1、数据集导入lm-evaluation-harness框架

Math500数据集结构

{ "problem": "",

"solution": "",

"answer": "",

"subject":"",

"level": ,

"unique\_id":""

}

大致思路为导入"problem""answer"部分，然后使用模型根据problem生成的generated和gt（answer）对比评估

1. 搭载EleutherAI/gpt-j-6B时爆显存

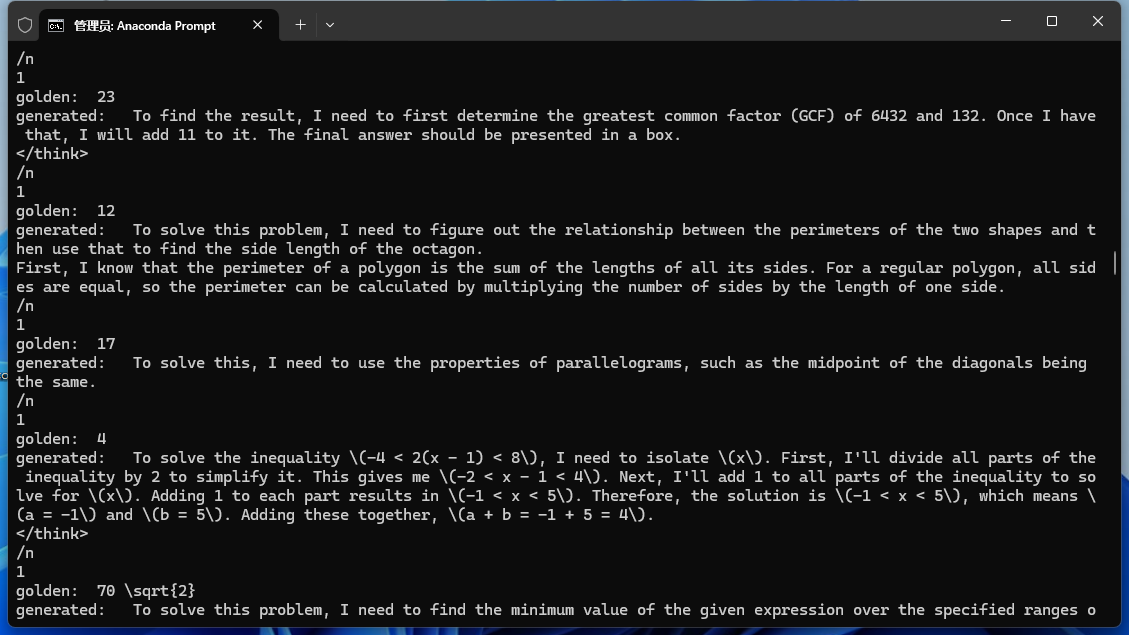
使用4060 8G进行搭载，gpt-j-6B需要20G显存以上，更换模型或者调取API。

lm-evaluation-harness框架支持hg、openai、vllm模型。选择调取DeepSeek-R1-Distill-Qwen-1.5B。

1. 该数据集答案可能包含数学表达式，表达式结构的不同会对结果的评估产生干扰

寻找原数据集中的处理方法，官方代码中给出了grading部分对于数学结果格式进行统一化处理，调用grader.py和math\_normalize.py函数对于gt和generated进行处理后再进行评估。

1. 模型输出带有大段文字分析，会对gt的比较分析产生影响



加入sys\_prompt让模型仅输出最终结果。

在math500（prm800k）的官方代码中寻找相关prompt，无果。

在math原始数据集的论文中寻找，最后找到使用“Final Answer：”的字样（但是原论文提到训练时便使用一部分gt格式为“Final Answer：”的训练集训练）先试试：

f9aebb966c3f479666fb0e10b4352be

发现结果中答案都被包裹在\boxed{ }，加入一个函数将答案提取出来。

根据输出可以发现答案都被很好的提取出来了，但是最终结果只有9个点。

e3ce990a71251038900cc65c0dfb74f

将提示直接加入doc\_to\_text中，结果略微提升，为9.6。

再看看math数据集的官方代码，看看是怎么让模型输出的。发现作者给模型的prompt为"Given a mathematics problem, determine the answer. Simplify your answer as much as possible."并给模型几个示例。用该方法尝试一下。结果没变化。

分析原因，deepseek r1直接输出结果会不会“降智”，如果加上推理再输出结果效果会不会更好。

尝试让模型推理后再输出结果，

79a3bb7fea2c9e718cd338f43c000ad

结果为0.04，再尝试了其他几种prompt后效果依然不佳。

结果不好的原因可能还是在于模型直接输出结果会对模型推理造成负面影响。