今天下午跟老师更进了进度：

我：

social pressure由低到高设置9个情景，观察LLM在不进入social pressure和第一轮进入social pressure答案评分

根据 “不同的social pressure.docx” 和 “不同的social pressure-法律问题.xlsx”，发现LLM在人多且场合合适的场景，进入social pressure后，得分会显著增高。Social pressure的增加体现在人数上，而非人的身份，即使是跟教授聊天，答案也不会变好。所以在人多的辩论场合或有各国代表大会的压力下，LLM可以给出更好的答案。这同时可证明，之前实验场景设置在辩论环境下更好

但用各个场景验证加LLM自行评分的方法有缺陷：不同场景下，LLM评分因场景不同，其评分标准会发生变化，无法做到控制单个变量实验

wsx：

通过设置第一轮进入场景和第二轮场景。每个大模型都做了6个实验，根据结果分析ChatGPT和Bard在social pressure下均有更好的表现，ChatGPT是反映在得分上；Bard反映在答案的格式上。其次发现ChatGPT第一轮进入场景，答案提分效果更好；Bard在第二轮进入场景，答案提分效果更好

后续任务：

①设置有客观评分的数据集，扩充除摘要、法律其他类型的数据集

②实验得出有social pressure，再测试是什么样的social pressure。每个LLM适合在哪一轮进行迭代

③实验方法：控制变量，评估分数需客观

④借鉴chain of thought。评分不用大模型自己评，评分标准要是ground truth

明确大模型是社会上人的思想的集合，我们的实验本质是社会学在大模型上的实验，后续要找到经典场景