**大语言模型部署实验报告**

2253893 苗君文

**一、部署过程**

1. **登录并使用魔搭平台，获取CPU云计算资源；**

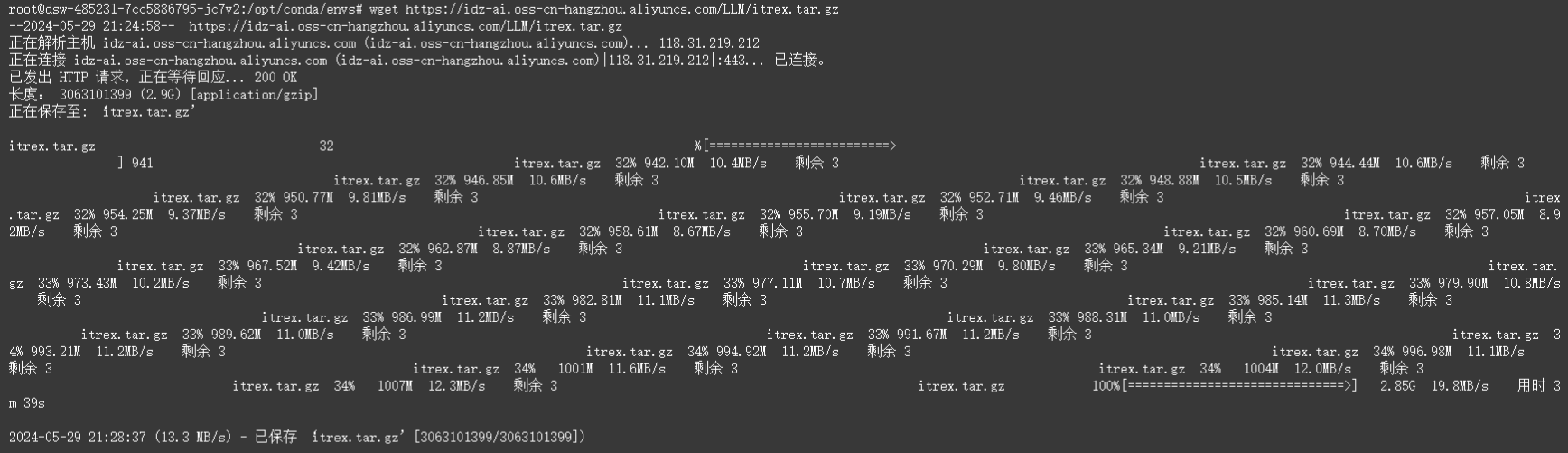


1. **通过Jupyter Notebook进行环境配置**

在环境目录/opt/conda/envs下新建文件夹 itrex



使用 wget 命令复制运行环境镜像文件并在云主机的 itrex 目录进行运行环境的恢复



解压文件



激活环境



安装对应的kernel

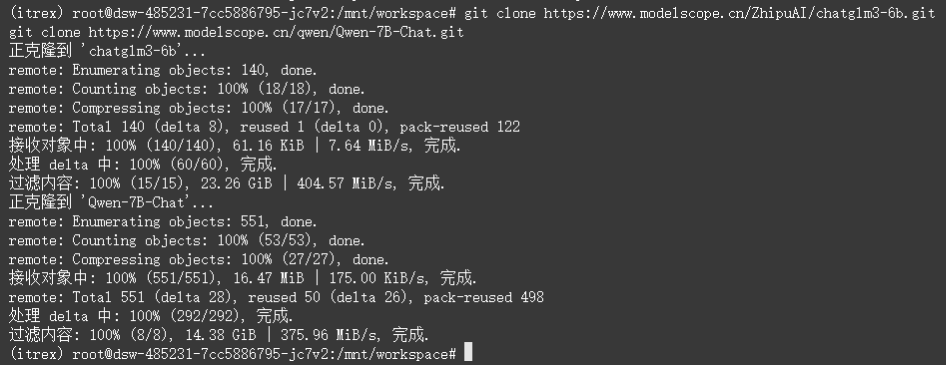


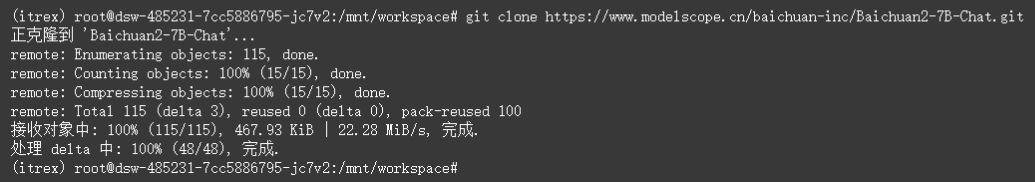
1. **LLM下载**

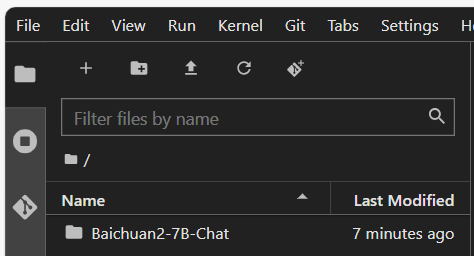
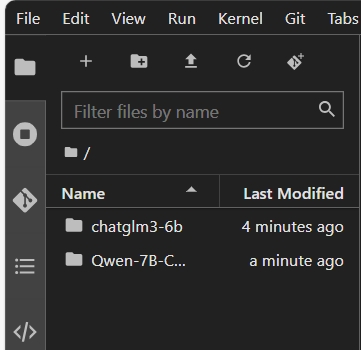
切换至工作目录



