GPT-Neo for commonsense reasoning-a theoretical and practical lens

ROHAN KASHYAP* and VIVEK KASHYAP*, Bangalore, India

Recent work has demonstrated substantial gains on pre-training large-scale unidirectional language models such as GPT-2, GPT-3 and GPT-neo followed by fine-tuning on a downstream task. In this paper, we evaluate the performance of the GPT-neo 1.3 billion model for commonsense reasoning tasks. We assess the model performance on 6 of the commonsense reasoning benchmark tasks and report the accuracy scores for these tasks. When fine-tuned using the right set of hyperparameters, we obtain competitive scores on 3 out of the 6 tasks, but struggles when the dataset size is significantly smaller. Our evaluation of the GPT-neo model shows that the low model performance on a few of these tasks is mainly because of the inherent difficulty in these datasets and since it fails to establish coherent patterns given its limited training samples.

We also investigate and substantiate our results using visualization and conduct numerous inference tests to better understand the model performance. Finally, we conduct thorough robustness tests using various methods to gauge the model performance under numerous settings. These findings suggest a promising path for exploring the results of much smaller language models than the GPT-3 175B model to perform tasks requiring natural language understanding.

CCS Concepts: • **Neural Networks → Language models**; • **Commonsense Reasoning**;

Additional Key Words and Phrases: Inteligence, Generalization

ACM Reference Format:
ROHAN KASHYAP and VIVEK KASHYAP. 2021. GPT-Neo for commonsense reasoning-a theoretical and practical lens. *ACM Trans. Graph.* 37, 4, Article 111 (August 2021), 6 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Commonsense reasoning encompasses a central role in neural language understanding. Recent years has seen the rise of large unidirectional language models (uses casual attention from left to right) like the GPT-2, GPT-3, and GPT-neo. Recent advances in natural language processing demonstrate the effectiveness of these huge pre-trained models on numerous downstream tasks, and also methods such as instructive fine-tuning methods have shown excellent few-shot learning capabilities. The parametric count of these models typically ranges from a few million to billions, the performance of whose follow strict scaling laws with the larger models being more sample efficient and are only mildly are prone to overfitting. Generalization power is the ability to mine previous experience to make sense of future novel situations. It describes a knowledge differential and characterizes the ratio between the known information (prior) and the space of possible future situations. Thus, it can be thought of as our ability to deal with novelty and uncertainty and acquire new skills through learning.

In this paper, we investigate the generalization capabilities of the GPT-neo 1.3 billion parameter model through the commonsense reasoning lens. We in particular, employ the supervised learning objective discussion throughout this paper and test the model on a suite of 6 tasks, namely hella swag, story cloze, cose, winograde, cosmos, and piqa.

The transformer architecture with causal self-attention mechanism has gained immense success in generative modeling tasks and when tuned adequately to the right temperature produce coherent and plausible text. Memory constitutes a big component of intelligence along with the model's ability to extrapolate and draw useful information from the training data to gain knowledge about the world. These huge models can easily interpolate between the convex hull of different words and capture the compositional relationships between them. The statistical patterns implicitly present in the pre-trained text corpus is known to provide the model with a few innate priors and experience and is a possible explanation for its performance on few-shot learning methods.However,these conclusions are well beyond the scope of this paper and thus remains a field of research to the deep learning community to explore and thus explain the reason for such performances.

Generalization power is the sensitivity to abstract analogies and the ability to mine previous experience to make sense of future novel situations. Intelligence thus requires that you adapt to novelty and convert past experience into future skills. Neural networks are long known for their ability to perform well as function approximators (excellent performance) on smooth continuous data manifolds without any discontinuities using gradient descent methods. Thus, this paper is an attempt in this direction to explore at least a few of these intuitive questions by resorting to the commonsense reasoning benchmark as a standard of measure in all our experiments.

