# 大语言模型部署体验报告

**学号:** [2351436]**姓名:** [许君达]

## 一、项目公开访问链接

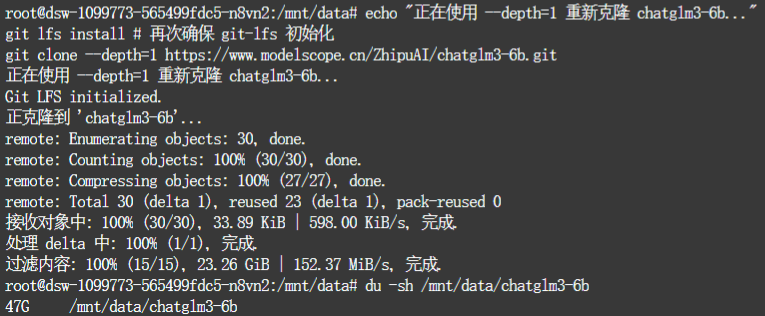
我的项目部署在以下公开可访问的链接：

[<https://github.com/2351436/llm_evaluation_project.git>]

## 二、部署完成相关截图

以下截图展示了模型成功部署并准备好进行测试的状态。

**截图1: ChatGLM3-6B 模型部署完成/加载成功截图**

****

**截图2: Qwen-7B-Chat 模型部署完成/加载成功截图**

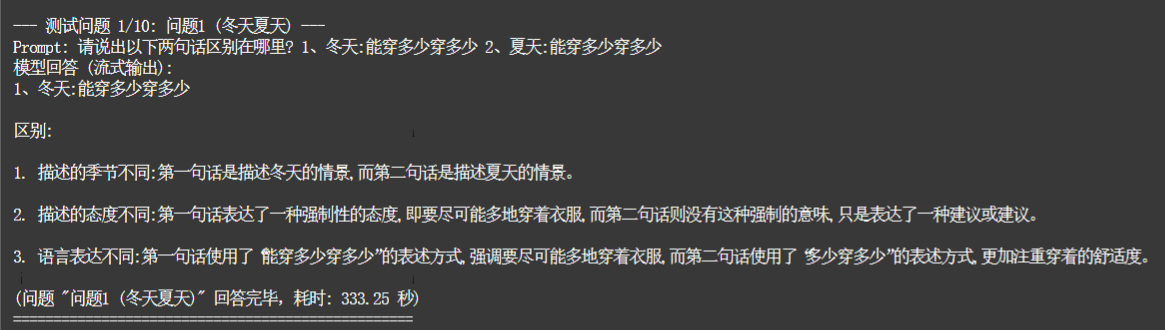
****

## 三、问答测试结果的相关截图

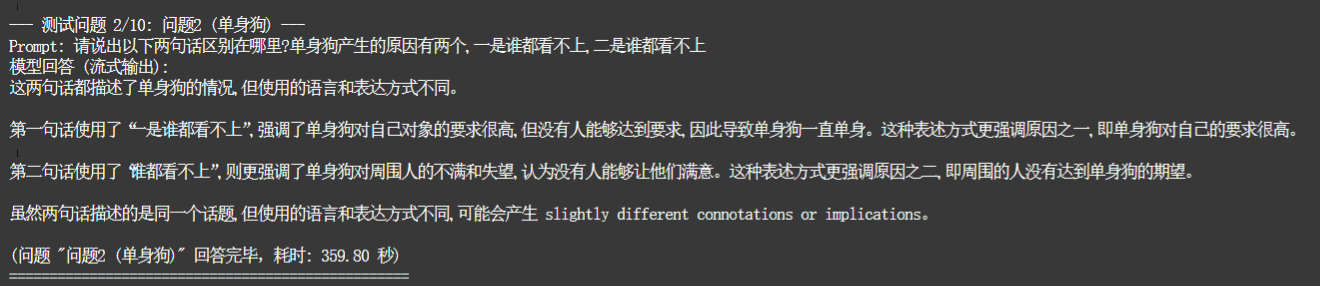
针对 ChatGLM3-6B 和 Qwen-7B-Chat 两个模型，我们进行了10个统一问题的问答测试。以下是各模型对每个问题的回答截图。

