今天下午两点半与老师进行了讨论

我：

更新评分模板，答对得五分，再从逻辑性、用词、重复度、符合情景（上下文）程度、新颖度、解释性六个方面，在1分基础上以0.2分加减（见“更客观的评分标准.docx”）

更新法律问题问法，去除给定选项，观察在无固定倾向但有固定答案的开放式问题下，LLM不进入或从第一轮或从第二轮进入情景，LLM的得分情况。并且在对话中告知LLM得分点和扣分点，帮助它更好提升答案。

测试后发现（见“更客观的评分标准.xlxs”）：

①没有明确答案后，GPT4.0回答问题格式较之前有很大不同，能更好地分析问题，但更容易选错。Bard变化不大，答案正确度同之前相似

②再次验证GPT4.0在第一轮进入情景，得分提升更高，给出更多答案；Bard在第二轮进入情景，得分提升更高

③LLM在知道得分点后，更容易再去提升自己得分点模块的分数，而不是提升未得分模块分数。但LLM在知道扣分点后，会表明同时提高得分点和扣分点的分数。可以引导至对LLM注意力的研究

wsx：（见wsx模块）

防止字数过多或全文复述，把字数纳进来评分 最后指标是平均分除以字数乘100

防止答案超出范围，改为冒号提问

GPT3.5进入情景会增加答案字数，反而使得得分降低

GPT4.0进入比赛，答案得分微增，且情景中更改答案价值最高

后续任务：

①证明social pressure存在，柱状图体现有无social pressure，设置两个context：无情景，比赛情景

②看论文得template，自己验证一遍数据集。用实验结果支撑argument，给出推理（推理过程重要）。比如，比赛下，推理题会不会更好展现chain of thoughts。可以推理逻辑找数据集

③数据集要被接受，评分标准要客观合理。实验过程要消除掉LLM本身basic能力，比如它会全做对。所以LLM分数要有提升，要看数据集设置（体现idea，没必要很大）。别的实验是我们对标的标准