Analysis and Mitigation of Dataset Artifacts in OpenAI GPT-3

Mingi Ryu, Kenta Nakajima
University of Texas at Austin
mryu@utexas.edu, kenta.nakajima@utexas.edu

Abstract

With the recent release of public beta, we took full advantage of OpenAI's Models-as-a-Service (MaaS) offering of GPT-3 to analyze and mitigate dataset artifacts in a model that has one of the highest number of parameters. Recent studies in dataset artifacts and adversarial attacks suggest that state of the art (SoTA) NLP models are susceptible to spurious correlations in training datasets. We decided to investigate GPT-3 on dataset artifacts taking advantage of its large scale and task-agnostic pre-training. We began by verifying few-shot capabilities of GPT-3 in order to lay the groundwork for analysis. Furthermore, we employed our approach to fine-tuning for the natural language inference (NLI) task. Using SNLI as a baseline, we carried out several experiments with Adversarial NLI (ANLI) to evaluate the performance and robustness of GPT-3. Our findings suggest that using adversarial datasets could mitigate dataset artifacts in GPT-3 at a negligible overall performance cost.

1 Introduction

Until recently, the GPT-3 model has been limited to academic researchers and a certain group of people who were able to get the private beta access early on. Unlike other state of the art (SoTA) models such as BERT and ELECTRA, GPT-3 is not available to the public for download. In addition, it could cost up to 12 million dollars for one to train it by oneself (Turner, 2020). Fortunately, OpenAI removed the waitlist for GPT-3 for anyone to use on November 18, 2021 (OpenAI, 2021).

GPT-3 is an autoregressive language model with 175 billion parameters that greatly improves on task-agnostic, few shot performance (Brown et al., 2020). For most of the tasks, GPT-3 can be applied without any gradient updates or fine-tuning, with few-shot demonstrations (examples) specified via text interaction (prompt) with the model (Brown et al., 2020).

Due to advantages in scale and task-agnostic training, GPT-3 is less vulnerable to spurious correlations in training data and can sometimes be competitive with state-of -the-art fine-tuned models even with “few-shot learning” (Brown et al., 2020). At the same time, it’s also found that GPT-3 can struggle at certain tasks such as natural language inference and reading comprehension due to its autoregressive language modeling design (Brown et al., 2020).

With respect to dataset artifacts, GPT-3 zero/one/few-shot has the potential to avoid the spurious correlations found in training datasets (Brown et al., 2020). Even with traditional fine-tuning, the greater expressiveness of the model should improve robustness in all tasks and be less susceptible to dataset artifacts (Brown et al., 2020). Considering these unique characteristics, GPT-3 is more suitable for investigating dataset artifacts than SoTA models where such problems have already been studied extensively due to the ease of access.

Dataset artifacts, where statistical irregularities allow a model to perform beyond what should be achievable without access to the context, is a well documented issue in most of the NLP tasks (Poliak et al., 2018). For tasks such as NLI, it's possible to train a hypothesis-only model that can outperform some of the models trained with premise as context (Poliak et al., 2018). More specifically, a hypothesis-only model can achieve 67% accuracy for SNLI due to annotication artifacts such as lexical choice and sentence length (Gururangan et al., 2018).

These artifacts are quite common across various tasks and datasets. For question answering (QA), a large language model such as BERT and GPT can perform well on many of the QA benchmarks without having access to context as external knowledge (Roberts et al., 2020). This brings into question whether these models are simply exploiting artifacts in both pre-training and fine-tuning datasets to answer questions instead of retrieving information from the provided context.

In order to mitigate such artifacts, one of the promising approaches has been on constructing a much harder benchmark (challenge set) that is curated and generated adversarially where both the model and the annotators operate in a loop. For NLI, adversarial NLI (ANLI) uses adversarial human-and-model-in-the-loop procedure to construct a new dataset (Nie et al., 2020). For QA, adversarial QA (AQA) uses human annotation with a model in the loop process to construct a new set (Bartolo et al., 2020). These challenge sets have demonstrated improvement in robustness and mitigation of dataset artifacts while maintaining near SoTA performance.

In the following sections, we describe our experiment approaches in regards to few-shot learning, fine-tuning, and dataset artifacts. Then, we present our analysis and results on dataset artifacts followed by discussions and a conclusion.