git clone下载中文大模型至本地



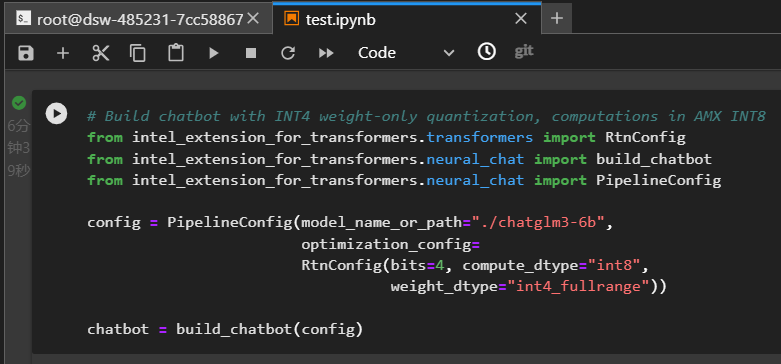




**二、构建chatbot并进行问答测试**

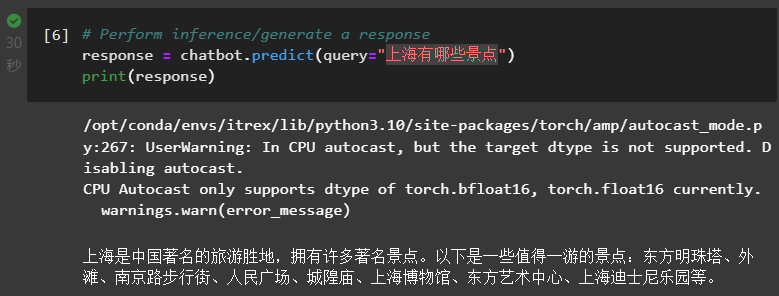
1. **基于chatglm3-6b模型**

量化并构建chatbot

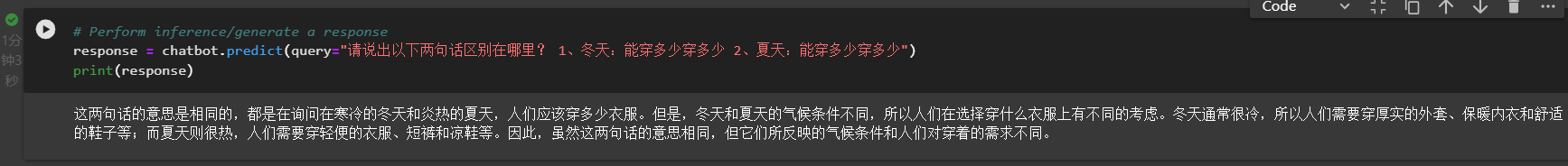


使用 chatbot 得到回答

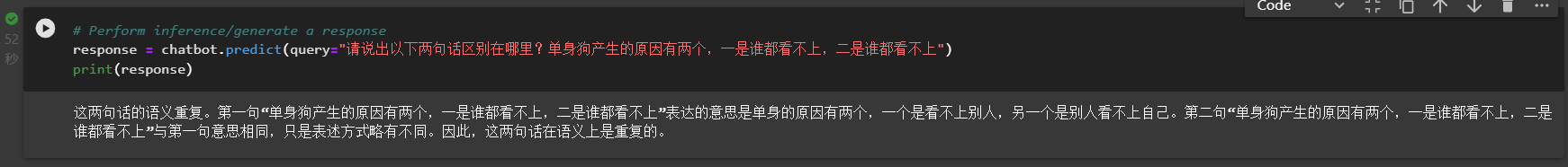
1）上海有哪些景点



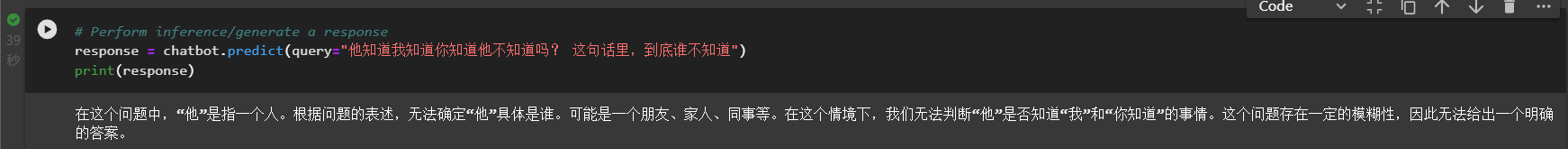
2）请说出以下两句话区别在哪里？ 1、冬天：能穿多少穿多少 2、夏天：能穿多少穿多少



