LLM Evaluation

Introduction & Usage

本项目是对Leveraging Large Language Models for Multiple Choice Question Answering中所提到的`CP`和`MCP`两种评估方式的实现，支持对模型`Qwen2-0.5B-Instruct`和`Qwen2-1.5B-Instruct`在数据集`ARC-Easy`和`ARC-Challenge`上进行评估，主要包括以下几个部分：

- `main.py`: 主程序，用于调用`CP`和`MCP`两种评估方式
- `eval.py`: 包括具体模型所对应的类，类中实现了对该模型的`CP`和`MCP`评估方法
- `utils`: 包含一系列辅助函数和类
    - `data.py`: 用于存放数据集对应信息以及加载数据集
    - `prompt.py`: 用于生成对应评估策略的提示信息
    - `acc.py`: 用于计算评估结果的准确率

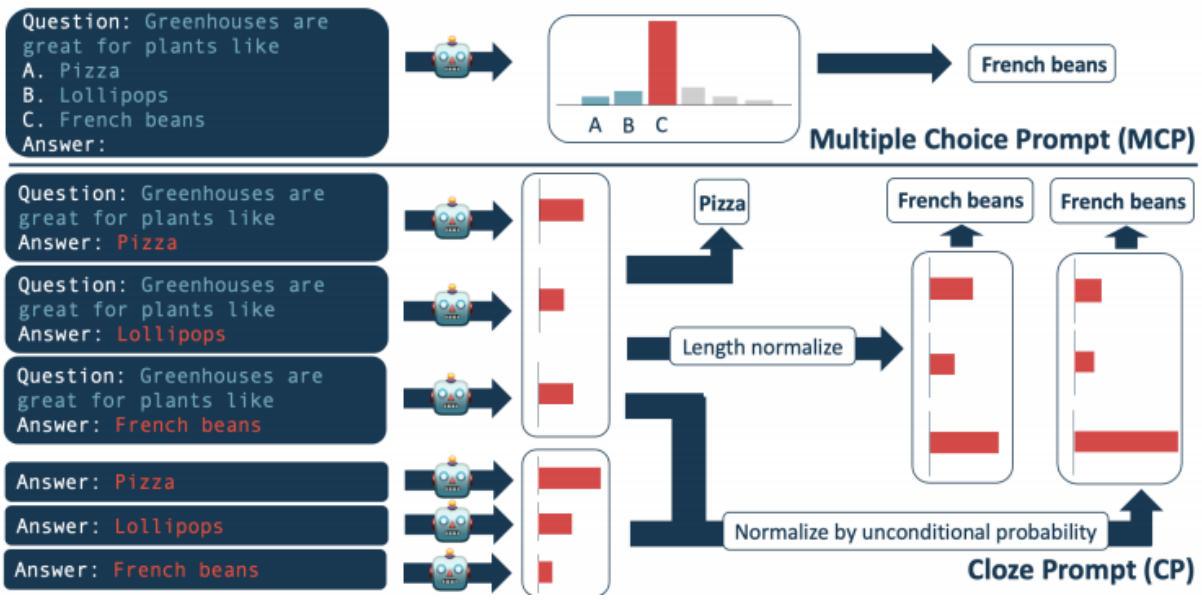
使用方法:

```
python main.py \
    --model MODEL_NAME \
    --dataset DATASET_NAME \
    --k_shot K_SHOT \
    --mcp/cp \
    --torler
```