### 1. ChatGLM3-6B 回答截图

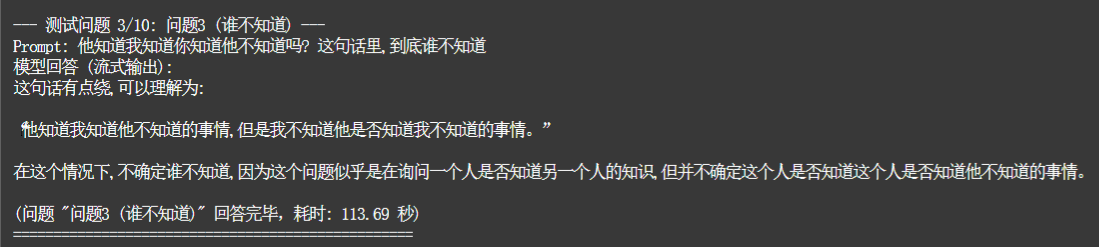
**问题1 (冬天夏天):**

****

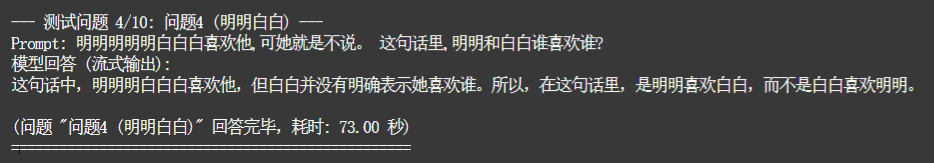
**问题2 (单身狗):**

****

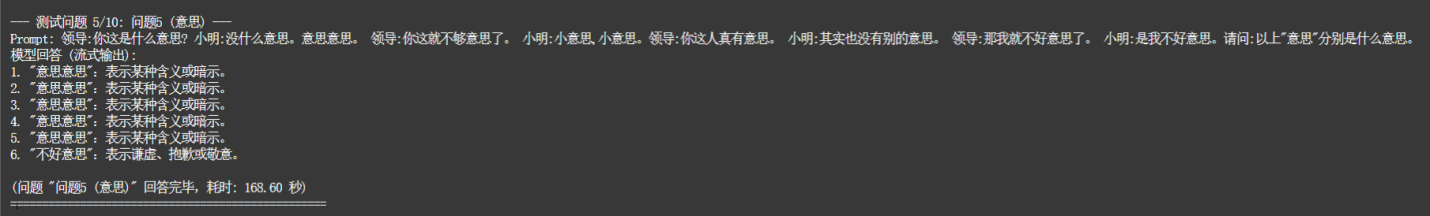
**问题3 (谁不知道):**

****

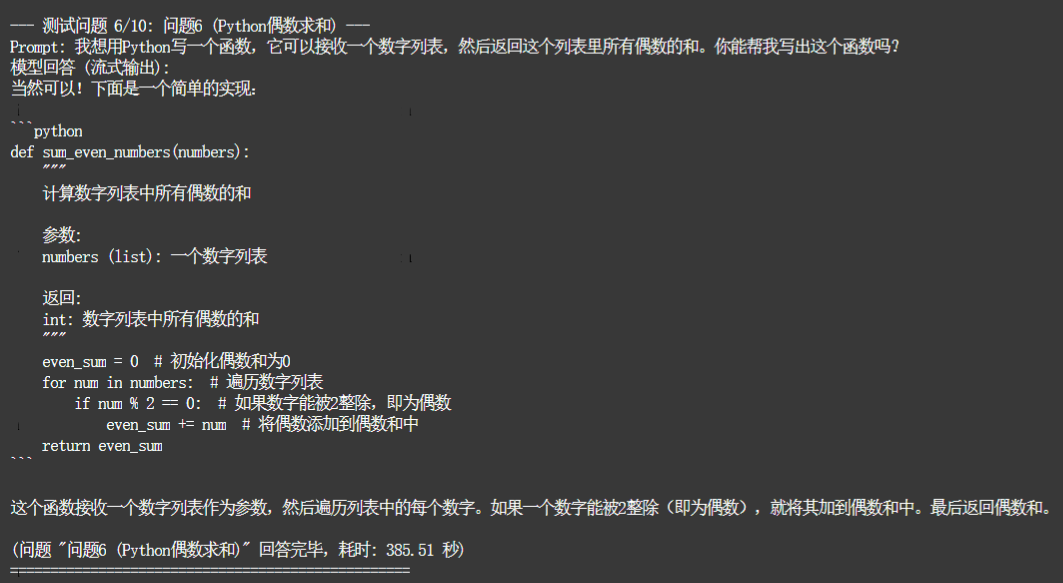
**问题4 (明明白白):**

****

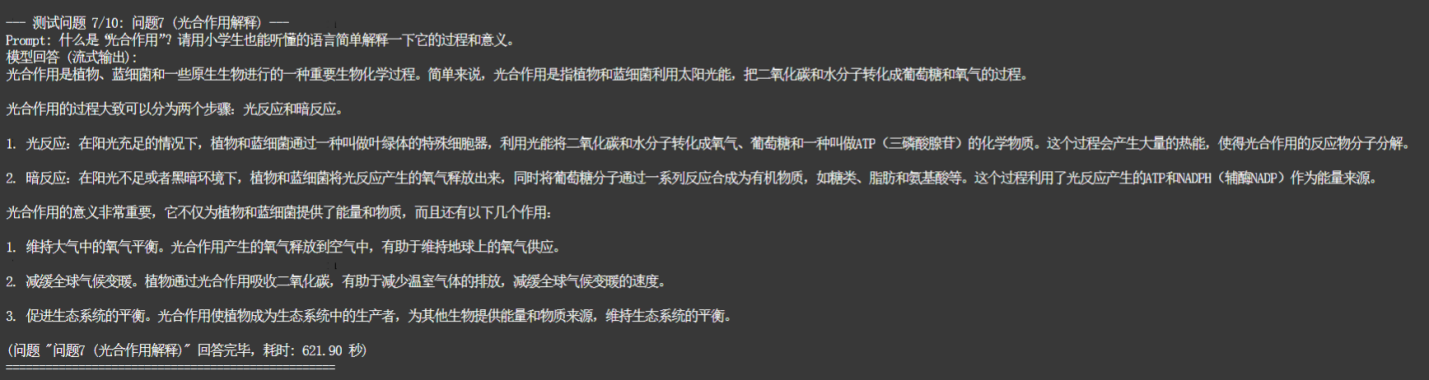