2 Implementation Approaches

In this paper, we employ two approaches to evaluating GPT-3: few-shot learning and fine-tuning. With few-shots, GPT-3 learns from given examples with respect to the specified format. In fine-tuning, we further train the model parameters for a specific task. We test the performance and robustness of GPT-3 on question answering (QA) and natural language inference (NLI) tasks. To better understand the characteristics of GPT-3 in both few-shot learning and fine-tuning, we test the general capabilities of the model against multiple datasets for each task.

To analyze the impact of dataset artifacts, we tested the model against adversarially generated datasets to test its robustness against spurious correlations or statistical irregularities. We then trained the models on the adversarial sets to evaluate whether these challenge sets can mitigate dataset artifacts found in standard datasets.

Since there's no available baseline performance for most of the downstream tasks, we explored many different datasets to determine the viability and feasibility of each task for GPT-3.

Once we had some understanding of GPT-3 baseline performance, we carried out a series of experiments to determine which dataset would be the most suitable for analyzing the impact of dataset artifacts in the model.

Table 1: SQuAD 2.0 (dev) Benchmark

Models	EM	F1
GPT-3 (Zero-shot, 175B)	52.6	59.5
GPT-3 (One-shot, 175B)	60.1	65.4
GPT-3 (Few-shot, 175B)	64.9	69.8
BERT	80.005	83.061
XLNet	87.926	90.689
RoBERTa	86.820	89.795
DeBERTalarge	88.0	90.7

Exact match (EM) and F1 scores for GPT-3 and SoTA models.

Table 2: ANLI Benchmark

Models	R1	R2	R3
GPT-3 (Zero-shot, 175B)	34.6	35.4	34.5
GPT-3 (One-shot, 175B)	32.0	33.9	35.1
GPT-3 (Few-shot, 175B)	36.8	34.0	40.2
RoBERTa (Large)	72.4	49.8	44.4
XLNet (Large)	70.3	50.9	49.4

GPT-3 and SoTA benchmarks for all rounds of ANLI in place of SNLI.

We chose question answering (QA) and natural language inference (NLI) as our main tasks to consider for analyzing dataset artifacts. QA is one of the well regarded and the most application tasks in NLP and NLI is one of the canonical tasks for natural language understanding. Since benchmarks for fine-tuned performance of GPT-3 are not available, we use few-shot performance as a proxy to reason about relative task difficulty.

Table 1 shows the comparison of the QA performance of GPT-3 against SoTA models. In Table 2, we compare the NLI performance of GPT-3 against SoTA models. The first three rows correspond to the performance of GPT-3 with different numbers of examples (zero-shot, one-shot and few-shot from the top, respectively). The GPT-3 results are directly from the GPT-3 paper (Brown et al., 2020) and all other results are from paperswithcode.com. The comparison of these cases indicates the impact of few-shot learning on the performance of GPT-3.

Furthermore, it is noteworthy that, although the accuracy is not as high as other SoTA models listed below the fourth item, the GPT-3 (few-shot) can show competitive performance even without fine-tuning on task specific dataset. On the other hand, Table 2 shows that GPT-3 (few-shot) struggles to do better than random guesses whereas the SoTA models perform substantially better. This suggests that the performance of GPT-3 can vary significantly based on the task.

2.1 Hyperparameters and Datasets

At the time of this writing, OpenAI offers 4 different tiers (or Engines) of GPT-3 that mainly differ in cost and performance ratio. More details about the model can be found in Engines - OpenAI API (https://beta.openai.com/docs/engines).

Based on our few-shot experiments, Ada engine was too weak for performing the few-shot question and answering and the pricing for Davinci engine simply felt too expensive. For fine-tuning experiments, we exclusively use the “Curie” engine to evaluate the general performance of OpenAI's GPT-3.

For hyperparameters, we tried to be consistent with the examples provided by OpenAI since we are mostly concerned with the datasets rather than the models. Likewise, all fine-tuning was done with the default hyperparameters and appropriate classification metrics configuration for the task.

Lastly, all datasets were sourced from HuggingFace Datasets library. https://huggingface.co/datasets.

