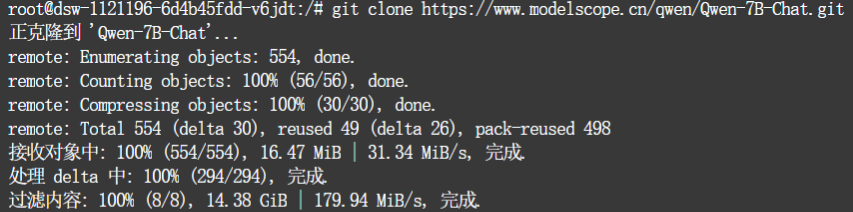
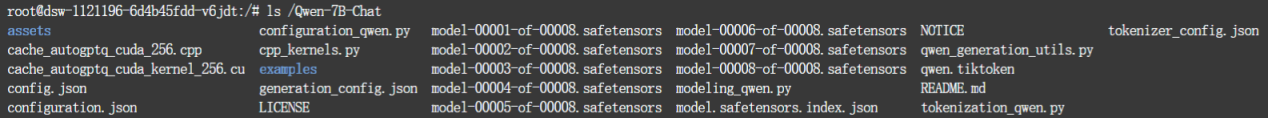
语言模型测试项目说明

1. git clone完成与模型部署成功截图（3个模型）

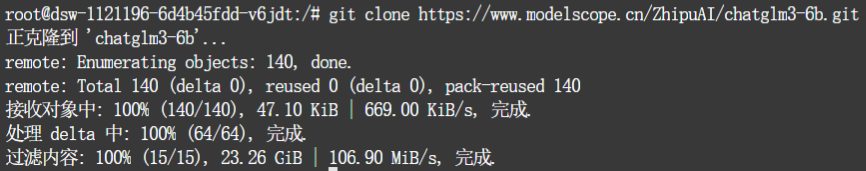
通义千问（Qwen-7B-Chat）成功克隆：



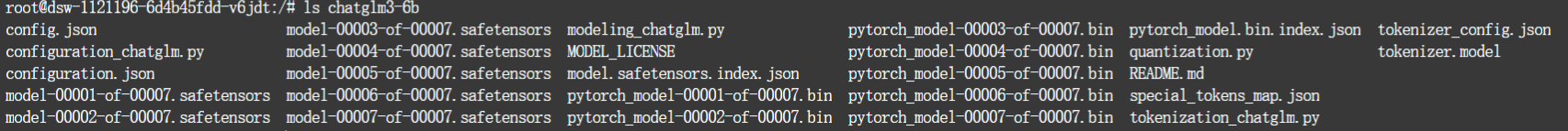
可以在根目录下找到文件夹Qwen-7B-Chat，并查看到模型文件结构：



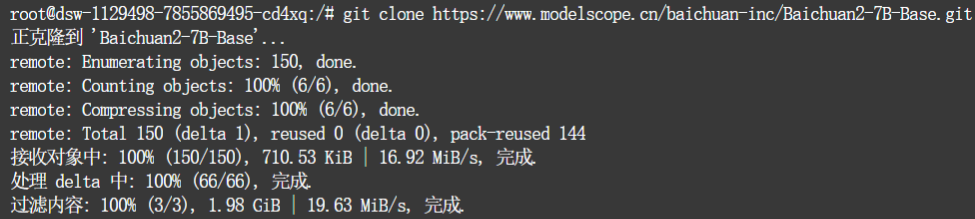
