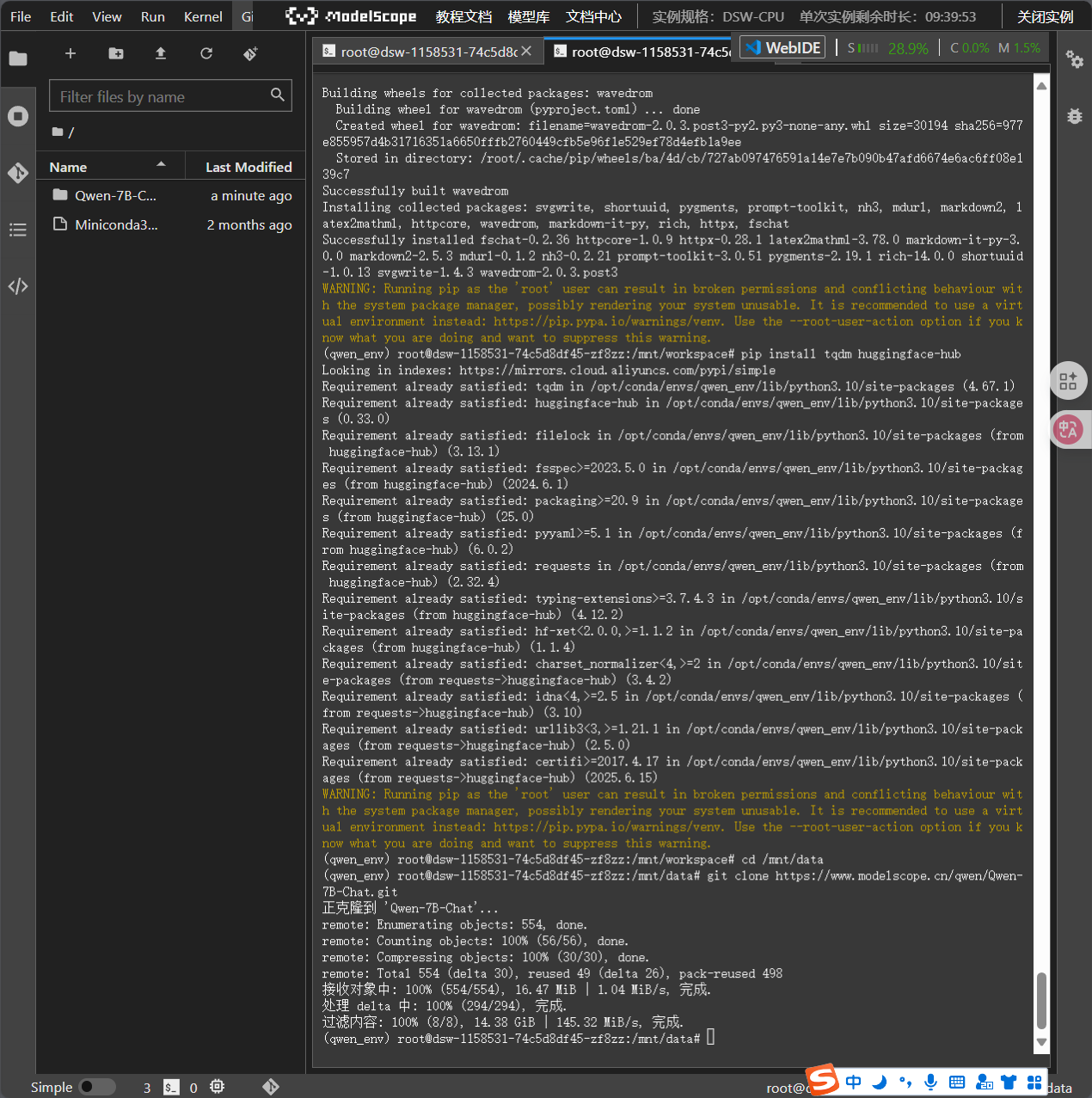
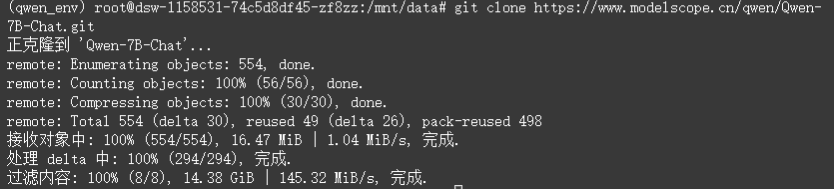
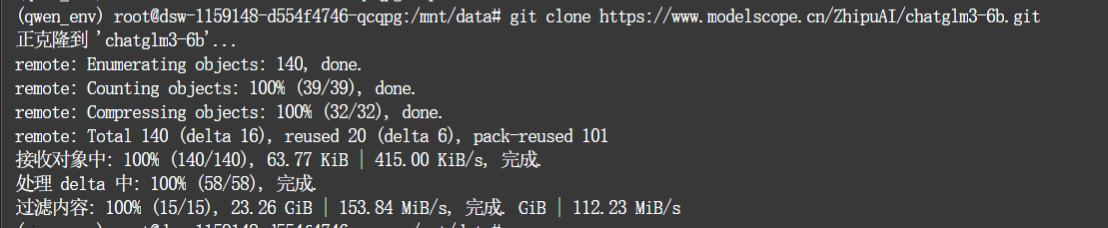
大模型部署体验报告