2.2 Few-shot Learning

GPT-3 is well known for zero/one/few-shot learning. It can learn on the spot from zero to few examples and perform surprisingly well. On the other hand, it can also behave quite unpredictably at times due to the nature of text generation. For instance, it can hallucinate some information it learned during pre-training, which makes it difficult to trust whether the model is actually performing the task without bias. Also, it can sometimes decide not to follow the example at all and go on its merry way of composing examples without labels.

Question Answering

In this section, we consider one of the most well regarded NLP tasks, the question answering (QA) and explore the following datasets:

- SQuAD 1.1 (Rajpurkar et al., 2016)
- SQuAD 2.0 (Rajpurkar et al., 2018)
- Adversarial QA (AQA) (Bartolo et al., 2020)

These datasets are chosen for their popularity and the subtle differences to understand the general capabilities of GPT-3 in QA.

For implementation, we followed the approach described in the GPT-3 paper (Brown et al., 2020) and some of the examples available on the OpenAI API website (https://beta.openai.com/examples/). We started off with the prompt from OpenAI's QA example shown on the website. Then, we experimented with different formats such as question only prompt. Since Table 1 showed competency of GPT-3 on the QA task, we started with zero-shot learning first.

In zero-shot learning, GPT-3 was not able to produce any meaningful results for any of the datasets. Even with a well engineered prompt, the model failed (or refused) to generate a valid text. In few-shot learning, GPT-3 performed well in many different scenarios. While it was able to generate correct answers even without the context, we noticed that the generated text varied significantly from the gold label in most cases. This meant that the exact match (EM) metric was going to be an unreliable signal in analyzing dataset artifacts.

Furthermore, the model failed to generate text when we structured the prompt in a way that can be used to compare the standard (SQuAD) and adversarial datasets (AQA). In order for us to use adversarial examples against the standard test set, we need to structure the prompt with multiple contexts rather than a single context with multiple questions. The change in prompt structure rendered the model incapable of generating a valid text. Due to aforementioned reasons, we decided to focus on the other task, natural language inference (NLI).

Natural Language Inference

In this section, we consider one of the hardest NLP tasks for GPT-3, the natural language inference (NLI) where we explore the following datasets:

- SNLI (Bowman et al., 2015)
- ANLI (Nie et al., 2020)

Stanford Natural Language Inference (SNLI) dataset is one of the most well known NLI dataset and Adversarial NLI (ANLI) is an improvement on top of SNLI with increased difficulty and complexity.

For implementation, we drew K examples from the training set as conditioning, delimited by 1 or 2 newlines depending on the task as outlined in the GPT-3 paper (Brown et al., 2020). We also explored different prompt engineering techniques similar to question answering and OpenAI tweet sentiment classifier (https://beta.openai.com/examples/default-adv-tweet-classifier.).

Only with the few-shot learning, we were not able to make the model to generate a valid text (one of Entailment, Neutral, or Contradiction) regardless of how many examples we threw at it. This further motivated us to employ fine-tuning to lay a solid foundation for investigating dataset artifacts.

2.3 Fine-tuning

On the other hand, fine-tuning GPT-3 allowed GPT-3 to perform the tasks where the few-shot learning had failed. With fine-tuning, we were able to verify the feasibility of using GPT-3 on the SNLI dataset. We found that the default number of four epochs was more than enough to consistently reach the training token accuracy of 100. However, some degree of prompt engineering such as adding the newline end token was required to reliably get a high score on training sequence accuracy.

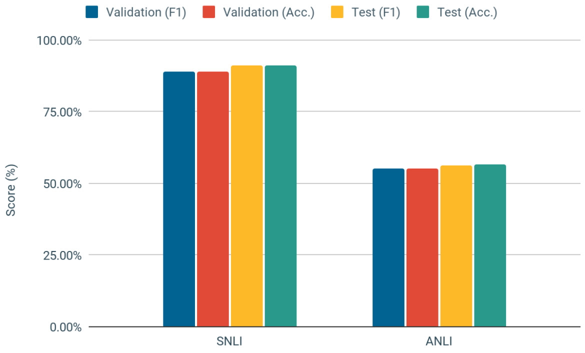

Figure 1: F1 and accuracy for SNLI and ANLI

With SNLI, we fine-tuned only 10% of the training dataset (55K) due to the imposed token limit. As shown in Figure 1, the reported validation F1 score was 88.974. The F1 score and accuracy for the test set was 91.33 and 91.30 respectively. With ANLI, we fine-tuned only on the Round 1 dataset in order to stay under the token limit. As shown in Figure 1, the reported validation F1 score was 55.0766 and test F1 score was 56.398, which is considerably higher than the published GPT-3 (few-shot) round 1 F1 score of 36.8.

