MMLU 基准测试

本文评估了一个语言模型在 MMLU（大规模多任务语言理解）基准测试上的表现。

MMLU 基准测试是对模型在广泛学科范围内执行各种任务能力的全面测试。它包含涵盖数学、历史、法律和医学等多个领域的多项选择题。

数据集链接：

- Papers with Code
- Hugging Face Datasets

```python
import torch
from datasets import load_dataset
import requests
import json

# 加载 MMLU 数据集
subject = "abstract_algebra"  # 选择你的科目
dataset = load_dataset("cais/mmlu", subject, split="test")

# 格式化提示，包含少量示例
def format_mmlu_prompt(example, few_shot_examples=5):
    prompt = "以下是关于{}的多项选择题（含答案）。\n\n".format(subject.replace("_", " "))

    # 添加少量示例
    few_shot_dataset = load_dataset("cais/mmlu", subject, split="validation")
    for i in range(few_shot_examples):
        ex = few_shot_dataset[i]
        prompt += f"问题：{ex['question']}\n"
        prompt += "选项:\nA. {}\nB. {}\nC. {}\nD. {}\n".format(*ex['choices'])
        prompt += f"答案：{ex['answer']}\n\n"

    # 添加当前问题
    prompt += f"问题：{example['question']}\n"
    prompt += "选项:\nA. {}\nB. {}\nC. {}\nD. {}\n".format(*example['choices'])
    prompt += "答案:"
    return prompt

# 评估循环
correct = 0
total = 0
```

```
for example in dataset:
    prompt = format_mmlu_prompt(example)

    # 向 llama-server 发送请求
    url = "http://localhost:8080/v1/chat/completions"
    headers = {"Content-Type": "application/json"}
    data = {
        "messages": [{"role": "user", "content": prompt}],
        "max_tokens": 5,
        "temperature": 0,
    }

    response = requests.post(url, headers=headers, data=json.dumps(data))

    if response.status_code == 200:
        output_text = response.json()["choices"][0]["message"]["content"]
        predicted_answer = output_text.strip()[0] if len(output_text.strip()) > 0 else ""
    else:
        predicted_answer = ""
        print(f"错误: {response.status_code} - {response.text}")

    # 与真实答案比较
    if predicted_answer.upper() == example["answer"]:
        correct += 1
    total += 1

# 计算准确率
accuracy = correct / total
print(f"科目: {subject}")
print(f"准确率: {accuracy:.2%} ({correct}/{total})")
```