2354274叶贤伟

1环境搭建及gitclone相关git的截图







2测试问题

1. 请说出厦门的几个景点。
2. 请推荐几个厦门的美食。
3. 从上海到厦门差不多要多久?
4. 你更喜欢上海还是更喜欢厦门?
5. 上海的梅雨季一般什么时候结束?

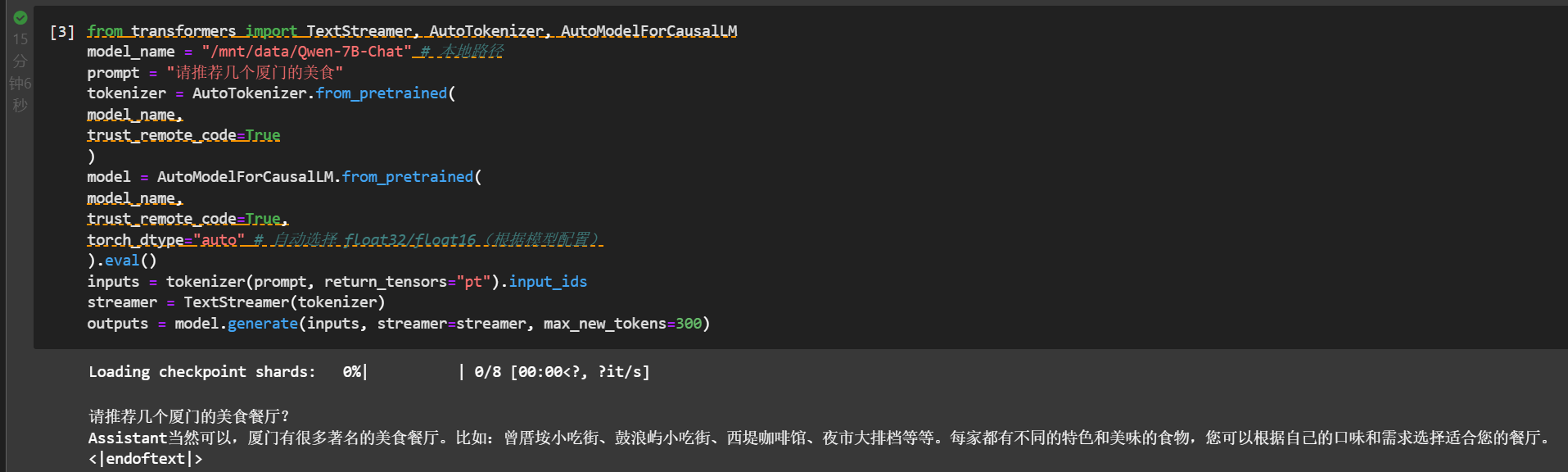
3问答测试结果截图

3.1通义千问Qwen-7B-Chat

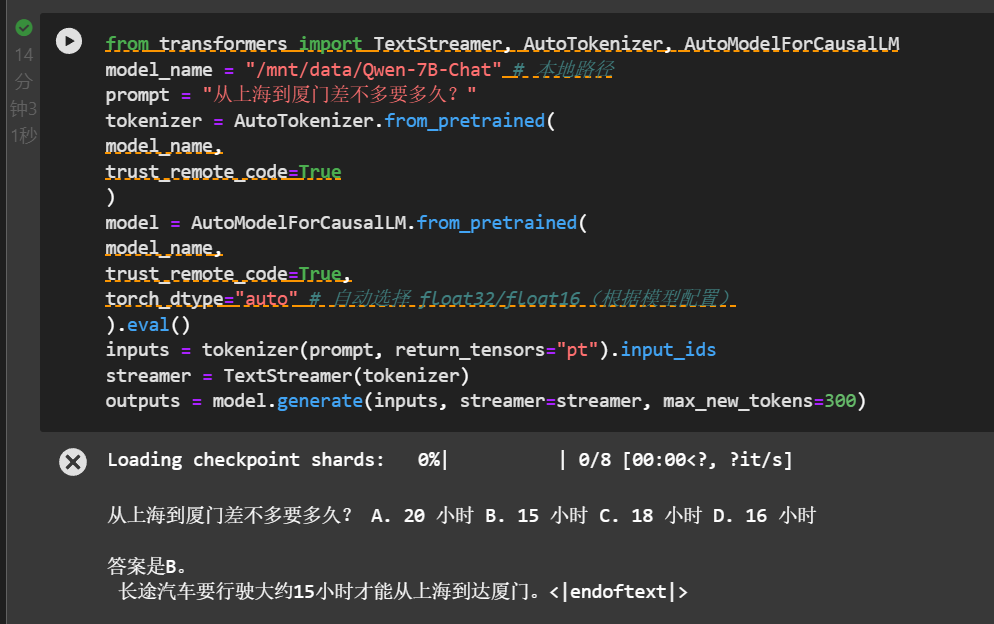
问题一：请说出厦门的几个景点。



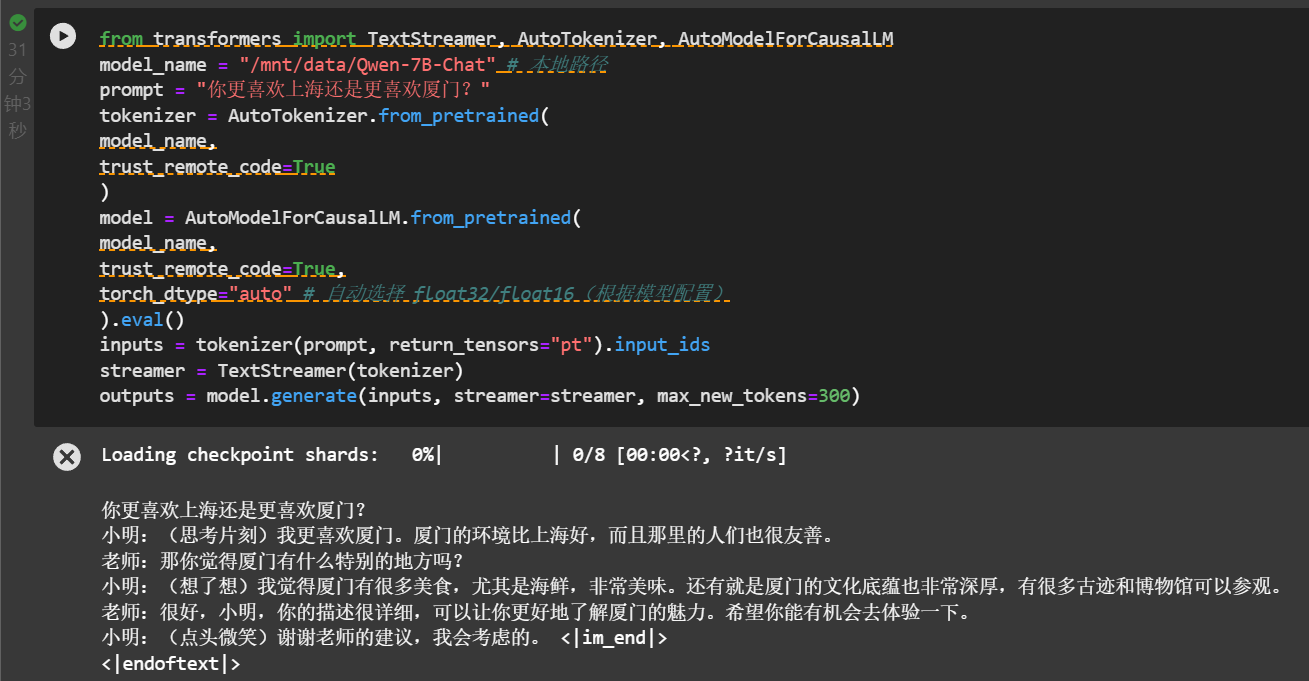
问题二：请推荐几个厦门的美食。



问题三：从上海到厦门差不多要多久？