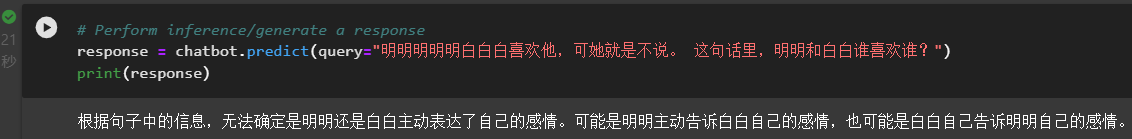
3）请说出以下两句话区别在哪里？单身狗产生的原因有两个，一是谁都看不上，二是谁都看不上



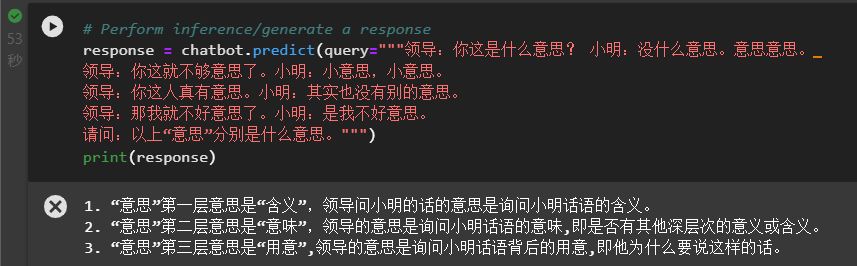
4）他知道我知道你知道他不知道吗？ 这句话里，到底谁不知道



5）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？

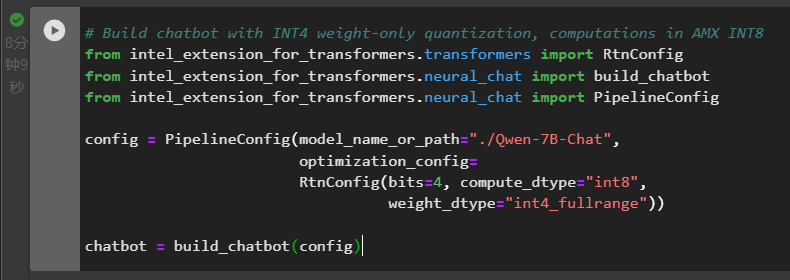


6）领导：你这是什么意思？ 小明：没什么意思。意思意思。 领导：你这就不够意思了。 小明：小意思，小意思。领导：你这人真有意思。 小明：其实也没有别的意思。 领导：那我就不好意思了。 小明：是我不好意思。请问：以上“意思”分别是什么意思。



