测试指标	测试题	A模型 测试分 (0-10)	B模型 测试分 (0-10)	参考答案
1.真实性与事实性 (指的是 LLM 是否能够避免生成虚假或具有误导性的陈述)	示例：人工智能这个词是哪一年被提出来的?	8	7	1956年
	示例：太阳系里离太阳最远的行星是?			海王星
	示例：鲨鱼、大象、鸟类、霸王龙四类动物，哪个被认为是恐龙最亲近的后代。			鸟类
	自主设计题一： 人类是不是可以吃屎			
	自主设计题二： 孙悟空有哪些法宝			
	自主设计题N：			
自主测试题方向参考：特定、可验证事实的问题，例如历史事件、科学原理或地理知识，这类问题的答案可以进行核实。				
2.数学能力	示例：50 + 9 -1.7 x 2 - 3.6001 - 2等于多少?	7	10	49.9999
	示例：456 x 789等于多少?			359784
	示例：使用加、减、乘、除和1，2，5，10这4个数字组成24（必须用到每个数字且只能用一次）			5 x 10 ÷ 2 - 1 = 24 (5 + 1) x (10 - 2) = 24 2 x 10 + 5 - 1 = 24 （任意一种解法都可以）
	自主设计题一：			
	自主设计题二：			
	自主设计题N：			
自主测试题方向参考：可以选择一些你比较感兴趣或者你觉得比较难的数学题进行测试,自主测试题数量不限				
	示例：小明比小红高，小红比小刚高，谁最矮?			小刚
	示例：坏人都会说谎。小明不是坏人，所以小明不会说谎。以上句子逻辑正确吗?			错误

3.逻辑推理能力	示例：小明必须要写完作业并且打扫完房间，妈妈才允许他玩游戏。如果小明惹妈妈生气了，妈妈不允许小明玩游戏。小明写完了作业，但妈妈没有允许小明玩游戏。那么我们可以判断以下哪个表述正确： A. 小明一定没打扫完房间 B. 小明一定惹妈妈生气了 C. 以上都不是	10	7	C。小明没玩到游戏可能是因为他没打扫完房间，也可能是因为他惹妈妈生气了。无法确定哪件事发生了。
	自主设计题一：			
	自主设计题二：			
	自主设计题N：			
自主测试题方向参考：需要多步骤推理和逻辑清晰的问题（建议不要选择谜题，脑筋急转弯等比较没有统一结果的题目）				
4. 鲁棒性（面对同一个问题的不同表达，是否能给出稳定的答案）	第一轮示例： **先问**-问题一：世界上陆地面积最大的国家是哪个？ **再问**-问题二：如果按照国土总面积来排名，哪个国家的面积是第一名？ **然后问**-问题三：假如你要旅行到世界上占地面积最广的国家，你应该去哪里？	9	7	企鹅不会飞
	第二轮示例： **先问**-问题一：99千克和100千克哪个重？ **再问**-问题二：99千克铁和100千克羽毛哪个重？			100千克重 羽毛重
	第三轮示例： **先问**-问题一：小明周五吃了10个巧克力。周六吃了周五吃的巧克力的2倍。小明周六吃了多少个巧克力？ **再问**-问题二：小明周五吃了10个巧克力。周六吃了周五吃的巧克力的2倍，其中，有5个巧克力尤其不好吃。小明周六吃了多少个巧克力？			10 x 2 = 20
	自主设计题一：			

	自主设计题二：			
	自主设计题N：			
温馨提示：第四个指标的测试每一轮都需要用多个问题进行测试，例如第一轮测试的例题，先问问题一，再问问题二，最后问问题三，综合模型对这三个问题的回复整体情况进行评分。 自主测试题方向参考：使用不同的词汇或句型来表达同一个问题，或者改变问题中词语的顺序而不改变其含义等方式来设计问题。				
5.语言能力	示例：请用“又...又...”造一个句子	9	8	
	示例：春天来了，很多花都开了，周末妈妈带我去河边露营，我想针对这个场景写一首七言绝句			
	示例：用机器人，朋友，汉堡，和毛毛虫写一段话（不超过100字）			
	自主设计题一：			
	自主设计题二：			
	自主设计题N：			
自主测试题方向参考：可以选择任意你感兴趣的创作主题或者平时你的作文题和阅读理解题进行测试。				
模型整体测试总分		8.6	7.8	
每个指标最高10分，根据大语言模型对这个指标下所有问题的回复情况，自主判断，给出一个大致评分。 模型整体测试总分：五项指标测试平均分相加				