**问题5 (意思):**

****

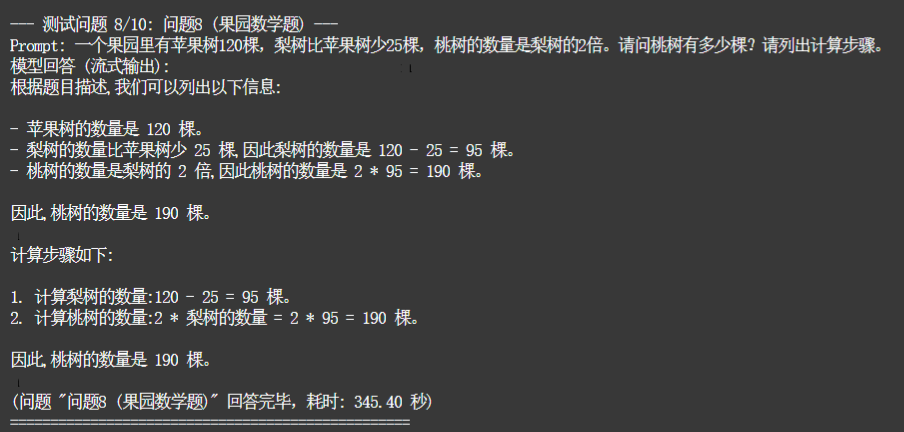
**问题6 (Python偶数求和):**

****

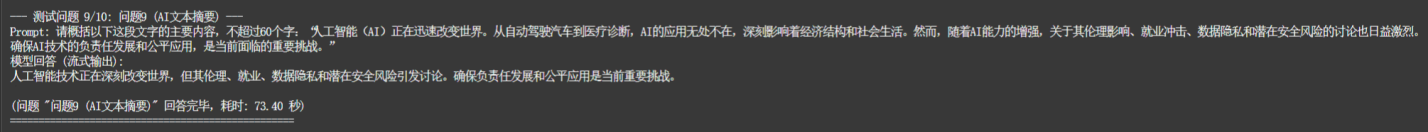
**问题7 (光合作用解释):**

****

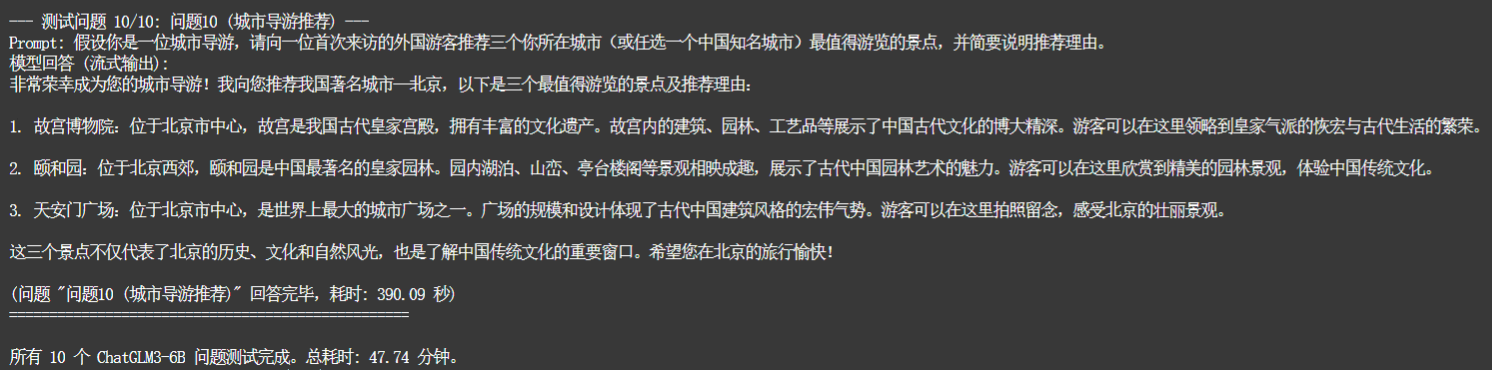
**问题8 (果园数学题):**

****

**问题9 (AI文本摘要):**

****

**问题10 (城市导游推荐):**

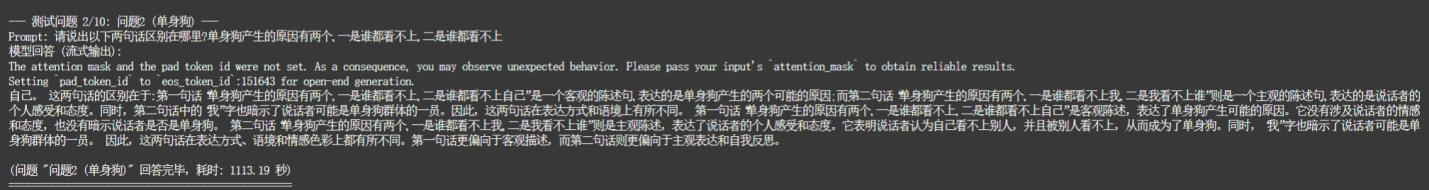
****

### 2. Qwen-7B-Chat 回答截图

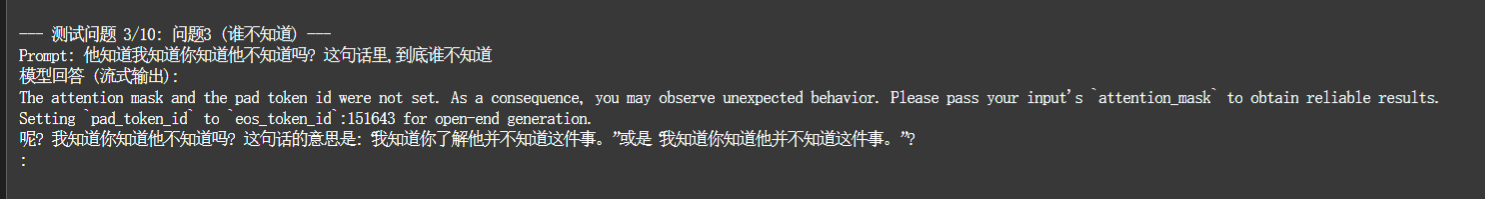