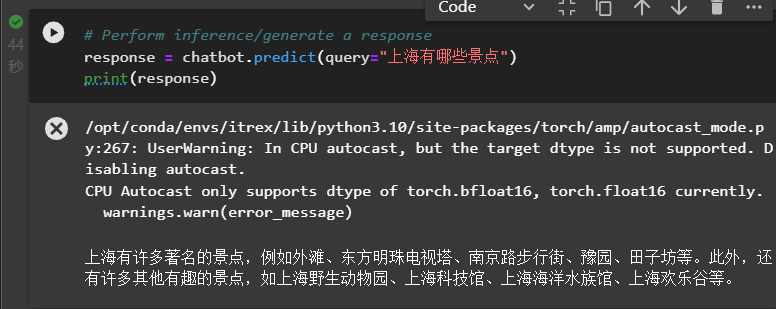
1. **基于Qwen-7B-Chat模型**

量化并构建chatbot

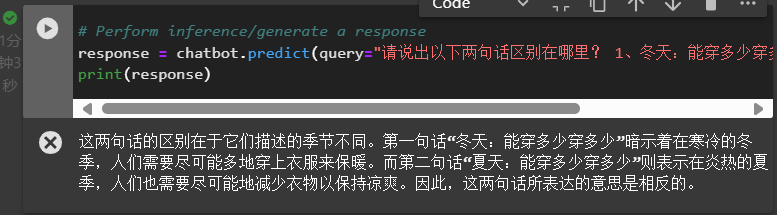


使用 chatbot 得到回答

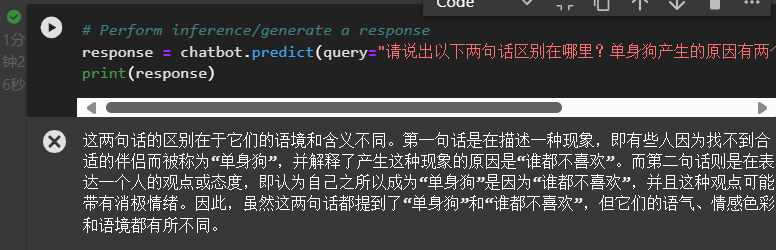
1）上海有哪些景点



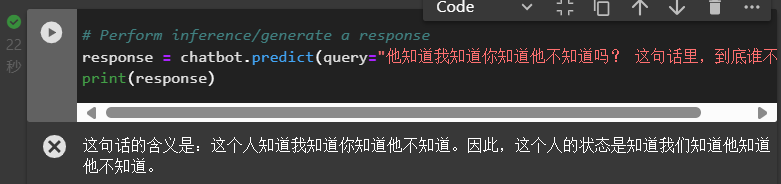
2）请说出以下两句话区别在哪里？ 1、冬天：能穿多少穿多少 2、夏天：能穿多少穿多少



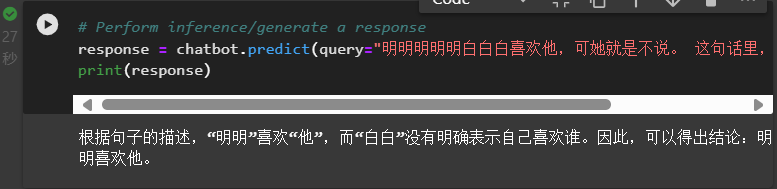
3）请说出以下两句话区别在哪里？单身狗产生的原因有两个，一是谁都看不上，二是谁都看不上