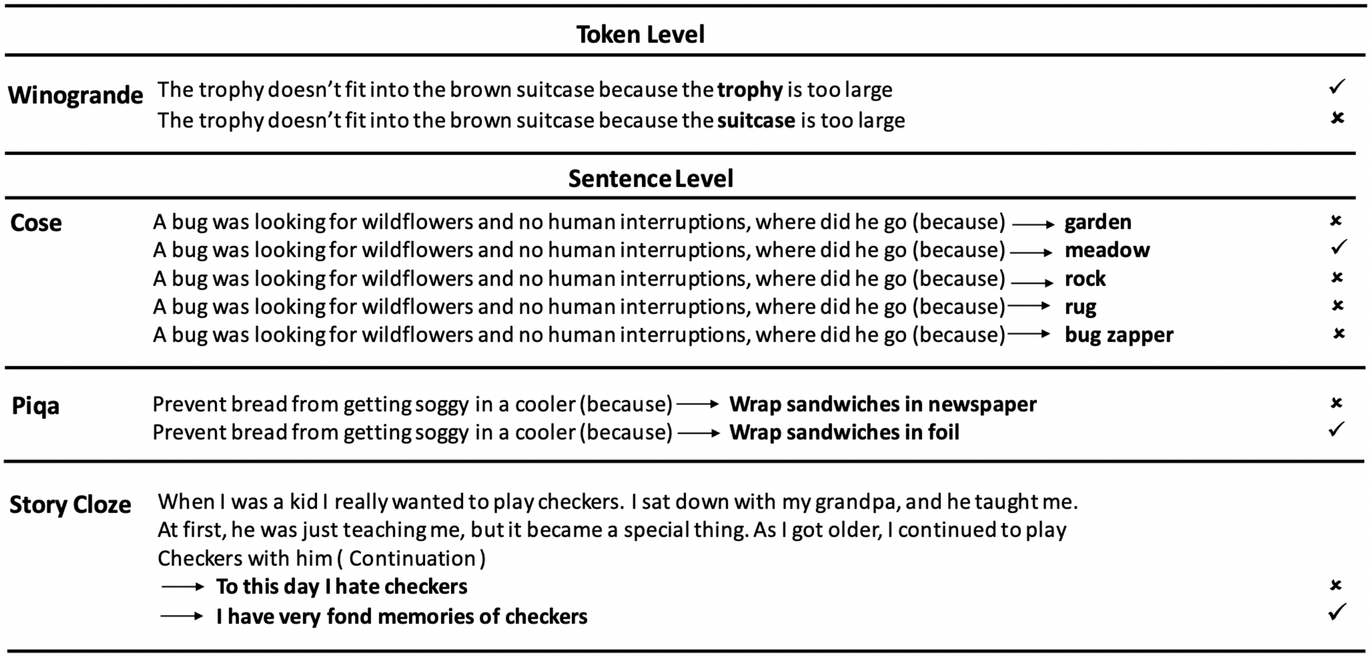
We demonstrate the effectiveness of our approach on a wide range of benchmarks for natural language understanding. The results also prove that the model leverages the learned word representations during unsupervised pre-training to significantly boost performance compared to the randomly initialized model. Their strong performance is dependent on the ability of these large models to quickly adapt to each of these tasks with just a few epochs of fine-tuning process.

One way to test GPT-neo's ability in the fine-tuning phase with limited dataset constraints is to give it simple commonsense reasoning tasks which requires recognizing a novel pattern that is unlikely to have occurred during pre-training and thus adapt quickly to the given task. We hope that the results presented in this paper will motivate further research along these lines and to better understand the model behavior under different task settings.

[*] Both authors contributed equally to this research.

Authors' address: ROHAN KASHYAP, rohanvk13@gmail.com; VIVEK KASHYAP, vivekvkashyap10@gmail.com, Bangalore, Bangalore, Bangalore, Karnataka, India, 560054.

0730-0301/2021/8-ART111 $15.00
https://doi.org/10.1145/1122445.1122456



	Token Level	
Winogrande	The trophy doesn't fit into the brown suitcase because the **trophy** is too large	✓
	The trophy doesn't fit into the brown suitcase because the **suitcase** is too large	✗
	Sentence Level	
Cose	A bug was looking for wildflowers and no human interruptions, where did he go (because) ⟶ **garden**	✗
	A bug was looking for wildflowers and no human interruptions, where did he go (because) ⟶ **meadow**	✓
	A bug was looking for wildflowers and no human interruptions, where did he go (because) ⟶ **rock**	✗
	A bug was looking for wildflowers and no human interruptions, where did he go (because) ⟶ **rug**	✗
	A bug was looking for wildflowers and no human interruptions, where did he go (because) ⟶ **bug zapper**	✗
Piqa	Prevent bread from getting soggy in a cooler (because) ⟶ **Wrap sandwiches in newspaper**	✗
	Prevent bread from getting soggy in a cooler (because) ⟶ **Wrap sandwiches in foil**	✓
Story Cloze	When I was a kid I really wanted to play checkers. I sat down with my grandpa, and he taught me. At first, he was just teaching me, but it became a special thing. As I got older, I continued to play Checkers with him (Continuation)	
	⟶ **To this day I hate checkers**	✗
	⟶ **I have very fond memories of checkers**	✓