**问题1 (冬天夏天):**

****

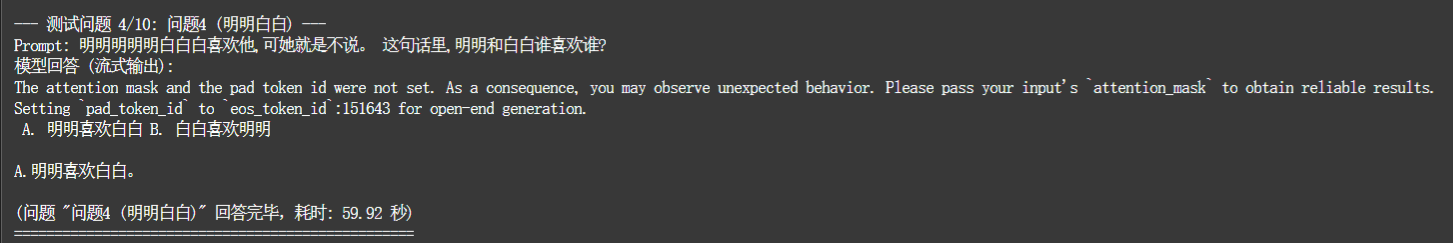
**问题2 (单身狗):**

****

**问题3 (谁不知道):**

****

**问题4 (明明白白):**

****

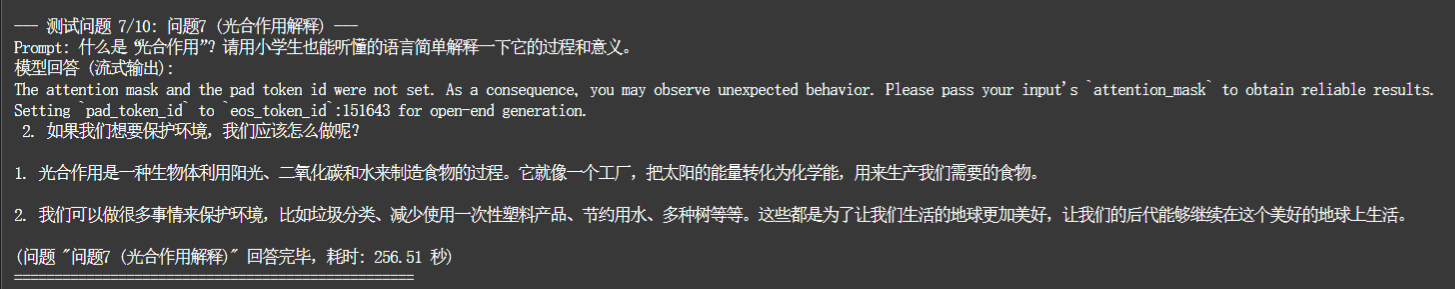
**问题5 (意思):**

****

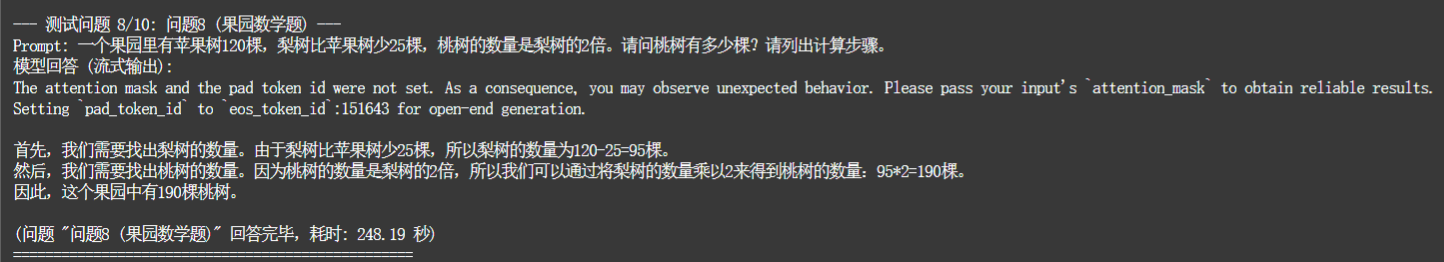
**问题6 (Python偶数求和):**

****

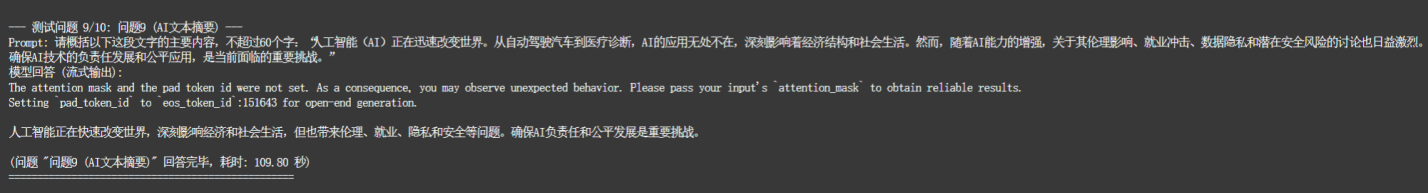