其中`MODEL_NAME=[qwen2-0.5B-ins, qwen2-1.5B-ins]`,`DATASET_NAME=[ac, ae]`

最后一个参数`--torler`是对`mcp`评估方式的结果的容忍性调整，详情见结果分析部分。

CP vs MCP



	CP	MCP
核心思想	将一个选择题的每一个选项分别作为答案部分构成多个自问自答的文本，将每个文本作为`prompt`输入给`llm`得到输入中答案部分`tokens`的概率（联合概率），选取概率最大的作为最终答案	直接按照选择题的形式将文本输入给`llm`，让大模型预测下一个`token`也即答案选项
相同点	两种策略本质上都是根据问题文本生成答案的条件概率	---
不同点	对于每一个选项的概率预测都是在没有其他答案对比下的	将所有选项都放在一起，能够对选项进行比较
优点	不受`MCSB`能力的影响	1.不受答案文本本身作为自然语言出现的概率对答案预测的影响；2.无需归一化；3.有对选项的比较；4.只需一次`prompt`
缺点	1.由于是根据选项`tokens`的概率来预测答案，会受到选项本身作为自然语言的概率的影响，例如一些在语法上不常见的内容，其分数就会较低；2.需要归一化；3.没有对选项的比较，4.需要多次`prompt`	受`MCSB`能力的影响，在`k-shot`值较小或模型参数较小的情况下不能很好的将选项字母（`A B C D`）和选项文本进行关联，例如对于正确选项`A. Paris`，模型可能认为这整体是一个备选项，从而输出`1`表达其认为第一个备选项是正确的而不是输出`A`，实例见结果分析部分

总结：CP相较于MCP最大的问题就是其缺点1，这是其评估分数一般低于MCP的原因之一，但实际上CP策略对于llm的挑战性更大，因为它本质上更加偏向于填空题，更加考验模型对于问题的理解和对于答案的推理，对模型能力的要求更高，而MCP策略就是模拟选择题的形式，更加考验模型对于选项的理解和对于选项的比较，对模型能力的要求相对较低。通俗一点来说，CP可能要求模型对每一个选项都有一个较为准确的判断，就好比做选择题时，我们不仅要知其然还要知其所以然，难度大。而MCP可能要求模型只要能够通过对比选项的方式找到正确答案即可，就好比做选择题时，我们只需要知其然而不一定要知其所以然，难度小。因此，CP更加能够反映模型的真实能力，但MCP更加能够反映模型的实际应用能力，同时MCP也更加适合在选择题的形式下进行评估。

Results Analysis

Qwen2-0.5B-Instruct在ARC-Challenge上的CP和MCP评估结果如下：

	CP-RAW	CP-UN	CP-LN	MCP	MCP-TORLER
0-shot	0.2244	0.2226	0.2346	0.0034	0.2926
1-shot	0.2329	0.2414	0.2329	0.0	0.2935
5-shot	0.2235	0.2226	0.2389	0.4863	0.4863

可以发现当直接使用MCP评估时，0-shot和1-shot的结果都是近乎0，这是因为在k-shot较小同时llm本身参数也较小的情况下，llm预测的下一个token很可能不是选项字母（A B C D）而是答案项系数，例如对于正确选项A. Paris，模型可能认为这整体是一个备选项，从而输出1表达其认为第一个备选项是正确的而不是输出A，这就导致了MCP评估方式的准确率很低，而带有torler参数则代表将答案项系数预测正确的也算作预测正确的准确率。

Qwen2-0.5B-Instruct在ARC-Easy上的CP和MCP评估结果如下：

	CP-RAW	CP-UN	CP-LN	MCP	MCP-TORLER
0-shot	0.2411	0.2369	0.2352	0.0050	0.2937
1-shot	0.2331	0.2285	0.2420	0.00126	0.3076
5-shot	0.2424	0.2335	0.2432	0.0004	0.2937

Qwen2-1.5B-Instruct在ARC-Challenge上的MCP评估结果如下：

	MCP	MCP-TORLER
0-shot	0.1058	0.3353

	MCP	MCP-TORLER
1-shot	0.1322	0.3430
5-shot	0.2883	0.4308

可以发现当模型参数增大后，不带torler参数的MCP准确率也进一步提升，说明模型的MCSB能力得到了提升，能够更好的将选项字母和选项文本进行关联。