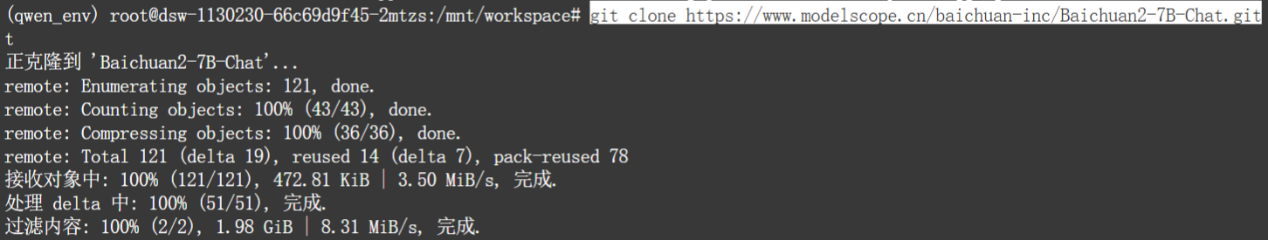
大语言模型横向对比分析报告

# 一、 项目背景

本项目旨在测试和评估三款主流的大语言模型：百川2-7B-对话模型 (Baichuan 2-7B-Chat)、通义千问Qwen-7B-Chat、智谱ChatGLM3-6B，在处理复杂中文语言任务时的综合能力。测试用例的核心是分析多义词“意思”在特定对话语境中的不同含义，以此来考察模型的语境理解能力、语义分析深度、任务遵循能力和逻辑推理能力。