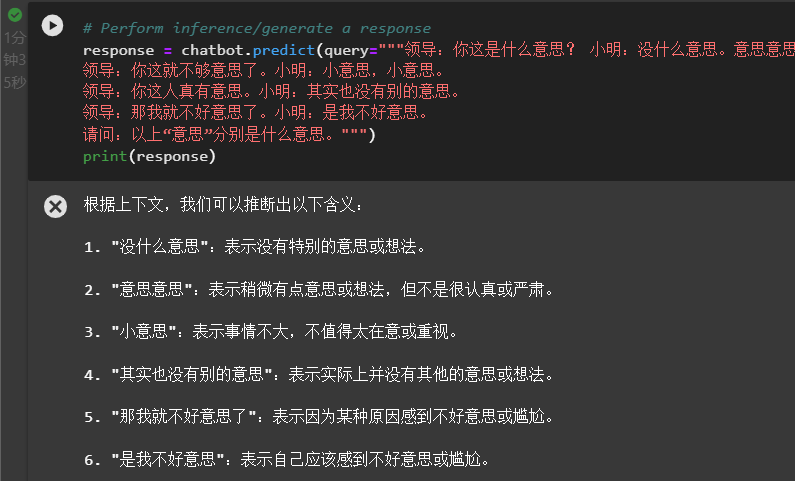
4）他知道我知道你知道他不知道吗？ 这句话里，到底谁不知道



5）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？

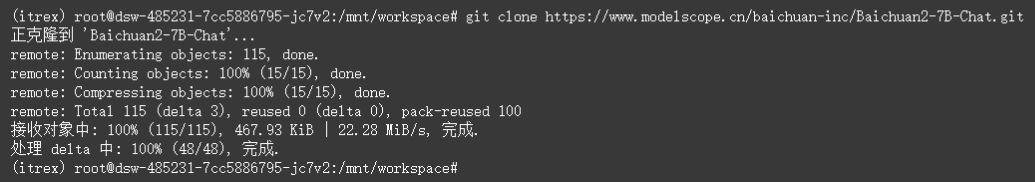


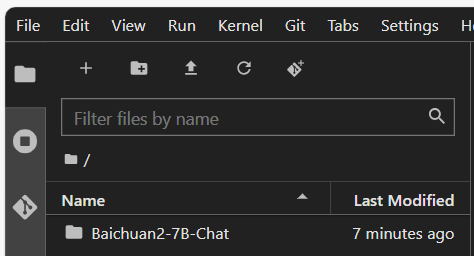
6）领导：你这是什么意思？ 小明：没什么意思。意思意思。 领导：你这就不够意思了。 小明：小意思，小意思。领导：你这人真有意思。 小明：其实也没有别的意思。 领导：那我就不好意思了。 小明：是我不好意思。请问：以上“意思”分别是什么意思。



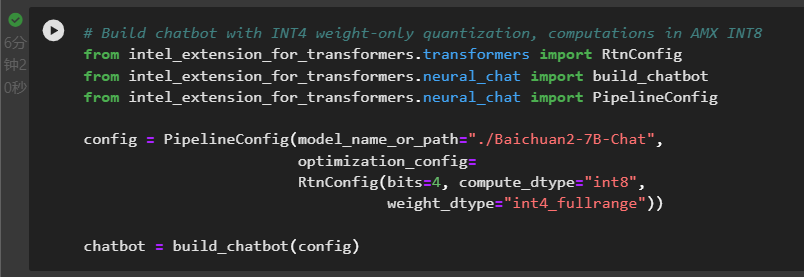
1. **基于Baichuan2-7B-Chat模型**

Git clone





量化并构建chatbot



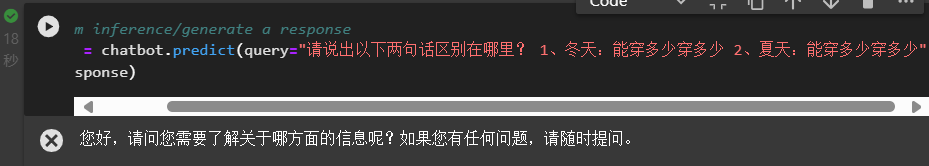


使用 chatbot 得到回答

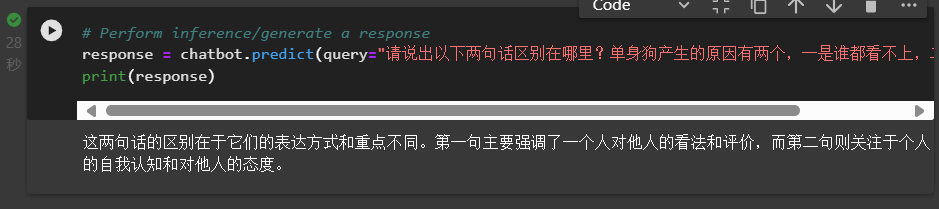
1）上海有哪些景点



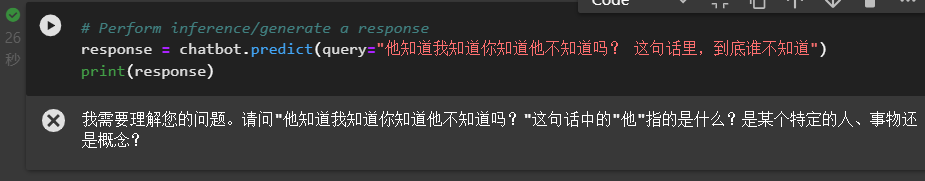
2）请说出以下两句话区别在哪里？ 1、冬天：能穿多少穿多少 2、夏天：能穿多少穿多少