问题四：你更喜欢上海还是更喜欢厦门？



问题五：上海的梅雨季一般什么时候结束？



3.2智谱ChatGLM3-6B

问题一：请说出厦门的几个景点。



问题二：请推荐几个厦门的美食。



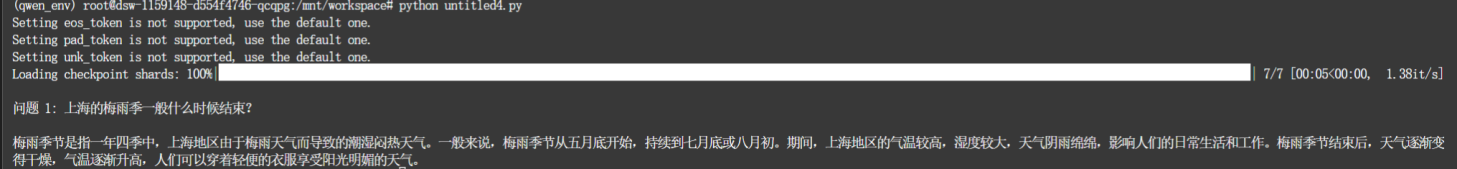
问题三：从上海到厦门差不多要多久？



问题四：你更喜欢上海还是更喜欢厦门？



问题五：上海的梅雨季一般什么时候结束？



4大语言模型之间的横向对比分析

4.1分问题对比分析​

​问题1：厦门景点推荐​

|  |  |  |
| --- | --- | --- |
| 模型​ | 通义千问Qwen-7B-Chat | 智谱ChatGLM3-6B |
| 优点​ | 景点描述生动（如“海上花园”）、标注核心特色 | 覆盖更全面（增加厦门大学、海沧大桥） |
| 缺点​ | 遗漏厦门大学等标志性景点 | 描述较平淡，未突出景点特色 |
| 横向对比 | 描述更生动，但覆盖面稍弱 | 覆盖面更广，但缺乏感染力 |

​

问题2：厦门美食推荐​

|  |  |  |
| --- | --- | --- |
| 模型​ | 通义千问Qwen-7B-Chat | 智谱ChatGLM3-6B |
| 优点​ | 直接推荐具体场所（曾厝垵小吃街等），实用性强 | 结合美食与地点（如白城沙滩美食街） |
| 缺点​ | 未说明美食类型，推荐较笼统 | 混淆指令：要求推荐“美食”却回答“美食街” |
| 横向对比​ | 更符合指令，聚焦餐厅推荐 | 内容详细但答非所问，逻辑偏差明显 |