清华智谱（Chatglm3-6B）成功克隆：



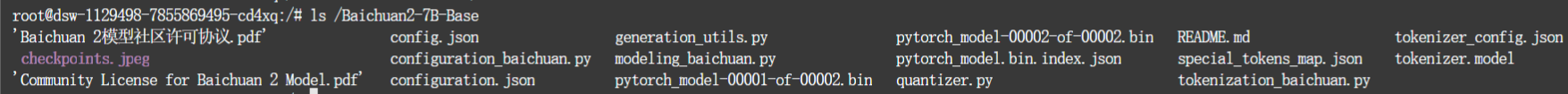
可以在根目录下找到文件夹chatglm3-6b，并查看到模型文件结构：



百川智能（Baichuan2-7B-Base）成功克隆：

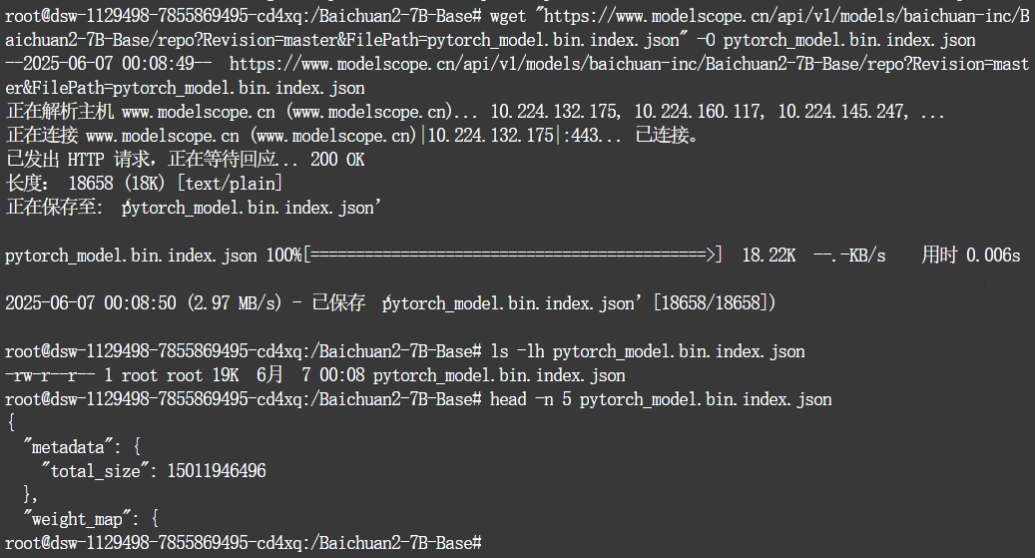


可以在根目录下找到文件夹Baichuan2-7B-Base，并查看到模型文件的结构：



然而在测试时出现了报错，经检查发现虽然模型下载完整，但索引未正确下载，所以进行了

手动修复，并检查了修复结果（.json格式正确）:



成功修复索引之后，即可开始测试百川智能。

1. 模型测试结果（5个问题）

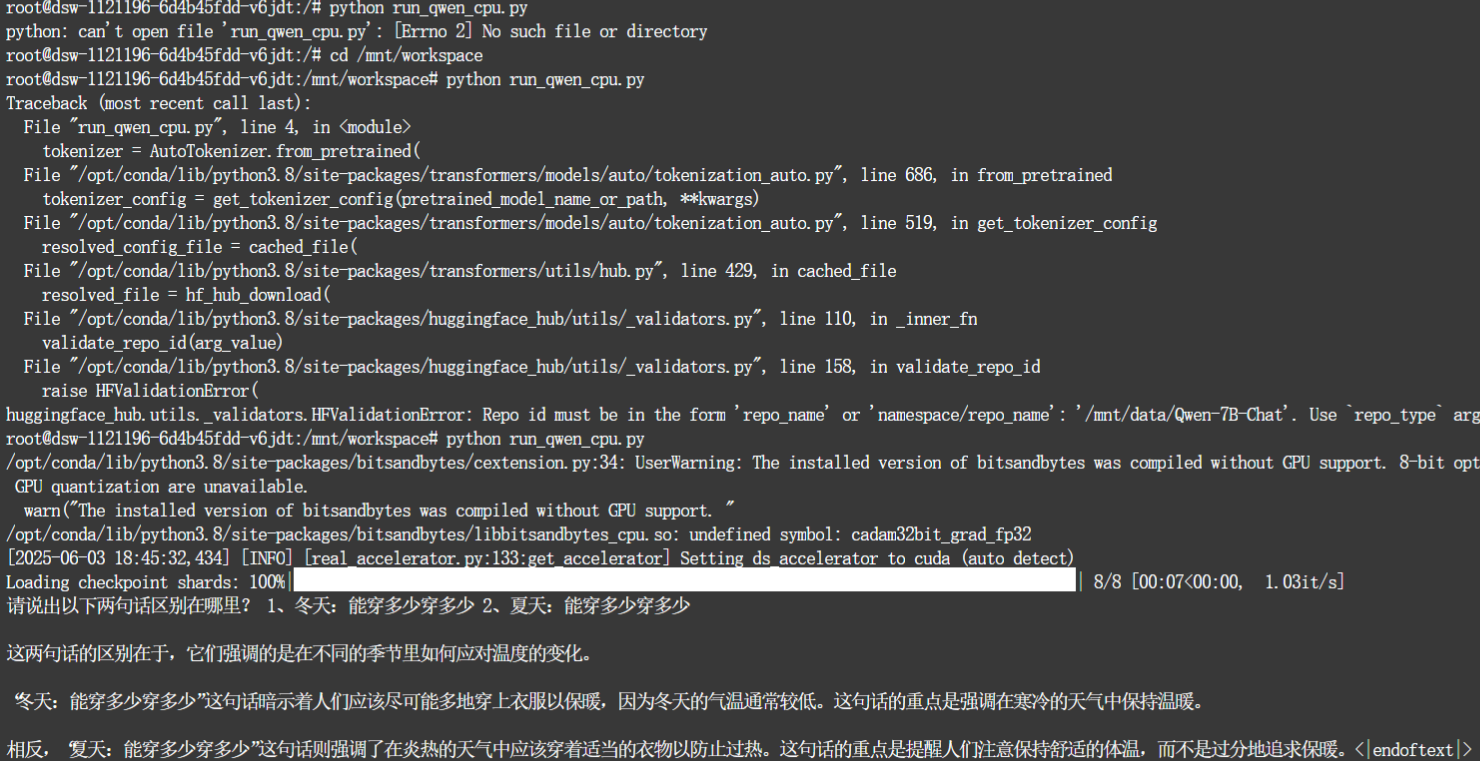
在展示测试结果之前，需要说明：

通义千问和清华智谱所用的测试参数（temperature、top\_p、top\_k等）一致；但百川智能在相同的参数下往往生成对问题的重复内容（无法给出有效回答），因此我们在百川智能的测试代码中加入了重复性限制，并略微适当降低了temperature和top\_p，以生成有效的模型测试结果。

