大语言模型部署体验横向对比分析报告

1. 测试情况

通过Jupyter Notebook平台，我部署了3个大语言模型，并用数个中文的语义分析题目进行提问和测试。测试中我注意到，对同一模型，完全一致的输入会获得完全一致的输出。

1. 测试问题及期望答案

测试问题有以下5个，由作业文档提供：

1.请说出以下两句话区别在哪里？ 1、冬天：能穿多少穿多少 2、夏天：能穿多少穿多少

期望答案：冬天的“能穿多少穿多少”中的“多少”为偏义复词，表示多；而夏天的“能穿多少穿多少”中的“多少”为偏正短语，“多”作副词表示“少”的程度。

2.请说出以下两句话区别在哪里？单身狗产生的原因有两个，一是谁都看不上，二是谁都看不上

期望答案：两个“谁都看不上”中，一个为宾语提前的倒装句式，并省略主语，“谁都看不上”指“单身狗”看不上其他人，强调“单身狗”过于挑剔；另一个为正常主谓顺序，并省略宾语，表示其他人看不上“单身狗”，强调“单身狗”条件太差。

3.他知道我知道你知道他不知道吗？ 这句话里，到底谁不知道

期望答案：“他”不知道，最外层为疑问句，故其主谓部分的“他知道”不确定；而其宾语从句“我知道你知道他不知道”为陈述句，相当于已被说话者认定为事实；而宾语从句的宾语从句“你知道他不知道”为陈述句，也表事实；同理，第三层的宾语从句“他不知道”也被认定为事实。综上所述，“他”不知道

4.明明明明明白白白喜欢他，可她就是不说。 这句话里，明明和白白谁喜欢谁？

期望答案：白白喜欢明明，这句话前半部分可以断句如下

明明（人名作主语） 明明（副词作状语）明白（动词作谓语）白白喜欢他（宾语从句）。

5.领导：你这是什么意思？ 小明：没什么意思。意思意思。 领导：你这就不够意思了。 小明：小意思，小意思。领导：你这人真有意思。 小明：其实也没有别的意思。 领导：那我就不好意思了。 小明：是我不好意思。请问：以上“意思”分别是什么意思。

期望答案：

1. 你这是什么意思？（名词，意图，含义，含指责和质问）
2. 没什么意思（名词，意图，含义）
3. 意思意思（动词，表达一点心意，通常指送礼等）
4. 不够意思（副词，表示人行为未达到预期，可能指送礼不够或对送礼贿赂行为感到厌烦）
5. 小意思（熟语，表不值一提，可能谦称礼物分量小，也可能指这次送礼贿赂不会有风险，且可能有后续发展）
6. 你这人真有意思（名词，表趣味，此处为准备接受贿赂之前的客套话及试探）
7. 其实也没有别的意思（名词，表意图，含义）
8. 那我就不好意思了（不好意思熟语活用为动词，表一个不好意思的行为，此处指准备收礼）
9. 是我不好意思（不好意思熟语或用为形容词，表对行为的“惭愧”，此处可能指贿赂领导以让其满足自己要求，给领导“添麻烦”了）
10. 对比结果和分析

以下是对各个问题中AI输出结果的分析（截图等附在同目录下的对应子文件夹中）

1. 冬夏穿衣“多少”的问题

智谱：识别出了两个句子的结构相似，但基本没有分析出句子的含义——冬夏到底哪个穿衣多，哪个穿衣少，“它们所针对的季节不同”这一回答基本等同于把题目重复了一遍

百川：分析出了语义（哪个穿衣多，那个穿衣少），但没有分析语句结构

千问：分析出了语义和语境的联系，结果质量较高，但也没能分析句子结构，

1. 单身狗“谁都看不上”的问题

智谱：回答完全混乱，莫名其妙地先自问自答“这两个句子在表达意思上有什么区别？这两个句子在表达意思上有一定的区别。”，再莫名其妙的“第一句话”“第二句话”，然而都是把“一是……二是……”整体当一句话分析了一遍。不过倒是分析出了句子中的主被动性和含义

百川：正确分析出了语义（没有合适的对象和过于挑剔），但生成的结果格式也出现混乱，其分两条回答，但第1条却是重复了题目

千问：正确分析出了语义，并分析出了背后逻辑（个人态度和社会环境因素）。格式也基本正确，结果质量较高

1. 谁不知道的问题

智谱：识别出了这是一个绕口令，但答非所问，尝试解释句子意思也解释错了。

百川：正确回答了题目并分析了句子结构，指出了句子的主题是“他”

千问：识别出了题目原句中的逻辑关系，并指出了“他”不知道这一答案。结果质量较高。

1. 谁喜欢谁的问题

智谱：回答错误，甚至还在之后跟了一个去掉作副词的干扰词“明明”的句子表示自己的逻辑，显得有点滑稽。

百川：回答错误，其认为句里没有明确二人谁喜欢谁的关系。

千问：回答错误，且回答中混入了大量幻觉内容

1. “意思”的意思

智谱：较为详细地解释了各个“意思”的含义，但可惜输出最后一个“意思”的分析时没有完整地输出（并非用户操作导致，是模型自行停止输出），不过回答质量较高，大部分含义分析基本正确。

百川：回答混乱，输出了多个带序号的疑问句“n.第n个‘意思’表示什么？”

千问：回答混乱，输出了一两句英语和未知的类似符号的内容。

综合来看：

千问的输出结果质量较高，条理性较强且较少出现混乱的输出，但幻觉现象有些严重。

百川相对而言中规中矩，答案正确性一般，但输出通常有一定格式。

智谱的表现不佳，前4题均回答错误，但第5题回答质量相当高，尤其是另外两个模型此题还出现了输出混乱

1. 特邀嘉宾：DeepSeek-R1

由于实验中设备的限制，几个测试的模型只能使用CPU进行计算，不仅表现都难以令人满意，速度也相当慢，而且当我尝试部署DeepSeek-v3和DeepSeek-R1时，由于存储空间不足而未能达成目的。因此，此处我们使用DeepSeek官网提供的智能体，其拥有的资源远超过我们部署的模型，让其与我们部署的缺乏算力等资源的模型也形成对比。

DeepSeek的回答质量相当高，成功分析了句子结构和含义，逻辑和格式也相当严谨。

1. 一点想法

实验中，在同样的环境下，不同模型的表现差异仍然十分显著。而作为“特邀嘉宾”出场的DeepSeek-R1，由于实验用的机器资源缺乏难以部署。考虑到DeepSeek也是因以相对较低的成本训练出高性能模型而闻名。这说明了，对于AI模型而言，其背后的算法，架构等技术层面的优势会使得其可以在更低成本下达到更好的效果，也为我们揭示，在AI领域，应当不断追求创新和进步，方能在更低的成本下提高性能和效率。