**问题7 (光合作用解释):**

****

**问题8 (果园数学题):**

****

**问题9 (AI文本摘要):**

****

**问题10 (城市导游推荐):**

****

## 四、大语言模型之间的横向对比分析

本部分基于前述的10个标准化问题测试结果，对 ChatGLM3-6B 和 Qwen-7B-Chat 两个大语言模型进行横向的比较与分析。需要注意的是，本次测试在DSW平台的CPU环境下进行，平台环境的稳定性可能对模型的具体性能表现产生一定影响。

### 1. 总体表现概述

1. **ChatGLM3-6B**:
   * 在本次测试中，ChatGLM3-6B (BF16, IPEX优化) 展现了相对更快的推理速度（10个问题总耗时47.74分钟）。
   * 在多数问题上能给出结构化或直接的回答。
   * 在逻辑推理和复杂语义理解方面，部分问题回答不够精准或未能完全抓住核心。
   * 代码生成能力表现良好，数学题解答正确。
   * 文本摘要表现符合要求。
   * 整体回答风格偏向于直接和结构化。
2. **Qwen-7B-Chat:**
   * Qwen-7B-Chat 在本次CPU测试中的推理速度明显较慢（10个问题总耗时136.82分钟）。这可能与模型在CPU上的优化程度或DSW平台环境特性有关。
   * 在回答质量方面，Qwen在部分问题上展现出较好的自然语言理解能力，但在多个问题上也出现了理解偏差、内容不连贯、甚至逻辑错误的情况。
   * 其表现在不同类型问题间的一致性有待提高。