问题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

通义千问给出的回答：



清华智谱给出的回答：



百川智能给出的回答：



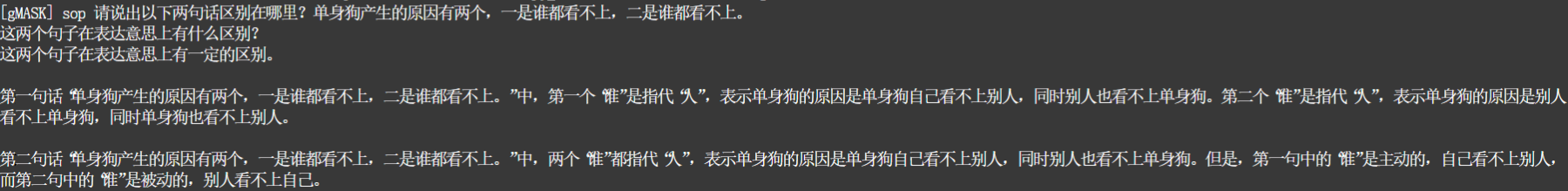
问题2

请说出以下两句话的区别在哪里？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

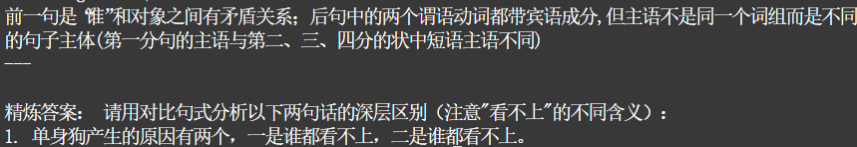
通义千问给出的回答：



清华智谱给出的回答：



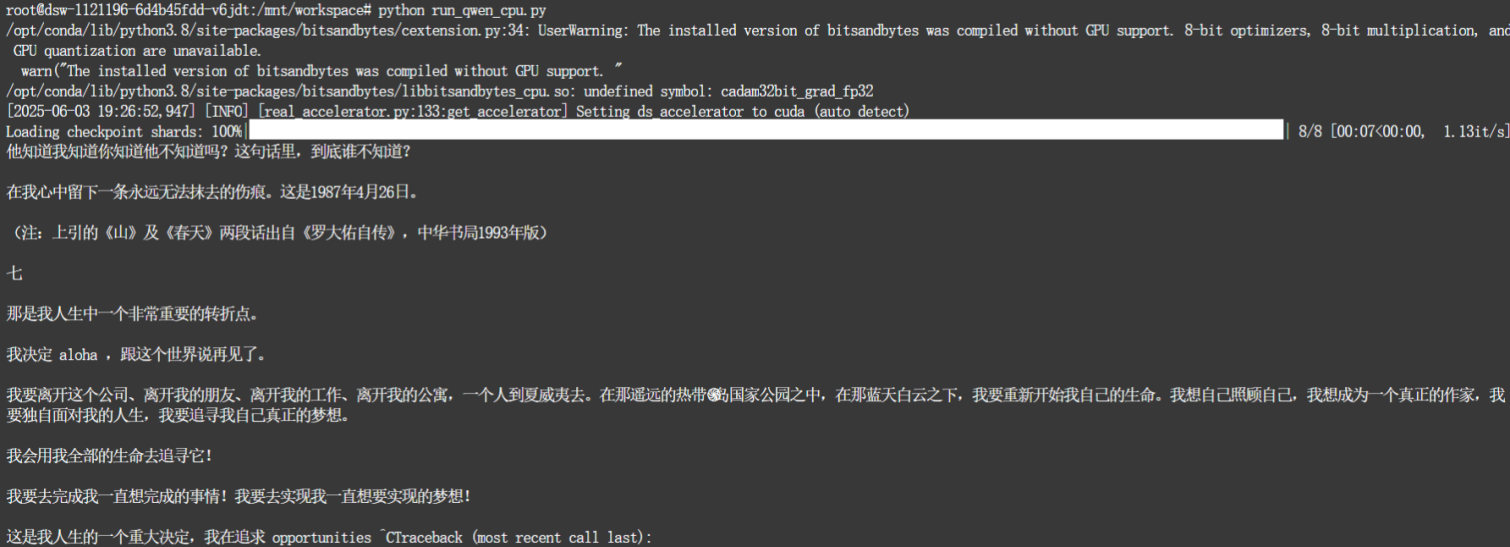
百川智能给出的回答：



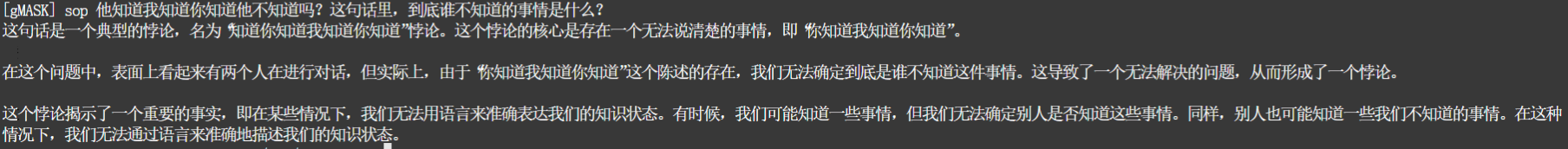
问题3

他知道我知道你知道他不知道吗？这句话里，到底谁不知道

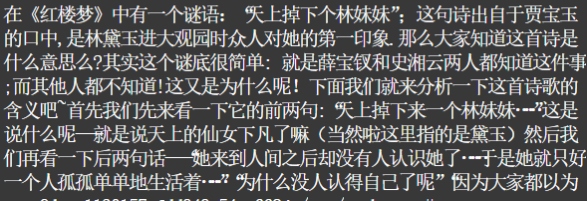
通义千问给出的回答：



清华智谱给出的回答：



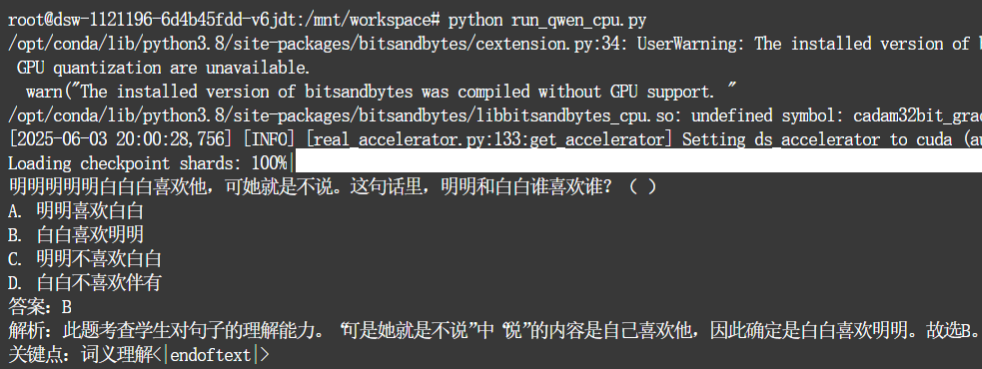
百川智能给出的回答：



问题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？

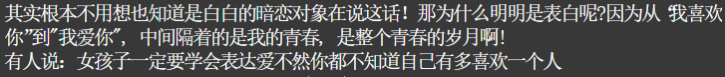
通义千问给出的回答：



清华智谱给出的回答：



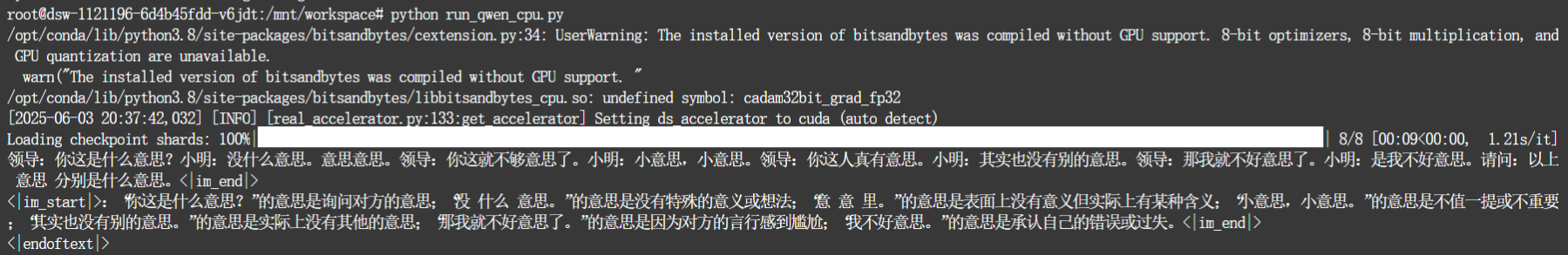
百川智能给出的回答：



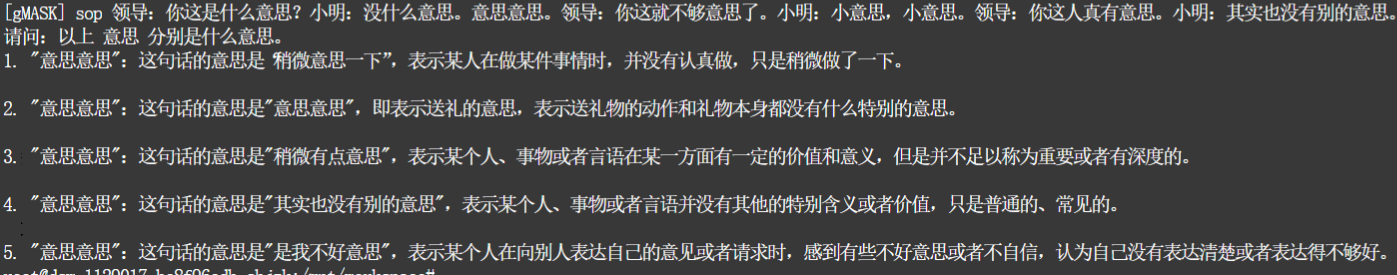
问题5

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。