Fig. 1. An example is specified for each of the tasks used for training the GPT-neo model with its answer choices indicated by ✓

2 DATASET

Commonsense tasks provide a way to understand the model's capability to interpret complex structural patterns that are not explicitly mentioned in the input text and are not part of the pre-trained data. Commonsense reasoning ability of large pre-trained models largely depends on the dataset quality and requires high levels of abstract reasoning capabilities to solve these tasks. These tasks include sentence completion, story generation, wherein all these tasks involve choosing one correct completion from several options(multiple choice); we enter the input by concatenating context and question for every choice possible. We concatenate the "context+question+choice" for each example to obtain k different such sentences (k choices) and pass it as input to the model. We compute the probability of each choice and compute the softmax scores from their logit values.

For our experiments, we consider a total of 6 diverse standard commonsense reasoning benchmark datasets and examine how well the model adapts to each of these tasks individually. It is often seen that large-scale language models suffer significant degradation in their performance specifically on these tasks since it requires the model to draw implicit, hidden correlations present within the text sentences and extrapolate beyond the given training examples, i.e., understanding what is not stated yet is obviously true.

For example, in the sentence," The corner table in a restaurant ordered a beer? The waiter served them with the drink".Humans can easily establish that it is the people at the corner table that ordered the beer and "them" refers to the people and not the corner table because we as humans have the hard-coded notion of what people is, but language models cannot quickly capture such implicit information or knowledge and hence fall back significantly on these tasks. These tasks, although quite trivial for humans, are extremely hard for machines that merely rely on statistical patterns without true understanding and abstraction capabilities. In this paper, we show that popular approaches to large-scale language pre-training, while highly successful on many abstract tasks, it fall short when a physical model of the world is required.

2.1 Winogrande

Winogrande[4] is inspired by the "Winograd Schema Challenge" with increased task complexity. It consists of predicting the word or participant to which noun in the sentence it refers to. Including train and validation, it consists of 44k examples.

Example:" Robert woke up at 9:00 am while Samuel woke up at 6:00 am, so he had less time to get ready for school". To predict what "he" actually refers to in the given context,it requires us to know the time he woke up. Thus, the model needs to understand every little detail in the example.

2.2 Cose

Cose[5] is a multiple-choice question answering dataset proposed for evaluating natural language processing(NLP) models for commonsense reasoning capabilities. Cose helps to predict the right choice of reasoning for the human actions performed.

It consists of 5 answer choices and a total of 10k sentences, including both train and validation sets. One example is:" A beaver is taking logs from a Pacific beach, where is it located?" and the model is expected to output the answer as Washington.

2.3 Piqa

PIQA[6] (Physical Interaction Question Answering) dataset is used to evaluate language representations and their knowledge about our physical world. This dataset describes how an object is built, used, or operated, which requires physical reasoning capabilities to select the right choice.

In total, it consists of over 16,000 training QA pairs with an additional 2K and 3k held out for validating and testing, respectively. One example is:" Make an outdoor pillow" Choice1:Blow into a tin can and tie with rubber band Choice2:Blow into a trash bag and tie with a rubber band.

2.4 StoryCloze

Story Cloze[7] dataset involves selecting a plausible ending to a long story which is framed as a multiple- choice question with two answer choices. It consists of 4k examples, which are based on everyday events of people (daily life). It helps evaluate the extent to which the model learns causal and temporal relations between different entities within the given context.

2.5 HellaSwag

HellaSwag[8] is a commonsense natural language inference task with its data format similar to SWAG but of higher data quality. Hellaswag dataset involves picking the best ending for a story.

For each question, a model is given a context from a caption and four choices to complete the sentence and predict what might happen next. HellaSwag contains a total of 50k sentences with an average length of 230-word tokens.

