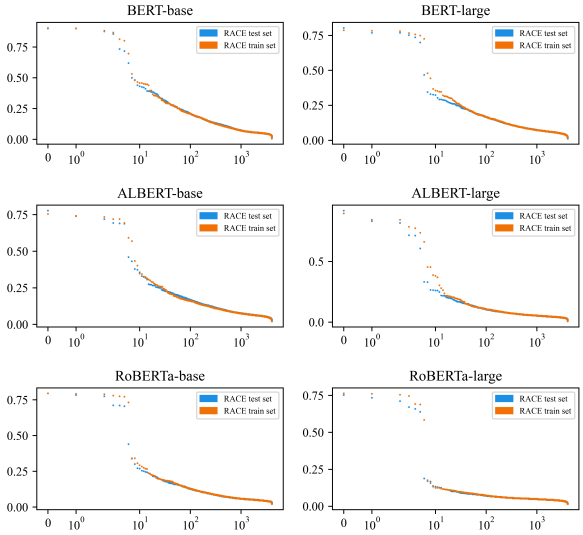

Figure 6: Interference score evaluated based on a subset of questions.

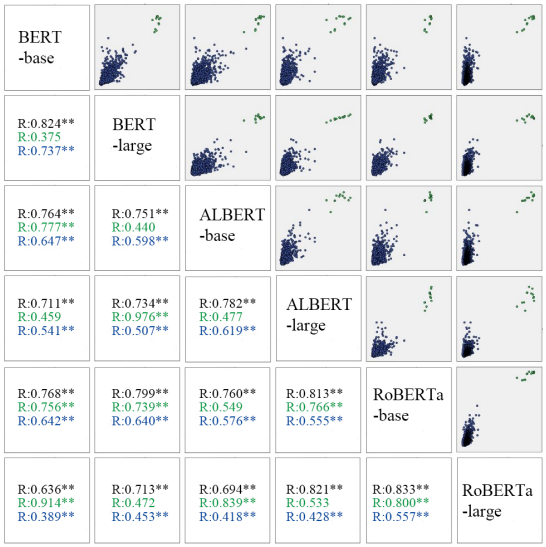

Figure 7: The scatter matrix diagram of the interference scores of the irrelevant options among models.

A Experimental Details

A.1 Fine-tuning Parameters

The parameters we used in the process of fine-tuning the pre-trained models were shown in Table 3.

A.2 Magnet Options for Validate

The 20 magnet options used for evaluating the interference scores in Section 4.3 were shown as following. The sentences selected from the RACE training set were shown in bold.

1. A, B and C
2. **all of A, B and C**
3. All of the above.
4. **Not all of it can be avoided.**
5. It's well beyond what the author could be responsible for.
6. **The passage doesn't tell us the end of the story of the movie**
7. didn't give the real answer
8. **make us know it's important to listen to people who offer a different perspective through his experience**
9. **give us a turning point in mind**
10. **not strictly stuck to**
11. You should purposely go out and make these mistakes so that you can learn from them and not have them ruin your entire life.
12. what's inside a person is much more important than his/her appearance.
13. **Not all of it is man-made Ming dynasty structure.**
14. **introduce the topic of the passage**
15. **The central command didn't exactly state what had caused the crash.**
16. **one good turn deserves another.**
17. the growing population is not the real cause of the environment problem.,
18. **misfortune may be an actual blessing.**

BERT-base	Correlation coefficient	accuracy	Average interference score
Pre-trained model	-0.023	0.315	0.0518
Partly fine-tuned model	0.069*	0.315	0.0214
Fine-tuned model	1	0.613	0.0713
RoBERTa-base	Correlation coefficient	accuracy	Average interference score
Pre-trained model	-0.021	0.225	0.3553
Partly fine-tuned model	0.088**	0.289	0.2282
Fine-tuned model	1	0.743	0.0569
ALBERT-base	Correlation coefficient	accuracy	Average interference score
Pre-trained model	-0.013	0.254	0.1483
Partly fine-tuned model	0.231**	0.39	0.1043
Fine-tuned model	1	0.702	0.0703

Table 4: Interference score of 1000 randomly selected irrelevant options for the same model architecture before and after fine-tuning. Correlation coefficient was counted between the interference score before and after fine-tuning (**$P < 0.01$, and *$P < 0.05$).