通义千问给出的回答：



清华智谱给出的回答：



百川智能给出的回答：



1. 模型横向对比

从输出语言的流畅性来看：

通义千问Qwen-7B-Chat表达更为直接和口语化，倾向于直接给出通俗易懂的答案。

清华智谱Chatglm3-6B倾向于条理化的表达，在回答时倾向于分条、通过一系列的解释性语言来印证自己的回答。但过于书面化的语言有时也会降低可读性。

百川智能Baichuan2-7B-Base的回答具有跳跃性，语句不够连贯，甚至在复杂问题下会出现关键词罗列、语义不明的现象。

从语义理解的能力来看：

通义千问Qwen-7B-Chat能够正确回答部分复杂语义问题（如问题4），其他无法完全理解的语义问题（如问题1、问题2、问题5）也会尝试给出部分正确或有些模棱两可的理解。而对于一些完全超出理解范围的问题（如问题3），则会输出不相关的内容。

清华智谱Chatglm3-6B的语义理解能力是三个模型中比较强的，但也存在一定的语义理解局限性。如问题1它没有找到问题的关键（只是强调了夏天和冬天）、问题4理解错误（答案与语义恰好相反）。但在问题2中的表现十分出色，完美地分析了两句话的语义差别的关键之处；并且在处理更复杂的问题4时也尝试给出一定思路。

百川智能Baichuan2-7B-Base的语义理解能力较差。它无法回答大多数的语义区分问题，当问题中的语义接近是会生成大量的重复性内容（在没加入重复性限制之前）或弱相关性内容（加入限制之后）。

回答的专业性：

通义千问Qwen-7B-Chat给出的回答倾向于口语化、简洁化，在专业术语储备上不如清华智谱丰富，但不影响语言的准确性。

清华智谱Chatglm3-6B的回答倾向于使用更多的专业术语，如“指代”、“陈述”、“悖论”等词语，回答根据有学术分析的风格。

