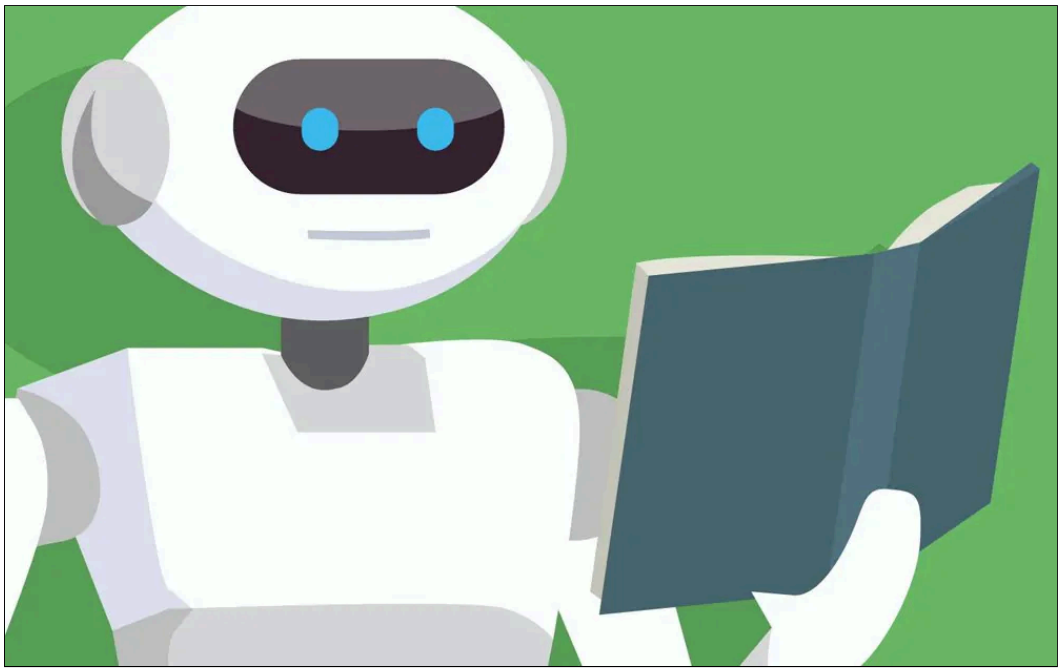
The Struggle to Stop AI From Cheating on Tests

Originally published at https://www.unite.ai/the-struggle-to-stop-ai-from-cheating-on-tests/



New research findings from a Chinese university offer an insight into why generative natural language processing models such as GPT-3 tend to 'cheat' when asked a difficult question, producing answers that may be technically correct, but without any real understanding of *why* the answer is correct; and why they demonstrate little or no ability to explain the logic behind their 'easy' answers. The researchers also propose some new methods to make the systems 'study harder' during the training phase.

The problem is twofold: firstly, we design systems that attempt to achieve results quickly and with an optimal use of resources. Even where, as with GPT-3, the resources may be considerably greater than the average NLP research project is able to muster, this culture of results-driven optimization still pervades the methodology, because it has come to dominate academic convention.

Consequently, our training architectures reward models that converge quickly and produce apparently apposite responses to questions, even if the NLP model is subsequently unable to justify its response, or to demonstrate how it arrived at its conclusions.

An Early Disposition To Cheat

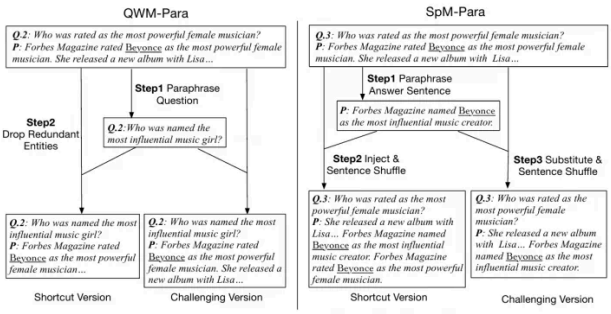
This occurs because the model learns 'shortcut responses' far earlier in the training than it learns more complicated types of knowledge acquisition. Since increased accuracy is often rewarded quite indiscriminately throughout training, the model then prioritizes any approach that will let it answer a question 'glibly', and without real insight.