19. may meet with difficulties sometimes
20. good answers are always coming when we think outside of the box

A.3 Magnet Options for Human Evaluation

The 10 magnet options used for human evaluating in Section 4.4.

1. all the above
2. Both B and C
3. do all of the above
4. A and B
5. not strictly stuck to
6. The passage doesn't tell us the end of the story of the movie
7. It's well beyond what the author could be responsible for.
8. You should purposely go out and make these mistakes so that you can learn from them and not have them ruin your entire life.
9. make us know it's important to listen to people who offer a different perspective through his experience
10. Not all of it is man-made Ming dynasty structure.

B Study of Interference Score

B.1 Comparison of Irrelevant Options from RACE Training and Test Set

Different models in Figure 2 were separately shown in Figure 6. It denoted that irrelevant options from the training and test sets had similar interference score. Only in BERT-large and ALBERT-large models, the interference scores of the irrelevant options from the training set were higher than those

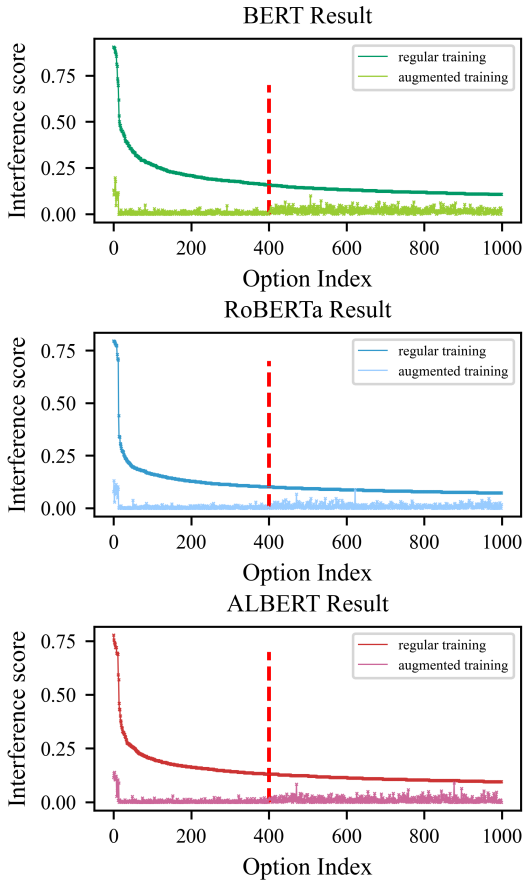

Figure 8: Interference score comparison of models evaluated based on a subset of questions.

from the test set in a certain range.

B.2 Comparison of Interference Scores Based on Different Models

The scatter matrix diagram of the interference scores of the irrelevant options among different models was shown in Figure 7. The detailed experimental process was described in Section 4.2. Here, text in black showed the correlation coefficient of all options; text in green showed the options of the option-combination series; text in blue showed the options except the option-combination series.

In general, the interference scores between models had high correlation coefficients. Models from the same architecture were more likely to have similar interference scores.

B.3 Comparison of Interference Scores During Fine-tuning

For each model architecture, the pre-trained model, partly fine-tuned model (fine-tuned the linear transformation mentioned in Section 2), and fully fine-tuned model were collected, and were used to evaluate the interference score of 1,000 randomly selected irrelevant options. The results were shown in Table 4. The subset of questions mentioned in Section 4.1 were used to evaluate the interference score.

C Augmented Training Result

The augmented training results were shown in Figure 8. In the figures, the left side of the red line contains the irrelevant options that were used in augmented training, and the right is the irrelevant options that were not involved in augmented training.