2.4 Dataset Artifacts

With the fine-tuned model as a baseline, we were able to replicate the dataset artifacts mentioned by Gururangan et al., Nie et al., and Poliak et al. We experimented with the common strategy of removing gender or number information, introducing negation, and adding a purpose clause in the hypothesis (Gururangan et al., 2018). We also took an adversarial approach to replacing certain words to fool the model to generate a wrong label (Nie et al., 2020). All of these approaches to analyze the dataset artifacts were successful to a certain degree, but the results varied significantly between datasets and the manual process of trial-error by hand was quite exhausting.

To further analyze the impact of dataset artifacts in GPT-3, we fine-tuned additional models on SNLI and SNLI+ANLI training sets. Then, we evaluated each model on SNLI-only, ANLI-only, and SNLI+ANLI test sets to analyze and compare model performance against the adversarial set.

For the SNLI+ANLI model, we fine-tuned GPT-3 on both SNLI and ANLI (Round 1). We used the entire ANLI (Round 1) training and downsample (25K) SNLI training set due to the token limit. For the SNLI-only model, we fine-tuned GPT-3 on only the SNLI but with a larger training set to get the same training set size (41,946). In both models, the training set for the SNLI portion were identical and were shuffled beforehand to ensure a random sample distribution. Same hyperparameters were used and same prompt structures were used to isolate any possible external variables.

For ANLI, only Round 1 was used. For SNLI, a downsampled (1K) test set was used to match the number of rows in the ANLI (Round 1) test set.

3 Results and Analysis

With few-shot learning only, we were unable to get any meaningful results due to several issues in both QA and NLI. In QA, few-shot performance looked promising at first with the example prompt from OpenAI. However, the performance quickly degraded with a different prompt structure, which was required in order to analyze the dataset artifacts. In NLI, GPT-3 failed to perform on few-shot learning despite our best attempts to make it work by trying many different prompt engineering techniques and best practices. Thus, the analysis of datasets artifacts only on few-shot GPT-3 is inconclusive.

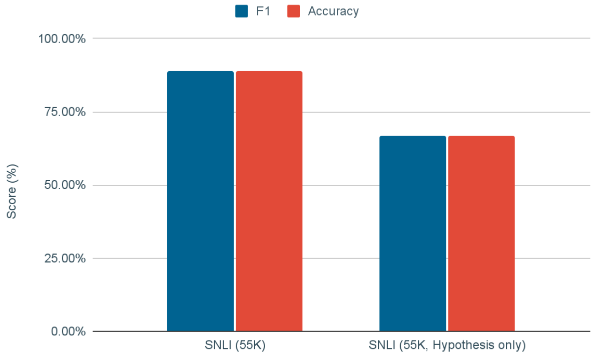

Figure 2: F1 and accuracy for hypothesis only SNLI

As shown in Figure 2, our result from the hypothesis-only experiment suggests that statistical irregularities and dataset artifacts in SNLI seem to allow GPT-3 to achieve suspiciously high performance without having access to the premise. Furthermore, we observed numerous instances of annotation artifacts such as lexical choice and negation that were presented by Gururangan et al. In Table 3, the first row demonstrates one of many instances of entailed hypothesis with a generic word “animal”. The second row demonstrates an instance of simple negation “not” with a contradicted hypothesis.

