Benchmarking LLMs

While benchmarking LLMs I used this repository ->

https://github.com/EleutherAI/lm-evaluation-harness/tree/main

Please go through the repository for proper understanding of how this thing works .

This project provides a unified framework to test generative language models on a large number of different evaluation tasks.

1. Create a virtual environment and activate it .
2. To install the lm-eval package from the github repository, we have to run these commands :

```
git clone https://github.com/EleutherAI/lm-evaluation-harness

cd lm-evaluation-harness

pip install -e .
```

3. Then run the following command like this :

```
lm_eval --model hf --model_args pretrained=facebook/opt-125m --tasks hellaswag --device cuda --batch_size auto:4 --output_path "results"
```

   a. In the --model_args parameter you can mention the name of model_id card here . Also you can put the path of the local directory if your model is already downloaded and present locally .
   b. In –device parameter :

      cuda**:** This instructs the script to use a Nvidia GPU (if available) for computations. The script might attempt to automatically choose a suitable GPU.

cuda:0 or cuda:1 (etc.): If you have multiple Nvidia GPUs, you can specify the specific GPU ID to use (e.g., cuda:0 for the first GPU).

cpu**:** This option forces the script to run on the CPU even if GPUs are available. This might be slower in comparison.

c. In the –tasks parameter mention the benchmark name you wanna measure like hellaswag , MMLU ,triviaqa etc.

d. In the --output_path parameter provide the path where you wanna save your results .

4. lm-eval --tasks list -> this command will give you all the benchmark names which you can evaluate .