# Evaluating Pre-Trained Models for Physical Commonsense: RoBERTa vs. GPT-2

Sriram Theerdh Manikyala
manikyala@arizona.edu

## Abstract

In this project, we compare the RoBERTa and the GPT-2 models in the context of the PiQA (Physical Interaction Questions and Answers) dataset, which tests a model's understanding of physical reasoning. The PiQA dataset is pre-processed and tokenized to be compatible with both models. RoBERTa is used as a multiple-choice model, while GPT-2 ranks possible answer choices according to their language generation loss. The performance of the model was recorded in terms of accuracy, taking into account that on teste results, the accuracies of 72.00% and 50.44% were achieved for GPT-2 and RoBERTa respectively. The two models are also visualized comparatives in a bar chart, which helps us understand their performance in relative terms. Though this paper demonstrates that transformer-based architectures achieve effective performance on physical reasoning tasks, it also highlights their limitations.

## 1 Introduction

Transformer-based models have achieved state-of-the-art results on a wide range of natural language processing (NLP) tasks since recent advances in this area. In this project we investigate how these type of models RoBERTa and GPT-2 can be applied to Sample PiQA dataset, which was created for checking of commonsense reasoning in the physical world. It consists of multiple-choice questions, where each question comes with two solutions, and the model should choose the more plausible one.

RoBERTa is a robustly optimized BERT model introduced in the paper RoBERTa: A Robustly Optimized BERT Pretraining Approach for Multiple-Choice Tasks. Conversely, a causal language model, such as GPT-2, is used to score options via the likelihood of the choice, as per the language generation loss. The models selected, highlighting their complementary architectures to compare their abilities on physical reasoning tasks

This includes data cleaning and tokenization using RoBERTa and the GPT-2 models with the PiQA dataset. Multiple-choice

questions designed to measure physical reasoning are used to evaluate these models. The findings shed further light on the limits and capabilities of these models, and help inform the broader discussion on the viability of transformer-based models in NLP.

# 2 Methodology

## 2.1 Data Acquisition

We use the PiQA (Physical Interaction Questions and Answers)table, a multiple choice dataset on physical reasoning. An easy-to-use interface is given by using Hugging Face datasets library to get the dataset from Hugging Face website and look up what the dataset contains. the dataset is divided into training and validation subsets, we use the validation set to evaluate the performance of our model.

## 2.2 Data Preprocessing

A preprocessing function is defined to format the input data that is given to the models. The function takes the question and choices for that question then combines them to make a structured input format. With this method, each question is always associated with the two answers, which is the expected format for multiple-choice tasks. Validation set will be passed through the preprocessing function.

## 2.3 Model Selection

We choose two pre-trained transformer-based models to evaluate:

**RoBERTa:** A better optimized BERT based model which has been fine-tuned for Multiple choice type tasks. It performs classification for which one of the 2 answer choices is right.

**GPT-2 :** A causal language model for scoring the answer choices according to their probability considering the question. It calculates the loss for each prediction and chooses the one with the lower loss, which reflects the more likely option.

## 2.4 Tokenization

Input data is tokenized for both of these models. RoBERTa tokenizes the input by merging the question along with each answer option, and then forward the model for classification. With GPT-2, both the question and answer choices are tokenized separately, and the model computes loss for each choice. The Tokenization is done using the AutoTokenizer from Hugging Face Transformers library.

## 2.5 Model Evaluation

The models are then used for prediction on the validation set after preprocessing and tokenizing the data. In the case of RoBERTa, it outputs logits which are used to predict the right answer (the one with the higher score). For the labels consideration in GPT-2, observation yielded a low loss for the best-representing answer and the optimizer aligned the lower value to preferred answers. Accuracies are computed of the predictions made by both of models against the actual labels in the validation set.

# 3 Performance Comparison

The models are evaluated through their calculated accuracy. The bar chart shows the comparison of performance between RoBERTa and GPT-2 on the test dataset. So we can effectively show both models ability to solve for the physical reasoning tasks.

The objective of the project is to evaluate the performance of RoBERTa and GPT-2 on commonsense reasoning tasks based on the PiQA dataset.

Table 1: Model Performance Comparison

| **Model** | **Accuracy** |
|---|---|
| RoBERTa | 50.44% |
| GPT-2 | 72.00% |

# 4 Conclusion

This work studied the performance of two pre-trained transformer-based architectures, RoBERTa and GPT-2, on the PiQA task, a dataset that challenges commonsense reasoning in physical contexts. Both models were able to train on the dataset and produce predictions but to very different degrees of accuracy. RoBERTa, however, had an accuracy of 50.44%, and GPT-2 had an accuracy of 72.00%.

These outcomes indicate RoBERTa's relatively good performance given that it is a model fine-tuned for multi-choice characteristics, whereas the language generation-associated model, GPT-2 had comprehended better the tasks that require physical reasoning. This means that both of them are useful in different aspects and one of them will be more useful for the current problem and appropriate architecture.

The findings are part of the growing interest in exploring transformer models for commonsense reasoning and the potential of using such transfer learning combined with task-specific fine-tuning procedures to improve the performance on the less covered benchmark datasets.

# References


Y. Bisk, M. B. Lin, A. O. K. N. C. M. *PiQA: A Dataset for Commonsense Physical Reasoning*. `https://huggingface.co/datasets/ybisk/piqa`, Accessed: December 2024.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, V. Zettlemoyer, and L. Stoyanov, *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv preprint arXiv:1907.11692, 2019. `https://arxiv.org/abs/1907.11692`, Accessed: December 2024.

Facebook AI, *RoBERTa-Large Model*, `https://huggingface.co/FacebookAI/roberta-large`, Accessed: December 2024. Referred code documentation: `https://huggingface.co/docs/transformers/main/en/model_doc/roberta#transformers.RobertaForMaskedLM`, Accessed: December 2024.