Table 3: Annotation Artifacts in SNLI

Premise: A white duck expanding its wings in the water.
Hypothesis: There is one animal in this picture.
Label: Entailment
SNLI-only: Entailment
SNLI+ANLI: Entailment

Premise: Two men working on the roof of an apartment building with a nice looking skyline behind them.
Hypothesis: Two men not working on the roof of an apartment building
Label: Contradiction
SNLI-only: Contradiction
SNLI+ANLI: Contradiction

Table 4: F1 score on SNLI or ANLI test set

Training Data	SNLI (1K)	ANLI (1K)
SNLI (55K)	91.330	36.317
ANLI (17K)	70.006	56.398
SNLI (42K)	89.350	33.502
SNLI (25K) + ANLI (17K)	88.926	58.647

The left rows are the training sets that were used for fine-tuning and the top columns are the test sets that were used to evaluate the fine-tuned models.

With fine-tuning the few-shot learning, we found that augmenting the adversarial datasets did improve the test score on the adversarial set, but did not improve the test score on the standard set. However, the test score on the standard set only dropped by 0.42 whereas the test score on the adversarial set increased by 25.15 [Table 4]. This trade off seemed favorable considering that the SNLI-only model was not doing much better than random guesses.

In order to further investigate the results, we took the first 100 examples which were categorized incorrectly with SNLI but were improved by ANLI [Figure 3]. Especially, ANLI contributed to the improvement of lexical inference and number reasoning cases. Among these 100 examples selected for the analysis, 13 cases were related to lexical inference and 21 cases involved numerical reasoning.

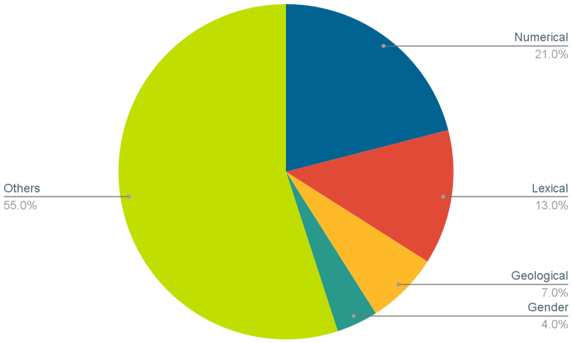

Figure 3: Categorization of improved examples with ANLI

Table 5 shows a few examples where fine-tuning with ANLI have resulted in improved robustness over the SNLI-only model. The first row demonstrates an instance of lexical inference for generic words and the second row demonstrates an instance of numerical reasoning. In the lexical inference example, SNLI-only fails to recognize that “individuals” and “men” are synonymous. In the numerical reasoning example, SNLI-only fails to recognize that “twenty-four unarmed Union soldiers” contradicts “25 Union soldiers”. Geological inference and gender inference are also detected as a part of the improvement.

Table 5: Improvements of SNLI+ANLI over SNLI-only

Premise: Two individuals dressed up like animals are posing for the camera. **Hypothesis:** Two men dressed as basketball players are running. **Label:** Contradiction **SNLI-only:** Neutral **SNLI+ANLI:** Contradiction
Premise: The Centralia Massacre was an incident during the American Civil War in which twenty-four unarmed Union soldiers were captured and executed at Centralia, Missouri on September 27, 1864 by the pro-Confederate guerrilla leader William T. Anderson. Future outlaw Jesse James was among the guerrillas. **Hypothesis:** The Centralia Massacre was the execution of 25 Union soldiers during the American Civil War. **Label:** Contradiction **SNLI-only:** Entailment **SNLI+ANLI:** Contradiction

Apart from performance of ANLI (17K) on SNLI, the benchmark results matched our expectations [Table 4]. We expected ANLI (17K) to perform relatively well on the SNLI test set since the ANLI training set was an improvement over the SNLI training set. While the lower number of rows in ANLI training set seemed to be the likely cause at first, the relatively high performance on the ANLI test set suggested that the GPT-3 model was still picking up on certain dataset artifacts in ANLI.

Out of all of the fine-tuned models, SNLI (55K) performed the best on the SNLI test set due to having a relatively large training set [Table 4]. Both SNLI (55K) and SNLI (42K) performed close to chance on the ANLI test set due to the difficulty of the challenge set. This suggests that the model relied heavily on dataset artifacts in SNLI rather than learning the task via NLU.

4 Discussion

4.1 Few-shot Learning

For QA, the prompt engineering had a notable impact on the text generation results. We found that adding more examples didn't necessarily help with the text generation. On the contrary, naively selecting examples introduced the possibility of the model cheating based on one of the examples the example question was similar to the query question. The presence of similar questions in many of the QA dataset posed an additional difficulty in accurately evaluating few-shot performance of GPT-3.

