MMLU BENCHMARK

LOADING

CONFIGURATION

seed	42
top_p	0.0
top_k	0
temperature	0.0
num_generations	1
max_prompt_len	3840
max generation_length	10

TOKENIZER

```
tokenizer = AutoTokenizer.from_pretrained(
            directory,
            use_fast = True,
            padding_side = "left")
```

MODEL

```
pipe = pipeline(
       'text-generation',
       model=directory,
       torch_dtype=torch.bfloat16,
       device=0,
       pad_token_id=
       tokenizer.eos_token_id)
```

DATA

```
dataset = load_dataset("cais/mmlu", "all")
```

Dev Data	57 subs * 5 Q&A	285 Q&A
Test Data	57 subs * varied	14042 Q&A

PROMPT FORMATTING

5-SHOT W/ DEV DATA (SAME SUB)

sample section of length 2:

```
<|start_header_id|>user<|end_header_id|>

Given the following question and four candidate answers (A, B, C and D),
choose the best answer.

Question: Find all c in Z_3 such that Z_3[x]/(x^2 + c) is a field.
A. 0
B. 1
C. 2
D. 3
Your response should end with "The best answer is [the_answer_letter]"
where the [the_answer_letter] is one of A, B, C or D.<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>

The best answer is B.<|eot_id|>
```

TEST DATA (SAME SUB)

```
<|start_header_id|>assistant<|end_header_id|>

The best answer is
```

PROMPT TRUNCATION

while token length + 1 > **max_prompt_len:**
    if section length > 2:
        clip 1st 2 sections

EVALUATION

BATCH PROMPT INFERENCE

```
batch_size = 8
outputs = pipe(batch_prompts,
          do_sample=False,
          top_k=top_k,
          top_p=top_p,
          temperature=temperature,
          max_new_tokens=
          max prompt length,
          num_return_sequences=
          num_generations)
```

OUTPUT PARSING

Find all with "**The best answer is ([A-D])**" and count the matches
num = (section length - 1) / 2

if matches count is (num + 1): answer is last
else: answer is 'E'

MACRO AVERAGING ACCURACY

sub accuracy =
total correct per sub / total Q&A per sub

accuracy = average of sub accuracies

RESULTS

Model Accuracy **49.2**

META Accuracy **49.3**