2.6 CosmosQA

CosmosQA[9] dataset is a reading comprehension task that requires the model to infer and understand correlations that are not explicitly mentioned and requires "reading between the lines" for commonsense inference. It consists of 35,600 examples with four answer choices.

It requires contextual commonsense reasoning to under- stand people's narratives and references in the reading comprehension to deduce logical connections between attributes, such as the cause (why?) and its effects(How?).

3 TRAINING PROCEDURE

We evaluate the performance of GPT-Neo(1.3B parameters) pretrained model, on a range of small and mid-sized downstream classification tasks. We are primarily interested in two primary quantities:(1) Accuracy of the downstream task (2) Cross-entropy loss, which is important to account for the model convergence.

Our main goal is not to demonstrate state-of-the-art results, but to show that, GPT-Neo model is competitive with its counterpart such as GPT-3[10] even though it has a much smaller parameter count by a factor of 134x. We also report scores from bidirectional language models such as BERT-large and GPT-3 few-shot results and obtain comparable results. For model quality, we report only top-1 model accuracy on downstream tasks.

4 MODEL

The GPT-neo model was released by Eleuther AI (open source model). The GPT-neo model is a transformer-based decoder-only autoregressive language model that uses a causal self-attention mechanism to learn contextual representations of individual word tokens.This provides the model a more structured memory for handling long-term dependencies compared to alternatives like recurrent networks, resulting in robust transfer performance across diverse tasks.

We train a 24-layer decoder-only transformer model, wherein it performs the masked multi-head attention (768-dimensional states and 16 attention heads) operation over the input context tokens followed by position-wise feed-forward layers (2048 dimensional inner states) to produce an output distribution over target tokens. The GPT-Neo was pre-trained using the standard language modeling objective to predict the next token using a causal attention mask, and this helps the model gain knowledge about the world and language constructs and learn useful coherent patterns in the training data. The objective is to maximize the likelihood function given below:

$$L_1(\mathcal{U}) = \sum_i \log P(u_i|u_{i-k}, \ldots, u_{i-1}; \Theta)$$

where k is the size of the context window, and the conditional probability P is modeled using the unidirectional GPT-Neo model parameters.These parameters are trained using stochastic gradient descent. We ensure that there is no overlap between the dataset used during the train and inference time.

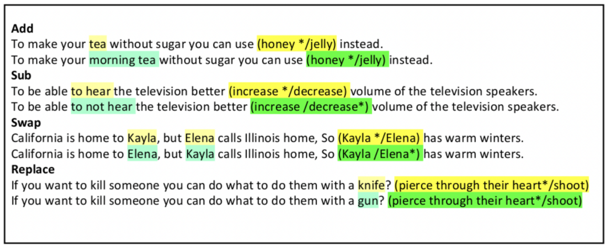
We adapt the model parameters to the supervised target task. First, the input tokens are passed through the model to obtain the final transformer block's activation, i.e., the contextual representation is then fed to a dense classifier to obtain the model predictions. The structure of the GPT-neo model is very much similar to the GPT-3 model except that it was pre-trained on the Pile dataset. The Pile consists of a diverse collection of 22 high-quality datasets derived from numerous sources and shows a significant performance boost compared to the GPT-3 models pre-trained on the common crawl dataset. It includes several datasets such as Books3, Pile-CC, DM-Mathematics, OpenWebText2, English Wikipedia and Arxiv and shows significant gains when fine-tuned on downstream tasks.

For our experiments, we utilize the GPT-Neo 1.3B parameter version model for adequate comparison to the BERT-large model [11] in terms of parameter count.

5 FINE-TUNING DETAILS

We perform all our fine-tuning experiments using the Adafactor, batch size of 16, which accounts for a batch size of 2 per device, linear learning rate schedule with a linear warmup, and we do not use gradient clipping. Following the common practice, we use the JAX framework for model training and evaluation and obtain significant training speed boost-up using Google Cloud TPU-VM's. We train the models for 3 epochs and use a fixed random seed throughout our experiments.