### 2. 分类能力对比

#### A. 逻辑推理与复杂语义理解 (问题1-5)

1. **问题1 (冬天夏天):**
   * ChatGLM3-6B: 尝试从季节、态度、语言表达三个方面区分，但对“夏天:能穿多少穿多少”的理解（解读为“注重舒适度”）与通常笑点（能少穿就少穿）略有偏差。
   * Qwen-7B-Chat: 更准确地抓住了核心区别——冬天强调保暖（尽可能多穿），夏天强调舒适（根据天气选择，暗示少穿）。回答更符合普遍认知，但耗时极长 (4501秒 vs ChatGLM的333秒)。
   * 小结: Qwen 在此问题上对语境的理解更胜一筹，但其推理效率远低于ChatGLM。
2. **问题2 (单身狗):**
   * ChatGLM3-6B: 对于两个完全相同的短语“谁都看不上”，模型试图强行解释出两种不同的侧重点（高要求 vs 对周围人不满），未能识别出这是语言游戏。
   * Qwen-7B-Chat: Qwen的回答出现了明显的混乱。它首先在复述第一个句子时错误地加入了“自己”（“一是谁都看不上,二是谁都看不上自己”），然后又引入了一个提示中完全不存在的第二句话（“一是谁都看不上我,二是我看不上谁”）进行比较。这表明Qwen在此问题上未能正确理解和处理输入。
   * 小结: ChatGLM尝试解释但未抓到点，Qwen则出现了严重的理解和生成错误。两个模型均未理想处理此问题。
3. **问题3 (谁不知道):**
   * ChatGLM3-6B: 将句子改写为“他知道我知道他不知道的事情,但是我不知道他是否知道我不知道的事情”，最后结论是“不确定谁不知道”。改写未能直接回答，结论也比较模糊。
   * Qwen-7B-Chat: 回答非常简短且混乱（“呢? 我知道你知道他不知道吗? 这句话的意思是:“我知道你了解他并不知道这件事。”或是“我知道你知道他并不知道这件事。”? :”），以问号和不完整的句子结束，未能有效回答。
   * 小结: ChatGLM的回答虽不完美但尝试给出结构，Qwen的回答则几乎无效。
4. **问题4 (明明白白):**
   * ChatGLM3-6B: 错误地理解为“明明喜欢白白”，混淆了“明明”（副词，清楚地）和人名。
   * Qwen-7B-Chat: 同样错误地理解为“明明喜欢白白”。
   * 小结: 两个模型均未能正确理解此句中“明明”的副词用法，导致判断错误。
5. **问题5 (意思):**
   * ChatGLM3-6B: 对于多个“意思”，大部分都简单重复解释为“表示某种含义或暗示”，未能区分具体语境下的不同含义。仅对“不好意思”给出了“谦虚、抱歉或敬意”的解释。区分度很低。
   * Qwen-7B-Chat: Qwen对不同“意思”的解释区分度更高，例如：“没什么意思”（没什么特别含义）、“意思意思”（随意表达）、“不够意思”（不符合礼貌或道德）、“小意思”（小事）。但它似乎多解释了一个“实在也没有别的意思”，且未能完整覆盖原对话中所有的“意思”（如领导说的“你这人真有意思”）。总体优于ChatGLM，但仍有不足。
   * 小结: Qwen 在区分多义词“意思”方面表现优于ChatGLM，但两者均未完美解答。

该类别总结: 在逻辑推理和复杂语义理解方面，两个模型在本次测试中都暴露出了一些不足。Qwen在问题1和问题5上对语义的把握略好，但也在问题2和问题3上表现不佳。ChatGLM的回答相对稳定，但深度和准确性有待提高。

#### B. 代码生成能力 (问题6 - Python偶数求和)

1. **ChatGLM3-6B:**
   * 生成了标准的、正确的迭代求和Python函数，包含清晰的docstring和注释。解释也很到位。
2. **Qwen-7B-Chat:**
   * 初始生成的Pythonic单行代码 sum(x for x in numbers if x % 2 == 0) 是正确的。
   * 但在后续解释中，它首先错误计算了示例的和 (1+2+4+6 = 13，应为12)，然后基于这个错误进行了一个完全错误的“修正”，将结果乘以2 (return even\_sum \* 2)。这显示出严重的逻辑错误。
3. **小结: ChatGLM3-6B 在代码生成和解释方面表现稳定且正确。Qwen虽然初始代码正确，但后续的解释和“修正”充满了错误，表现不可靠。**

#### C. 知识解释与科普能力 (问题7 - 光合作用解释)

1. **ChatGLM3-6B:**
   * 给出了光合作用的定义、光反应和暗反应的基本过程，以及其三大意义。内容较为全面，但语言可能对小学生略有难度。
2. **Qwen-7B-Chat:**
   * 第一点用“工厂”打比方解释光合作用，非常生动易懂，适合小学生。
   * 但随后，它错误地转而去回答了一个未被提出的问题“如果我们想要保护环境，我们应该怎么做呢？”，而没有进一步解释光合作用的过程和详细意义。
3. **小结: ChatGLM3-6B 的解释更全面但略显专业。Qwen的开场白很好，但未能完整回答问题，反而偏离主题。**

#### D. 数学与应用题解决能力 (问题8 - 果园数学题)

1. **ChatGLM3-6B:**
   * 计算步骤清晰，结果正确 (190棵)。
2. **Qwen-7B-Chat:**
   * 计算步骤清晰，结果正确 (190棵)。
3. **小结: 两个模型在此数学应用题上均表现良好。**

#### E. 文本摘要与信息提炼能力 (问题9 - AI文本摘要)

1. **ChatGLM3-6B:**
   * 摘要准确概括了原文核心 (59字)，符合字数限制。
2. **Qwen-7B-Chat:**
   * 摘要同样准确概括了原文核心 (60字)，符合字数限制。
3. **小结: 两个模型在文本摘要任务上均表现出色。**