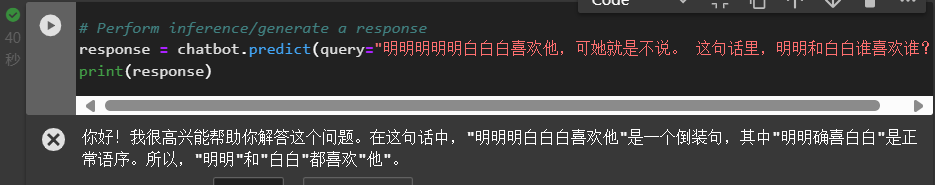
3）请说出以下两句话区别在哪里？单身狗产生的原因有两个，一是谁都看不上，二是谁都看不上



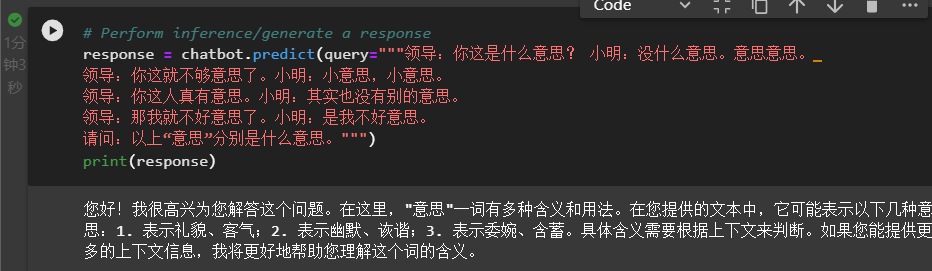
4）他知道我知道你知道他不知道吗？ 这句话里，到底谁不知道



5）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？



6）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。



**三、横向对比分析：ChatGLM3-6B、Qwen-7B-Chat、Baichuan2-7B-Chat**

1. **模型架构对比**
2. 智谱ChatGLM3-6B

参数量: 6亿（6B）

基础模型: GLM（General Language Model），基于自回归生成模型架构，具有良好的推理和生成能力。

优化技术: 利用自监督学习技术，并针对中文语言进行了特别优化。

1. 通义千问Qwen-7B-Chat

参数量: 7亿（7B）

基础模型: Transformer架构。

优化技术: 包含了诸如混合精度训练和模型剪枝等技术以提高性能和效率。

1. 百川2-7B-对话模型

参数量: 7亿（7B）

基础模型: Transformer架构，特别针对对话场景进行了优化。

优化技术: 使用了对话优化技术，如对话状态跟踪和多轮对话管理，以提升对话体验。

1. **训练数据对比**
2. 智谱ChatGLM3-6B

数据量和类型: 主要使用中文数据，同时也包含了部分英文数据。特别注重中文语料的丰富性和准确性。

数据质量: 数据集经过精心挑选和处理，以确保模型对中文语言的理解和生成能力。

1. 通义千问Qwen-7B-Chat

数据量和类型: 使用了大量多语言数据集，包括大量中文数据，覆盖广泛的知识领域。

数据质量: 注重数据清洗和预处理，以确保训练数据的质量和多样性。

1. 百川2-7B-对话模型

数据量和类型: 侧重于中文对话数据，涵盖了多种对话场景，如客服、问答、社交等。

数据质量: 数据集经过严格的筛选和处理，以确保对话内容的相关性和上下文连贯性。

1. **针对具体问答的对比分析**
2. 上海有哪些景点

ChatGLM3-6B：回答准确，涵盖了主要的旅游景点，信息全面。

Qwen-7B-Chat：回答详细，提供了更多的景点，覆盖面广，包含了部分非主流但有趣的景点。

Baichuan2-7B-Chat：回答详细且丰富，包含了多个类型的景点，回答内容全面且多样化，简单介绍了每个景点值得游玩的理由。

1. 冬天：能穿多少穿多少 vs. 夏天：能穿多少穿多少

ChatGLM3-6B： 解释准确但有点重复，缺乏深入分析，没有解释清楚“多少”和“多‘少’”在两种语境下的区别。

Qwen-7B-Chat：描述了季节不同，冬天需要尽可能多穿以保暖，夏天需要尽可能少穿以保持凉爽。解释清晰，准确地指出了反向意义，且逻辑清晰、理由充分。

Baichuan2-7B-Chat：表现不佳，未能理解问题，未能提供有效回答，未能识别问题的具体含义。

1. 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

ChatGLM3-6B：解释准确，指出了一个是不被别人看上，一个是看不上别人，但是对句子结构的认识有重复，认为出现了两句话。。

