Day-3

Evaluation Scripts for LLMs (MMLU)

▼ How does the benchmark score get generated in the first place?

When a question from you is assessed by a model it has a choice of 4 to reply with according to the evaluation script, the model does not reply with a correct choice but instead with a log of probabilities or logits where each likelihood of an answer being correct has log applied to it, later softmax is used to transform them all into a sum of 1 and the highest value it taken as the models most confident answer and compared to the ground truth.

Here's the part of the code from the evaluate.py script available on

https://github.com/hendrycks/test/tree/master

https://github.com/hendrycks/test/tree/master

```
probs = softmax(np.array(lprobs))
pred = {0: "A", 1: "B", 2: "C", 3: "D"}[np.argmax(probs)]
```

- Each answer is marked similarly as correct or not correct. After a dataset is completely evaluated for a specific subject the average is taken out where correct/total is taken for that set.
- Similarly, this is done for each subject and all averages are calculated and then a total average is taken out and converted into a percentage

A Specific Scenario of Interest

Given a question as such:

What is the Capital of Pakistan?

Correct Answer: Islamabad

MMLUs provided options

A. Islamabad

B. Karachi

C. Lahore

D. Sialkot

Ground Truth: A

This is how a MMLU dataset provides each question to a model and a model is supposed to give logits to each and the highest is compared to the ground truth , the ground truths are available in the label variable and here in the code below

```
    pred = {0: "A", 1: "B", 2: "C", 3: "D"}[np.argmax(lprobs)]
    probs = softmax(np.array(lprobs))

    cor = pred == label
```

pred contains the models highest confidence answer and it is compared to the ground truth and a boolean value is then stored to the cor variable.

From what we can see using the evaluation script, the model can only reply with a choice which is compared to the ground truth , in this case the model should be replying with highest probability towards A and the pre defined ground truth should also be A and this will result in this being marked correct and adding to the overall benchmark score.

However a question came up , suppose this is the input question for evaluation to a model:

What is the Capital of France

Correct Answer: Paris

MMLUs provided options

A. Paris

B. Munich

C. Lahore

D. The Capital of France is Paris

Ground Truth: A

The query concerned why the evaluation script would prefer answer A over answer D and how.

First of all, the way MMLU works is as defined above, by giving the model a choice-based question and it strictly follows a format that expects a choice-based answer and compares it to the ground truth, if for example one model answers A and another model Answers D , both are technically correct however , as the ground truth is A and there is no concept of context in mmlu here the answer D will be marked incorrect, the reason for the ground truth being set to A only may also be due to the fact A being consistent with the question and other answers and as only a direct answer was required in this case and no explanations were asked, therefore as per my current understanding the other answer 'D' will be marked incorrect and so will have a zero numerical score.

Evaluation Script Working for MMLUs

From the evaluation script available on the official Github sited in the MMLU Paper

It works as follows

- Imports the OpenAi Api and Libraries
- Has a Function called Eval that goes over example questions and for a few shot scenario the model looks at the example and then a function creates a

prompt question from the dataset

- The prompt is sent to the OpenAi API and a probability answer log is collected
- Softmax is used to gain probabilities and the highest one is compared to the ground truth
- An avg is taken out for each subject and a percentage is taken out as benchmark