# 二、 测试部署与结果截图

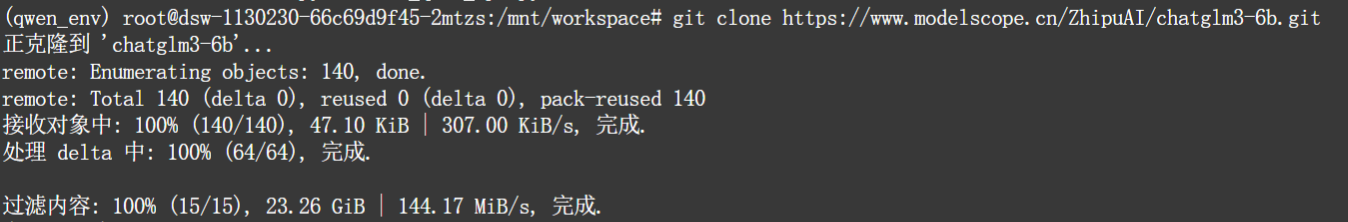
## 2.1 Git Clone 或模型部署截图



百川2-7B-对话模型



通义千问Qwen-7B-Chat



智谱ChatGLM3-6B

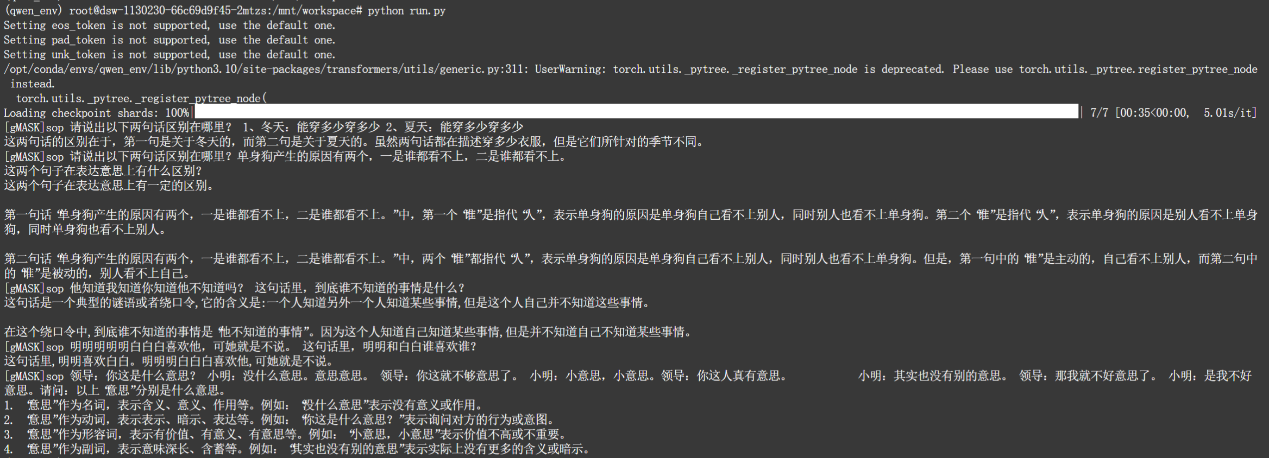
## 2.2 问答测试结果截图



百川2-7B-对话模型



通义千问Qwen-7B-Chat



智谱ChatGLM3-6B

# 三、 大语言模型横向对比分析

本次测试的核心任务是让模型解读一个包含多重“意思”的复杂对话。一个优秀的回答需要准确识别出每一处“意思”的具体含义，并结合上下文进行解释。

## 百川2-7B-对话模型 (Baichuan 2-7B-Chat)

表现概述：表现最优。该模型准确理解了测试任务的核心，即“在特定对话中分析多义词”。

优点：

任务遵循能力强：它没有偏离问题，而是严格按照用户的提问，对一个复杂的“套娃”句式和一段包含多个“意思”的对话进行了直接分析。

语境理解精准：模型成功地、逐条分析了对话中“意思”一词在不同语境下的7种不同含义，例如将“什么意思”解读为“质问、意图”，将“小意思”解读为“礼物不贵重”，将“真有意思”解读为“有趣”，分析得非常到位且符合人类的语言习惯。

分析结构清晰：回答以编号列表的形式呈现，每一条都对应一个具体的语境，逻辑清晰，一目了然。

缺点：在分析第一个“单身狗”的句子时，其解释略显简单，但瑕不掩瑜，核心任务完成得非常出色。

## 通义千问Qwen-7B-Chat

表现概述：表现一般。该模型具备相关的知识，但未能有效应用到具体问题中。

优点：

知识广度较好：模型罗列了“意思”一词的11种不同解释，从“语言文字的含义”到“象征性的表示”再到“情趣”等，展示了其庞大的知识储备。

初步分析尚可：对于第一个“单身狗”句子的分析，它引入了“自恋和自卑感”的心理学视角，有一定深度。

缺点：

任务脱节严重：这是该模型最主要的问题。它没有去分析给定对话中的“意思”，而是给出了一个“意思”的通用字典式大全。这说明它没能真正理解用户的“意图”，即在特定语境中解决问题，而是选择了知识的罗列。未能完成核心测试任务。

## 智谱ChatGLM3-6B

表现概述：表现较差。该模型的回答最为抽象和偏离，几乎完全没有解决用户提出的具体问题。

优点：

具备基础语言学知识：模型从“动词、名词、形容词”等词性角度对“意思”进行了分类，这体现了一定的语言学结构知识。

缺点：

完全偏离任务：与通义千问类似，但程度更甚。它不仅没有分析具体对话，甚至连罗列具体含义都没有做，而是给出了一个高度概括、过于学术化的词性分类。这种回答对于解决用户的实际困惑几乎没有帮助。

实用性差：这种回答方式反映出模型在理解和执行具体、实用的自然语言任务方面存在显著短板，无法将抽象知识与具体应用相结合。

## 横向对比总结

|  |  |  |  |
| --- | --- | --- | --- |
| **评估维度** | **百川2-7B-对话模型** | **通义千问Qwen-7B-Chat** | **智谱ChatGLM3-6B** |
| **任务遵循能力** | 优秀 | 较差 | 较差 |
| **语境理解能力** | 优秀 | 一般 | 较差 |
| **分析深度与准确性** | 优秀 | 一般 | 较差 |
| **知识广度** | 良好 | 优秀 | 良好 |
| **综合表现** | **优** | **差** | **差** |

# 四、 结论与评分建议

通过本次横向对比测试，可以得出以下结论：

百川2-7B-对话模型展现了卓越的语言理解和任务执行能力，不仅能读懂问题的字面意思，更能深刻领会其背后的语境和意图，是三者中当之无愧的最佳模型。

通义千问Qwen-7B-Chat和智谱ChatGLM3-6B虽然拥有一定的知识储备，但在关键的“按需应用”环节表现不佳。它们更倾向于“知识检索与展示”而非“问题分析与解决”，这暴露了它们在理解复杂、具体的用户指令方面存在的短板。