We also add dropout with a rate of 0.2 for regularization. We use



Add
To make your tea without sugar you can use (honey */jelly) instead.
To make your morning tea without sugar you can use (honey */jelly) instead.
Sub
To be able to hear the television better (increase */decrease) volume of the television speakers.
To be able to not hear the television better (increase /decrease*) volume of the television speakers.
Swap
California is home to Kayla, but Elena calls Illinois home, So (Kayla */Elena) has warm winters.
California is home to Elena, but Kayla calls Illinois home, So (Kayla /Elena*) has warm winters.
Replace
If you want to kill someone you can do what to do them with a knife? (pierce through their heart*/shoot)
If you want to kill someone you can do what to do them with a gun? (pierce through their heart*/shoot)

Fig. 2. Samples used for measuring the robustness of the GPT-neo models using various methods such as addition,subtraction,swapping and replacing.We use 2 sentences here to indicate the positive and negative samples and the model predicts accurately for the positive sentence and correct or incorrect predictions for the negative sentence.The * indicates the predicted answer choice.

the bytepair encoding (BPE) tokenization scheme with a vocab size of 50257 tokens.For all our tasks,we use a learning rate of 2e-5 to 3e-7 and a batch size of 16. We use Adafactor optimization scheme with a maximum learning rate of 2e-5. We slightly deviate from the standard approach used in the GPT-2 paper, wherein we use the maximum length of the input tokens for padding the input sequences into batches and not 512-this seems not to harm the model performance by any means and also results in a reduced memory footprint on the TPU accelerators. Our model finetunes quickly, and 3 epochs of training were sufficient for most cases. We use a linear learning rate decay schedule with warmup over 0.1% of training, and the value annealed 3e-7. We perform all our model fine-tuning using Google Cloud TPU- VM's using the TPU-V8 accelerators. We also experiment with numerous batch sizes ranging from 4 to 32 (a batch size of 32 results in OOM) and conclude that the 16 version outperforms the other values significantly.

6 RESULTS

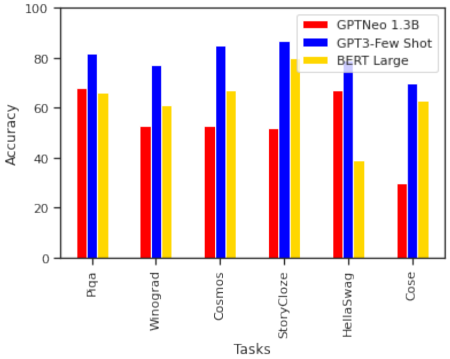
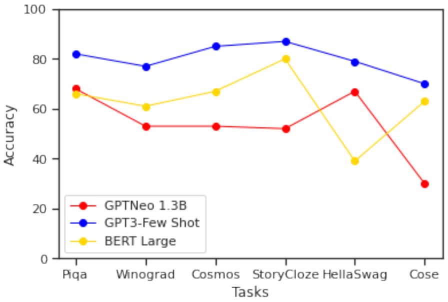


Fig. 3. The model performance for each of the tasks,trained for the same number of steps. The GPT-neo model obtains excellent results on the piqa, winogrande and hellaswag datasets while falls back on the other tasks by a considerable amount.

In this section, we test GPT-neo's performance on a suite of commonsense reasoning benchmark tasks that involve text or sentence completion, natural language inference, or selection preference. Figure 4 compares the GPT-neo models with the other models on all of our commonsense reasoning tasks. The GPT-neo 1.3B model outperforms its GPT-3 1.5B parameter model counterpart by a significant accuracy score on hellaswag, piqa and winograde datasets with an increase by +7% points. We achieve our best results on the cosmos-qa dataset with an accuracy of 81.1%, and it compares favorably to the 85.1% accuracy of the GPT-3 few-shot(top-1 on cosmos-qa original set) while it is still 20% worse than the human performance, which is the current state-of-the-art. We also report the model scores for bert models for appropriate comparison in terms of parameter count with the GPT-neo, GPT-3 model, and it is seen that unidirectional models generally demonstrate better gains on commonsense reasoning tasks. These tasks require the information stored in the model's parameters to answer general knowledge questions, and the degree to which the pre-training stage has helped in the improved model performance is an open research problem that we seek to establish in our subsequent research work.



6.1 HellaSwag

GPT-neo achieves an accuracy score of 67.8% on the evaluation dataset. We obtain competitive results for this task while significantly outperforming the 39% accuracy of the fine-tuned bert-large model and 54.7% of the GPT-3 1.5 billion model, which is still lower than the SOTA results of the GPT-3 175B few-shot model and 85.6% achieved by the fine-tuned multi-task model ALUM.