Qwen-7B-Chat：解释细致，不仅解释了含义不同及其背后原因，而且涵盖了语气和情感色彩的不同。

Baichuan2-7B-Chat：回答内容强调表达方式和重点不同，但解释不够具体，缺乏深入分析。

1. 他知道我知道你知道他不知道吗？

ChatGLM3-6B：对问题分析的侧重点不正确，没有给出明确答案，无法处理复杂的嵌套逻辑，表现不佳。

Qwen-7B-Chat：解释清楚，指出这个人知道其他人知道他不知道。逻辑分析清晰，能理解复杂嵌套。

Baichuan2-7B-Chat：未能提供有效回答，问题理解有误。无法处理复杂的嵌套逻辑，表现不佳。

1. 明明明明明白白白喜欢他，可她就是不说。

ChatGLM3-6B：无法确定明明还是白白喜欢他。未能准确理解句子结构和逻辑。

Qwen-7B-Chat：理解了问题的意思，但是分析有误，未能正确解析句子结构。

Baichuan2-7B-Chat：认为句子语序有误，分析角度错误，理解有误，未能正确解析句子。

1. 领导和小明的对话中“意思”分别是什么意思

ChatGLM3-6B：解释了不同语境下“意思”的含义。但是没有解释所有出现的“意思”，仅解释了三个，分析不够全面。

Qwen-7B-Chat：解释了不同语境下“意思”的含义。解释详细且准确，涵盖了各语境下的不同意思，但依旧缺少了一些“意思”的解释。

Baichuan2-7B-Chat：仅解释了“意思”可能表示的含义礼貌、客气、幽默等。解释泛泛而谈，缺乏具体上下文的详细分析。

1. **对比总结与评价**

ChatGLM3-6B：总体表现稳定，在信息完整性和解释准确性方面表现较好，对于具体的知识性问题和简单的语义问题回答较好，但在复杂逻辑和嵌套问题、复杂句子结构的处理上稍显不足。

Qwen-7B-Chat：回答详细且准确，能够很好地理解并解释复杂语句和逻辑关系，是三者中表现最优的模型。尤其在复杂问题和逻辑分析方面展现了强大的能力，适合需要深入解释和逻辑分析的应用场景。

Baichuan2-7B-Chat：在多样化和详细回答方面有一定优势，但在处理复杂逻辑和语义分析问题时表现欠佳，存在理解偏差，更适合娱乐问答和需要详细回答的基础对话场景。

用户可以根据具体应用场景使用不同的大语言模型：

对于一般信息查询和问答任务，推荐使用Qwen-7B-Chat，因其详细且准确的回答能力。

对于涉及复杂逻辑和多层嵌套的问答，首选Qwen-7B-Chat，其强大的逻辑分析能力能够准确解答；其次可以选择ChatGLM3-6B作为备选。

对于需要丰富多样的回答内容，可以参考Baichuan2-7B-Chat，但需注意其在复杂问题上的表现。

三款模型各有优势，适用于不同的应用场景。Qwen-7B-Chat在多数测试场景中表现最佳，尤其适合需要深入分析和解释的场景；ChatGLM3-6B在知识性问答和简单语义处理上表现出色，适合基础问答场景；Baichuan2-7B-Chat则在提供多样化回答方面具有优势，更适用于娱乐和基础对话场景。根据具体需求选择合适的模型，可以最大化发挥其性能优势。

1. **优化建议**

Qwen-7B-Chat：可以通过优化信息量管理，避免信息过载，同时进一步提升在非常复杂语境下的解析能力。

ChatGLM3-6B：可以加强对复杂逻辑和嵌套语义的处理能力，提升其在深度分析问题上的表现。

Baichuan2-7B-Chat：需要提升对复杂逻辑和语义结构的理解能力，同时保持其在多样化回答方面的优势。

通过进一步优化和增强这些方面，这三种模型都可以在各自的适用场景中提供更为优质的服务。