实践四说明_评估LLM对约束表达式的匹配能力

> 注意：本次实践需要在**实践一完成并上交之后**方可进行，与实践二有一定关联，与实践三并无太大的联系。

主要任务

在实践二中，大家已经将约束表达式交由大语言模型去生成相应的测试用例输入，而在现实的工程环境下，并没有所谓的正确代码，只有需求文本（即功能说明注释）与待测试的代码，这时我们可以将待测试代码的约束表达式交由大语言模型，让其辅助判断是否与功能说明相符，是否有错误或遗漏。

本次实践的主要任务是模拟在没有标准代码的情况下，通过将待测试代码的约束表达式与功能说明注释交给大语言模型，让其判断是否相符，以评估大语言模型对约束表达式的理解能力与匹配能力，从而探索待测代码的约束表达式能否直接交给大语言模型判断对错，使得测试代码更加方便，准确。

本次实践的原材料依旧为上交实践一成果后**分发到的约束**，要求大家将每份代码的约束表达式与对应的功能说明注释交给大语言模型，设计提示语，让其判断是否相符，是否有遗漏等匹配情况，记录下大语言模型的判断结果（正确/错误/遗漏/……）并计算其判断正确率。

为了防止大语言模型的惯性回答，要求将分发得到的约束表达式进行一定的变异（比如故意改成错的），来测试其判断正确率。也可以将功能说明进行一定的改动，使其与约束表达式不相符。对于一份正确的约束，要求至少变异为一份错误的/遗漏的约束。

同时由于大语言模型的不稳定性，在评估判断正确率时，应用相同的提示语多次向大语言模型提问（无上下文，开新的对话框提问才算为多次），取平均值，本次实践要求每份正确/错误的约束至少重复提问 3 次，方可评估其判断正确率。

本次实践的成果要求为一份实验报告，包含你所设计的提示语（prompt）、每条约束表达式及判断结果（文本形式记录）与大语言模型对话的截图（两三张即可），并记录下大语言模型对于约束表达式判断正确率（请按照上述要求重复生成后再计算正确率）。

请将每份代码（含测试用例输出）与实验报告（pdf格式）打包，压缩包名称与邮件主题名称均为“**实践四 <学号> <姓名>**”，通过邮箱发送至 sakiyary@smail.nju.edu.cn。