6.2 Piqa

Here we achieve an accuracy of 68.2% upon fine-tuning, which is slightly lower than the 75% accuracy of the GPT-3 1.5B model while it shows a significant gap of 14% compared to the SOTA results of GPT-3 175B few-shot model with an accuracy 82%.

6.3 Cosmos-qa

We next evaluate our model on the cosmos-qa dataset and obtain an accuracy of 58% while the bert-large model achieves an accuracy of 67%.

6.4 Winogrande

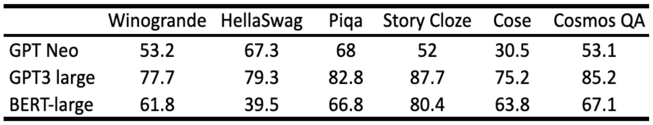
We evaluate the results of our model on the Winograde dataset and obtain an accuracy of 53%, which is equivalent to the results obtained for the GPT3 1.5B parameters. We also obtain near comparable results when compared to the 61% accuracy of the bert-large model while it is significantly lower than the 77.7% accuracy of the GPT3 few-shot model.

6.5 StoryCloze

We evaluate the results of our model on the StoryCloze 2017 dataset and achieve an accuracy of 81.1% on the train set while only 52.1% on the evaluation dataset. Shows 30.1% lower than the fine-tuned SOTA using a BERT-based model and GPT-3 model.

6.6 Cose

In the cose dataset, we get diminishing returns and obtain our lowest accuracy score of 30% on this challenging dataset benchmark. We mainly attribute this to the in- herent difficulty in the sentence structure, which requires inferring adequate correlations within the sentence and the lack of enough dataset examples in the training set. However, this is still 33% lower than the BERT-large model.



	Winogrande	HellaSwag	Piqa	Story Cloze	Cose	Cosmos QA
GPT Neo	53.2	67.3	68	52	30.5	53.1
GPT3 large	77.7	79.3	82.8	87.7	75.2	85.2
BERT-large	61.8	39.5	66.8	80.4	63.8	67.1

Fig. 4. Performance comparison of the GPT-neo,GPT-3 few-shot and BERT-large model for all the 6 tasks.The GPT-3 model outperforms by a significant margin,however the GPT-neo model is competitive in at least a few of these commonsense reasoning tasks.

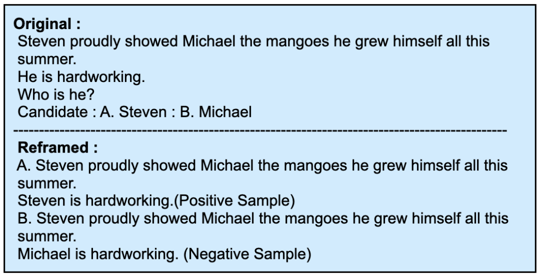
7 ROBUSTNESS TEST

To test GPT-neo's ability to learn novel symbolic manipulations, we perform the robustness test on a few examples. First, we follow the standard approach as studied in the [12] to evaluate the commonsense reasoning ability of the GPT- neo model. We test the model's capabilities for 4 main tasks, namely add, subtract, swap, and replace individual text tokens.

We are mainly interested in evaluating if our results are robust to the possibility that one of our assumptions might not be accurate. Each task involves giving the model the original sentence and the distorted one by some means of addition, or deletion, or replacing word tokens, and asking it to predict similar scores for both these sentences. We follow the conventional dual text instance method by constructing 2 similar sentences and thus inferring the robustness measure. We create 36 test samples with 6 examples from each of our commonsense reasoning datasets.

We do this by sampling adequate test sentences from the dataset and forming a pair of semantically equivalent sentences, testing for consistent results for all instances, and reporting the scores in figure3. Notice that in both these examples, we perform the robustness tests in context. The 4 tasks are as follows:

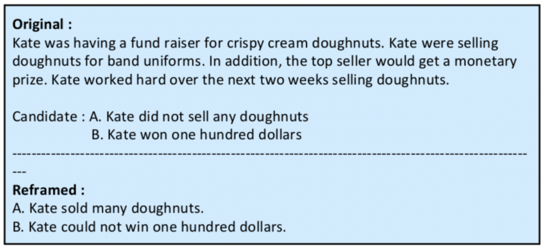
Addition:-The model is given 2 sentences with addition of words to its sentences, such that they are semantically equivalent and is required to output the same answer choice for both those sentences.Example:" good day->pleasant and a bright day."