Since shortcut learning will inevitably represent the *first* successes during training, the session will naturally tend away from the more difficult task of gaining a useful and more complete epistemological perspective, that may contain deeper and more insightful layers of attribution and logic.

Feeding AI The 'Easy' Answers

The second problem is that even though recent research initiatives have studied AI's tendency to 'cheat' in this way, and have identified the phenomenon of 'shortcuts', there has until now been no effort to classify 'shortcut'-enabling material in a contributing dataset, which would be the logical first step in addressing what may prove to be a fundamental architectural flaw in machine reading comprehension (MRC) systems.

The new paper, a collaboration between the Wangxuan Institute of Computer Technology and the MOE Key Laboratory of Computational Linguistics at Peking University, tests various language models against a newly annotated dataset which includes classifications for 'easy' and 'hard' solutions to a possible question.



Source: https://arxiv.org/pdf/2106.01024.pdf

The dataset uses paraphrasing as a criteria for the more complicated and deep answers, since a semantic understanding is necessary in order to reformulate obtained knowledge. By contrast, the 'shortcut' answers can use tokens such as dates, and other encapsulating keywords, to produce an answer that is factually accurate, but without any context or reasoning.

The shortcut component of the annotations features question word matching (QWM) and simple matching (SpM). For QWM, the model utilizes entities extracted from the supplied text data and jettisons context; for SpM, the model identifies overlap between answer sentences and questions, both of which are supplied in the training data.

Shortcut Data Almost 'Viral' In Influence In A Dataset

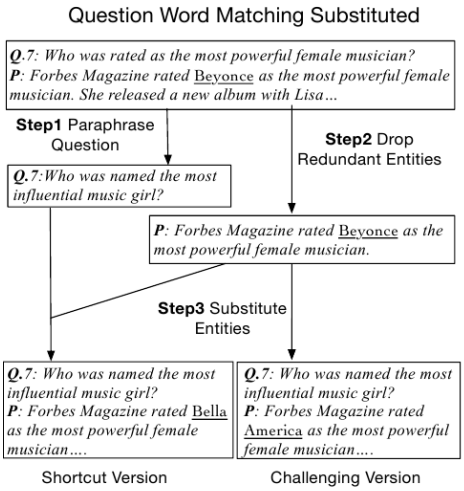
The researchers contend that datasets tend to contain a high proportion of shortcut questions, which make trained models rely on shortcut tricks.

The two models used in the experiments were BiDAF and Google's BERT-base. The researchers observe that even when trained on dataset variations with a higher proportion of 'difficult' questions, both models still perform better on shortcut questions than harder paraphrased questions, despite the small number of examples in the datasets.

This presents 'shortcut data' almost in the context of a virus – that there needs to be very little of it present in a dataset in order for it to be adopted and prioritized in training, under conventional standards and practices in NLP.

Proving The Cheat

One method the research uses to prove how the fragility of a shortcut answer is to substitute an 'easy' entity word for an anomalous word. Where a shortcut method has been used, the logic of the 'cheated' response can't be provided; but where the answer was provided from deeper context and semantic evaluation of a wider range of contributing text, it's possible for the system to deconstruct the error and reconstruct a correct answer.



Substituting 'Beyoncé' (a person) for 'America' (a location), reveals whether the model has any background logic for its answer.

Shortcuts Due To An Economic Imperative

Regarding some of the architectural reasons why shortcuts are so prioritized in NLP training workflows, the authors comment *'MRC models may learn the shortcut tricks, like QWM, with less computational resources than the comprehension challenges, like identifying paraphrasing'*.

This, then, could be an unintended result of standard optimization and resource-preserving philosophies in approaches to machine reading comprehension, and the pressure to obtain results with limited resources in tight time-frames.

The researchers also note:

> *'[Since] the shortcut trick can be used to answer most of the training questions correctly, the limited unsolved questions remained may not motivate the models to explore sophisticated solutions that require challenging skills.'*

If the paper's results are subsequently borne out, it would appear that the vast and ever-growing field of data preprocessing may need to consider 'hidden cribs' in data as a problem to be addressed in the long-term, or else revise NLP architectures to prioritize more challenging routines for data ingestion.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More