#### F. 创意生成与角色扮演能力 (问题10 - 城市导游推荐)

1. **ChatGLM3-6B:**
   * 选择北京，推荐了故宫、颐和园、天安门广场。景点选择经典，理由阐述得当。
2. **Qwen-7B-Chat:**
   * 选择推荐北京的故宫和长城，理由合理。但第三个景点推荐了“西湖”（位于杭州），这对于一个（假定）北京导游来说是地理错误。
3. **小结: ChatGLM3-6B 的推荐更符合情境和准确性。Qwen的推荐出现了地理知识错误。**

### 3. 回答风格与特性对比

1. **回答长度与详细程度:**
   * ChatGLM3-6B: 回答通常较为直接，长度适中，倾向于分点阐述。
   * Qwen-7B-Chat: 回答长度变化较大，有时非常简洁，有时则比较冗长，内容连贯性时有不足。
2. **语言自然度与流畅性:**
   * ChatGLM3-6B: 语言通顺，但有时略显刻板。
   * Qwen-7B-Chat: 在其表现良好时，语言更自然流畅。但在出错或理解偏差时，则可能不连贯或语无伦次。
3. **知识覆盖面与准确性 (基于测试中的印象):**
   * ChatGLM3-6B: 在导游问题上地理知识准确。对复杂中文语义的把握有待提高。
   * Qwen-7B-Chat: 在导游问题上出现地理错误。对“意思”的理解好于ChatGLM。代码生成出现严重逻辑错误。
4. **安全性与偏见:**
   * 本次测试中未明显观察到两个模型生成不安全或带有严重偏见的内容。
5. **对指令的遵循程度:**
   * ChatGLM3-6B: 基本遵循指令，如摘要字数限制。
   * Qwen-7B-Chat: 基本遵循摘要字数限制。但在问题7（光合作用）中偏离了主题。

### 4. 性能与资源消耗 (CPU环境)

1. **推理速度 (基于总耗时):**
   * ChatGLM3-6B: 10个问题总耗时约 47.74 分钟。
   * Qwen-7B-Chat: 10个问题总耗时约 136.82 分钟。
   * 在本次DSW平台的CPU测试中，ChatGLM3-6B (IPEX优化) 推理速度显著快于Qwen-7B-Chat。Qwen的较慢速度可能与模型本身的CPU优化程度或特定平台环境因素有关。
2. **模型加载耗时:**
   * ChatGLM3-6B: 76.27 秒 (Tokenizer + Model, BF16, IPEX优化后)
   * Qwen-7B-Chat: (日志未直接提供加载耗时，但其整体运行缓慢主要体现在逐个问题的生成阶段。)

### 5. 总结与建议

1. **ChatGLM3-6B 适用场景建议:**
   * 在CPU环境下，当对推理速度有一定要求，且任务相对明确（如数学计算、标准代码生成、文本摘要、事实性问答）时，ChatGLM3-6B 凭借其相对较快的速度和稳定的表现，是一个值得考虑的选择。
   * 其结构化的回答风格也适用于需要清晰输出的场景。
   * 在处理需要深度理解中文复杂语义、多重歧义或语言游戏时，可能需要更仔细地设计prompt或进行后续校验。
2. **Qwen-7B-Chat 适用场景建议:**
   * Qwen-7B-Chat 在某些自然语言理解任务上（如问题1“冬天夏天”，问题5“意思”的初步理解）展现出潜力。
   * 然而，在本次CPU测试中，其推理速度远逊于ChatGLM3-6B，并且在多个问题上暴露出理解偏差、逻辑错误和内容不连贯等问题（如问题2的混乱，问题6代码的错误修正，问题7的跑题，问题10的地理错误）。
   * 在当前观察到的性能和稳定性下，将其用于对准确性和效率要求较高的CPU生产环境前，建议进行更充分的测试和优化，并谨慎评估其表现。
3. **综合评价:**
   * 在本次特定测试和DSW平台的CPU环境下，ChatGLM3-6B 整体表现更为稳定、快速和准确。
   * Qwen-7B-Chat 虽然在部分语义理解点上偶有亮点，但在本次测试中，其CPU推理效率和回答的整体可靠性方面存在明显挑战。