Original :
Steven proudly showed Michael the mangoes he grew himself all this summer.
He is hardworking.
Who is he?
Candidate : A. Steven : B. Michael

Reframed :
A. Steven proudly showed Michael the mangoes he grew himself all this summer.
Steven is hardworking.(Positive Sample)
B. Steven proudly showed Michael the mangoes he grew himself all this summer.
Michael is hardworking. (Negative Sample)

Fig. 5



Original :
Kate was having a fund raiser for crispy cream doughnuts. Kate were selling doughnuts for band uniforms. In addition, the top seller would get a monetary prize. Kate worked hard over the next two weeks selling doughnuts.

Candidate : A. Kate did not sell any doughnuts
B. Kate won one hundred dollars

Reframed :
A. Kate sold many doughnuts.
B. Kate could not win one hundred dollars.

Fig. 6. Reframed sentences-we construct additional samples by modifying the original sentence or its answer choice which is then used to test the model during inference time.

Subtraction:-The model is given 2 sentences with its polarity reversed to negate it's semantic meaning wherein the sentences contradict each other and is expected to output adequate predictions.Example:" increase volume,decrease volume".

Swap:-The model is given 2 sentences with the role of each noun(name) swapped with one another and is expected to pick the right choices by contextual inference by identifying the roles of each noun respectively.Example:" Harry<->Hermione".

Replace:-The model is given 2 sentences by replacing its word tokens (one or more words), such that they are logically coherent in their meaning and should output the right choice depending on the context presented in the sentence.Example:" Knife->gun".

The piqa dataset performs consistently well on all the sentences for both positive and negative sentences with correct predictions for almost all text tokens when replaced with an equivalent word and achieves a score of 100%. The StoryCloze dataset also demonstrates competitive results and outputs excellent predictions when tested under different settings for each of the 4 tasks (exceptionally better off at replacing and swapping tasks). However, the Hellaswag sentences can perform well only half of the time, and the performance is somewhat degraded for the train sentences compared to its test sentences, and this behavior is quite unexpected among all the test suites of examples.

The performance on winograde, cosmos-qa, and cose is somewhat degraded with severe overfitting concerns and not being able adapt to the given context, suggesting that a significant portion of the correct answers is memorized. The inspection of incorrect answers reveals that the model often makes mistakes such as not being able to capture the contextual meanings and seems invariant at times to such changes. Overall, GPT-neo displays reasonable proficiency

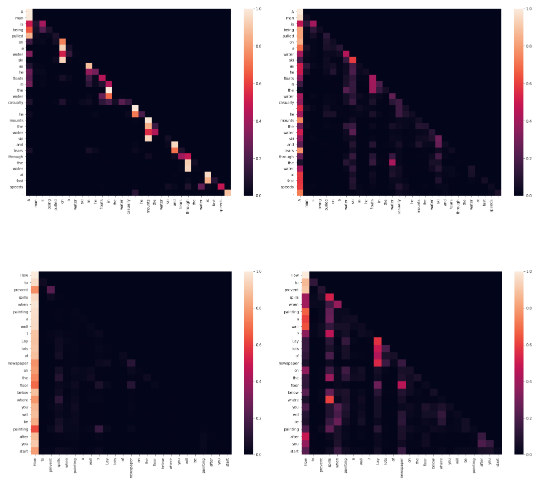


Fig. 7. Attention plots indicating the token-wise attention values for a given sentence.The plots include samples from the top,middle,bottom layers.

when probed at complex robustness zero-shot settings for at least a few of these tasks.

8 VISUALIZATION

In this section, we test the attention-head view using a heatmap that visualizes the attention of individual attention heads for each of its layers and report a few examples here (refer to figure 7). We are mainly interested in understanding the token dependencies of each token in the input sentence and their contextual representations across all its layers. It is commonly observed that the lower layers of transformer models capture global token dependencies (long-term dependencies), which enables it to capture and exchange the required knowledge across all its tokens. In comparison, the higher layers are highly specialized heads-a diagonal matrix that is very good at local and short-term dependencies of adjacent tokens. Deeper layers have no clear identifiable structure, and the structure is mainly invariant to the choice of specific hyperparameters. Our results align almost entirely with this standard, and we perceive these to be reasonable indications of the model's ability to adapt its weights and, hence, representations to its numerous tasks on the fine-tuning phase.