The NLI task requires re-reading or carefully considering a long passage and then generating a very short answer (Brown et al., 2020). This was something GPT-3 is quite poor due to lack of bidirectionality. Despite the inherent difficulty of the NLI task for GPT-3, it seems quite problematic that the model could not generate any relevant text.

4.2 Fine-tuning

Certain aspects of a prompt such as tasks descriptions and examples weren't necessary, but having a basic prompt structure remained crucial in making the model perform. More specifically, adding in an end token such as a newline ("\n") or triple hashtags ("###") was necessary for the text generation to stop after the label was generated. Otherwise, the model continued generating non-label text such as the "Premise:" and "Hypothesis:".

4.3 Dataset Artifacts

Despite our best efforts to isolate external variables in our experiments, we suspect that some of the results are biased due to unequal distribution of the number of tokens in the hypothesis between SNLI and ANLI. Sentence length is one of the common annotation artifacts (Gururangan et al., 2018) and premises in ANLI dataset had a consistently higher number of tokens per each dataset. This may explain why the ANLI (17K) model performed significantly worse on the SNLI test set despite ANLI training set being more "data-efficient" (Nie et al., 2020).

5 Conclusion

In this paper, we leveraged OpenAI's GPT-3 to investigate dataset artifacts in one of the largest NLP models currently available to the public. First, we started experiments with the few-shot performance of GPT-3 and the effects of prompt engineering in regards to dataset artifacts. To have a solid foundation for the performance analysis regarding text interaction and generation, we decided to analyze dataset artifacts in fine-tuned GPT-3 models. Our results suggest that GPT-3 is vulnerable to common dataset artifacts despite its large number of parameters and ability to generalize better. We found that fine-tuning GPT-3 with adversarial dataset in addition to the standard dataset helps with mitigating dataset artifacts and improving robustness of the model.

References

Bartolo, Max, A Roberts, Johannes Welbl, Sebastian Riedel and Pontus Stenetorp. "Beat the AI: Investigating Adversarial Human Annotation for Reading Comprehension." Transactions of the Association for Computational Linguistics 8 (2020): 662-678.

Bowman, Samuel R., Gabor Angeli, Christopher Potts and Christopher D. Manning. "A large annotated corpus for learning natural language inference." *EMNLP* (2015).

Brown, Tom B., 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, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff 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. "Language Models are Few-Shot Learners." *ArXiv* abs/2005.14165 (2020): n. Pag.

Gururangan, Suchin, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman and Noah A. Smith. "Annotation Artifacts in Natural Language Inference Data." *NAACL* (2018).

Nie, Yixin, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston and Douwe Kiela. "Adversarial NLI: A New Benchmark for Natural Language Understanding." *ArXiv* abs/1910.14599 (2020): n. Pag.

OpenAI. "OpenAI's API Now Available with No Waitlist." OpenAI. OpenAI, November 18, 2021. https://openai.com/blog/api-no-waitlist/.

Poliak, Adam, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger and Benjamin Van Durme. "Hypothesis Only Baselines in Natural Language Inference." *SEMEVAL (2018).

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev and Percy Liang. "SQuAD: 100,000+ Questions for Machine Comprehension of Text." *EMNLP* (2016).

Rajpurkar, Pranav, Robin Jia and Percy Liang. "Know What You Don't Know: Unanswerable Questions for SQuAD." *ACL* (2018).

Roberts, Adam, Colin Raffel and Noam M. Shazeer. "How Much Knowledge Can You Pack into the Parameters of a Language Model?" *ArXiv* abs/2002.08910 (2020): n. Pag.

Turner, Elliot. "Reading the Openai GPT-3 Paper. Impressive Performance on Many Few-Shot Language Tasks. the Cost to Train This 175 Billion Parameter Language Model Appears to Be Staggering: Nearly $12 Million Dollars in Compute Based on Public Cloud GPU/TPU Cost Models (200x the Price of GPT-2) PIC.TWITTER.COM/5ZTR4CMM3L." Twitter. Twitter, May 29, 2020. https://twitter.com/eturner303/status/1266264358771757057.