Evaluation of Commonsense Reasoning Using Pre-trained Language Models

Raghav Upadhyay
raghavupadhyay@arizona.edu

Abstract

This project evaluates the performance of pre-trained large language models on the PIQA dataset, a benchmark for physical commonsense reasoning. Two models, facebook/opt-1.3b and roberta-large-mnli, were tested in a zero-shot setting to compare their ability to handle physical reasoning tasks. Using a custom evaluation method, the models were assessed on 20 examples from the validation set. The results showed that facebook/opt-1.3b outperformed roberta-large-mnli with an accuracy of 70.00% compared to 55%. These findings highlight the varying capabilities of these models in solving tasks that require physical commonsense knowledge.

1 Introduction

LLMs (Large Language Models) are widely used for tasks like language translation, question answering, and text understanding. However, reasoning about physical situations remains a challenge for these models.

The PIQA dataset tests a model's ability to reason about physical scenarios. It presents questions where the model must choose the most reasonable solution to achieve a goal, based on everyday physical knowledge humans naturally acquire.

This project evaluates two pre-trained LLMs, facebook/opt-1.3b and roberta-large-mnli, on the PIQA dataset in a zero-shot setting (without additional training). The goal is to compare their accuracy on a small set of questions to understand their strengths and weaknesses in physical reasoning tasks.

2 Methodology

This study evaluates the zero-shot performance of two pre-trained large language models (LLMs), `facebook/opt-1.3b` and `roberta-large-mnli`, on the PIQA dataset. The methodology consists of the following steps:

2.1 Dataset

The **Physical Interaction Question Answering (PIQA)** dataset was used as the benchmark. It is designed to test a model's understanding of physical commonsense reasoning. Each question in the dataset consists of:

- A **goal** describing a physical task.
- Two potential **solutions** (*sol1* and *sol2*), where one is more plausible.
- A **label** indicating the correct solution.

A subset of 20 examples from the validation set was selected for evaluation to keep the computational load manageable.

2.2 Models

Two pre-trained LLMs were selected for evaluation:

- **`facebook/opt-1.3b`**: A general-purpose LLM optimized for efficiency.
- **`roberta-large-mnli`**: A model fine-tuned on natural language inference tasks, known for strong performance in reasoning tasks.

These models were used in a zero-shot setting, meaning they were not fine-tuned on the PIQA dataset.

2.3 Evaluation Function

A custom evaluation function was implemented to measure model performance:

1. For each question, the goal and two solutions were passed to the model as input.
2. The model generated a score for each solution, representing its confidence in the solution being correct.

3. The solution with the higher score was selected as the model's prediction.

4. The model's prediction was compared to the ground truth label to determine correctness.

Accuracy was calculated as the percentage of correctly predicted solutions.

2.4 Implementation

- The Hugging Face Transformers library was used to load the pre-trained models and perform inference.
- The evaluation function leveraged the `pipeline` feature to simplify text classification tasks.
- The experiments were conducted on a system with GPU acceleration where available.

2.5 Comparison

The accuracy of each model was calculated on the same set of 20 examples. The results were compared to analyze their relative performance in reasoning about physical scenarios.

This methodology ensures a fair and consistent comparison of the two LLMs while highlighting their strengths and limitations in physical commonsense reasoning.

3 Results

The evaluation compared the zero-shot performance of two pre-trained models on 20 examples from the PIQA validation set. The results are summarized in Table 1.

Model	Accuracy (%)
`facebook/opt-1.3b`	70.00
`roberta-large-mnli`	55.00

Table 1: Model Accuracy on PIQA Validation Subset

`facebook/opt-1.3b` achieved 70.00% accuracy, outperforming `roberta-large-mnli`, which achieved 55.00%. These results highlight the varying capabilities of the models in physical commonsense reasoning.

4 Conclusion

This project tested two pre-trained large language models, facebook/opt-1.3b and roberta-large-mnli, on the PIQA dataset to evaluate their ability to reason about physical scenarios. In a zero-shot setting, facebook/opt-1.3b performed better, achieving 70.00% accuracy compared to roberta-large-mnli's 55.00%.

These results show that while large language models can handle some physical commonsense reasoning tasks, there is still room for improvement. Fine-tuning on domain-specific datasets like PIQA may be needed to further enhance their reasoning capabilities. This study highlights the importance of developing models that better understand everyday physical knowledge.

References

- **PIQA Dataset**: https://huggingface.co/datasets/piqa
- **Hugging Face Transformers**: https://github.com/huggingface/transformers
- **Hugging Face Datasets**: https://github.com/huggingface/datasets
- **Meta OPT Model**: https://huggingface.co/facebook/opt-1.3b
- **Grammarly**: Rephrasing and grammatical corrections were performed using Grammarly (https://www.grammarly.com).