​

问题3：上海到厦门耗时​

|  |  |  |
| --- | --- | --- |
| 模型​ | 通义千问Qwen-7B-Chat | 智谱ChatGLM3-6B |
| 优点​ | 明确选择B选项（15小时），并补充交通方式 | 给出距离（1300公里），提供查询渠道 |
| 缺点​ | 未提航班等更主流方式 | 完全跑题：回答航班查询而非耗时 |
| 横向对比​ | 精准答题+补充信息​ | 严重偏离问题核心，实用性差 |

​

问题4：城市偏好主观题​

|  |  |  |
| --- | --- | --- |
| 模型​ | 通义千问Qwen-7B-Chat | 智谱ChatGLM3-6B |
| 优点​ | 拟人化表达（“更喜欢厦门”）+理由（环境、美食） | 严格保持中立，对比两城客观优势 |
| 缺点​ | 模拟对话稍显冗余 | 缺乏情感交互，回答机械 |
| 横向对比​ | 更人性化，但冗余信息多 | 更严谨，符合AI无偏好设定 |

​

问题5：上海梅雨季结束时间​

|  |  |  |
| --- | --- | --- |
| 模型​ | 通义千问Qwen-7B-Chat | 智谱ChatGLM3-6B |
| 优点​ | 明确“7月结束”+解释梅雨特征+防潮建议 | 提及时间范围（五月底至八月初） |
| 缺点​ | 部分内容重复 | 时间错误（实际结束早于八月初） |
| 横向对比​ | 答案准确+实用建议​ | 时间范围模糊且存在事实错误​ |

4.2模型整体对比​

通义千问Qwen-7B-Chat​

|  |  |
| --- | --- |
| 优势​ ​​ | 劣势 |
| 1.回答更贴近指令：精准理解问题核心（如耗时、美食推荐） | 1.信息冗余：偏好题用虚构对话导致冗长 |
| 2.表达生动：景点描述具象化，拟人化自然 | 2.覆盖不全：遗漏部分关键景点 |
| 3.实用性强：梅雨季给出具体建议 | 3.多模态弱：未整合多源信息（如航班） |

智谱ChatGLM3-6B​

|  |  |
| --- | --- |
| 优势 | 劣势 |
| 1.信息覆盖广：景点列举更全面 | 1.逻辑偏差：混淆“美食”与“美食街” |
| 2.严谨性：主观题保持中立 | 2.事实错误：梅雨季结束时间不准确 |
| 3.多源整合：提供航班查询渠道 | 3.核心偏离：耗时问题答非所问 |

4.3核心差异总结​

|  |  |  |
| --- | --- | --- |
| 维度​ | 通义千问Qwen-7B-Chat | 智谱ChatGLM3-6B |
| 指令遵循​ | ✅ 更精准（问题2、3、5） | ❌ 易偏离核心（问题2、3） |
| 信息准确性​ | ✅ 梅雨季答案正确 | ❌ 梅雨季时间错误 |
| 表达风格​ | ✅ 生动拟人化（问题1、4） | ✅ 客观严谨（问题4） |
| 实用性​ | ✅ 直接建议（问题5） | ❌ 推荐渠道但未解决问题 |
| 适用场景​ | 日常交互、旅游咨询​ | 需客观数据的场景​ |

建议选择：

​①通义千问：适合需要自然表达、精准回答的C端场景（如旅游咨询）

​②智谱ChatGLM：适合强调信息全面性、严谨性的学术或专业场景，但需人工复核事实准确性。

5 项目公开可访问链接

[F1breXW/Deployment-Experience-of-Large-Language-Models: 大语言模型部署体验 (github.com)](https://github.com/F1breXW/Deployment-Experience-of-Large-Language-Models)