9 CONCLUSION

We illustrate the commonsense reasoning capabilities of the GPT-neo 1.3 billion parameter model on a suite of 6 diverse tasks. We demonstrate the power of constructing contextualized representations and language understanding of pre-trained language models, particularly the GPT-neo model.

We discussed the commonsense reasoning ability of GPT-neo in attribution to its commonsense knowledge in its pre-training data and its fine-tuning phase. Its ability to perform competitively with GPT-3 175B zero-shot model on atleast a few of these tasks and even exceeds the GPT-3(1.5B model) on several tasks seems to be a promising and exciting area of research. However, the behavior of these language models is still yet to be fully understood and an open research question.

10 ACKNOWLEDGMENTS

We thank the anonymous reviewers for their review and suggestions.

REFERENCES

[1] Gao, Leo and Biderman, Stella and Black, Sid and Golding, Laurence and Hoppe, Travis and Foster, Charles and Phang, Jason and He, Horace and Thite, Anish and Nabeshima, Noa and others, "The Pile: An 800GB Dataset of Diverse Text for Language Modeling,"arXiv preprint arXiv:2101.00027,2020.
[2] Ashish Vaswani and Noam Shazeer and Niki Parmar and Jakob Uszkoreit and Llion Jones and Aidan N. Gomez and Lukasz Kaiser and Illia Polosukhin,"Attention Is All You Need,"arXiv:1706.03762,2017.
[3] François Chollet,"On the Measure of Intelligence,"arXiv:1911.01547,2019.
[4] Keisuke Sakaguchi and Ronan Le Bras and Chandra Bhagavatula and Yejin Choi,"WinoGrande: An Adversarial Winograd Schema Challenge at Scale,"arXiv:1907.10641,2019.
[5] Nazneen Fatema Rajani and Bryan McCann and Caiming Xiong and Richard Socher,"Explain Yourself! Leveraging Language Models for Commonsense Reasoning,"arXiv:1906.02361,2019.
[6] Yonatan Bisk and Rowan Zellers and Ronan Le Bras and Jianfeng Gao and Yejin Choi,"PIQA: Reasoning about Physical Commonsense in Natural Language,"arXiv:1911.11641,2019.
[7] Nasrin Mostafazadeh and Nathanael Chambers and Xiaodong He and Devi Parikh and Dhruv Batra and Lucy Vanderwende and Pushmeet Kohli and James Allen,"A Corpus and Evaluation Framework for Deeper Understanding of Commonsense Stories,"arXiv:1604.01696,2016.
[8] Rowan Zellers and Ari Holtzman and Yonatan Bisk and Ali Farhadi and Yejin Choi,"HellaSwag: Can a Machine Really Finish Your Sentence?,"arXiv:1905.07830,2019.
[9] Lifu Huang and Ronan Le Bras and Chandra Bhagavatula and Yejin Choi,"Cosmos QA: Machine Reading Comprehension with Contextual Commonsense Reasoning,"arXiv:1909.00277,2019.
[10] Tom B. Brown and Benjamin Mann and Nick Ryder and Melanie Subbiah and Jared Kaplan and Prafulla Dhariwal and Arvind Neelakantan and Pranav Shyam and Girish Sastry and Amanda Askell and Sandhini Agarwal and Ariel Herbert-Voss and Gretchen Krueger and Tom Henighan and Rewon Child and Aditya Ramesh and Daniel M. Ziegler and Jeffrey Wu and Clemens Winter and Christopher Hesse and Mark Chen and Eric Sigler and Mateusz Litwin and Scott Gray and Benjamin Chess and Jack Clark and Christopher Berner and Sam McCandlish and Alec Radford and Ilya Sutskever and Dario Amodei,"Language Models are Few-Shot Learners,"arXiv:2005.14165,2020.
[11] Jacob Devlin and Ming-Wei Chang and Kenton Lee and Kristina Toutanova,"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,"arXiv:1810.04805,2019.
[12] Xuhui Zhou and Yue Zhang and Leyang Cui and Dandan Huang,"Evaluating Commonsense in Pre-trained Language Models,"arXiv:1911.11931,2021.