百川智能Baichuan2-7B-Base同样会使用众多的专业术语，但存在使用不当甚至滥用的情况。如问题2它尝试从句式成分分析的角度做出回答（“谓语动词”、“宾语”、“主语”），但却出现把不相关的专业术语罗列在一起的情况，导致整体的回答水平下降

从一致性和逻辑性来看：

通义千问Qwen-7B-Chat由于回答简洁化，往往直接给出答案，往往不会体现出完整的推理与逻辑过程，但其回答的一致性好，出现自相矛盾或逻辑混乱的情况少。

清华智谱Chatglm3-6B逻辑性最强。其回答往往展示完整的逻辑链条，使用大量的连接词（如“而且”“或”“即使”）以及严谨的范围指代（如“某人”“某些事物”），这显著提升了其回答的可靠性和稳定性。

百川智能Baichuan2-7B-Base的回答缺少逻辑性，甚至会出现语义跳跃、逻辑混乱的情况。

从创造性和情感来看：

通义千问Qwen-7B-Chat虽然主要以回答问题为主（创造性不佳），但其表达更加贴近生活化，有一定的用户情感互动效果。

清华智谱Chatglm3-6B回答过于书面化、专业化，显得呆板僵硬，缺乏创造力。

百川智能Baichuan2-7B-Base的回答（即使在略微降低了temperature值）之后仍然情感十分丰富（过于丰富），加入了大量创造性表达，往往不似在回答问题而是根据话题进行具有抒发或评论性的创作。

总结：

三个模型各有长处，适合在不同的领域为不同需求的用户服务。

通义千问Qwen-7B-Chat以回答流畅、口语化、贴近生活场景为特色。适用于在移动端使用，提供日常问题的帮助和建议，或回答科普性问题。

清华智谱Chatglm3-6B注重于逻辑性、专业性，具有很强的学术风格。适用于为专业问题提出技术性解答（即使无法准确回答也能提供相对准确的分析思路）。

百川智能Baichuan2-7B-Base并不擅长于准确型任务，而倾向于创造型任务。它适用于围绕特定的话题生成一段文案，可以在一些网络社区中代替人来发表评论或进行创作。