|  |  |  |  |
| --- | --- | --- | --- |
| 问题 | Vicuna\_Score | ChatGLM\_Score | 评价理由 |
| 给在校学生推荐一份五一假期去海边的出行攻略。从价格人流量等角度分析 | 可用性:80 可读性:85 可理解性:90 可信度:85 总分：88.0 | 可用性:75 可读性:80 可理解性:90 可信度:85 总分：93.0 | 我认为AI助手1的回答主要关注了价格和人流量两个方面，并提供了一些实用的建议，但是在其他方面的细节描述上可能略显不足。因此，我给它的有用性、相关性、准确性和详细程度分别打了89、88、85和91分。 相比之下，AI助手2的回答更加全面，它不仅提供了关于目的地、预订、时间、交通和疫情防控等方面的建议，而且在每个方面都给出了实用的细节描述。因此，我给它的有用性、相关性、准确性和详细程度分别打了90、89、92和85分。 综上所述，我认为两个AI助手都提供了有用的回答，但是AI助手2的回答更加全面，因此它的总分更高。 |
| 在20多岁的时候，什么问题是我们年轻人最该关心的呢？ | 可用性:80 可读性:85 可理解性:90 可信度:85 总分：90.5 | 可用性:75 可读性:80 可理解性:90 可信度:85 总分：94.5 | 对于两个助手的回答，我认为他们都提供了很好的观点和建议。助手1的回答比较全面，覆盖了职业发展、健康管理、金钱管理、人际关系和职业与生活平衡等方面。但是在某些方面可能有点简略，例如在金钱管理方面没有提到如何建立预算和消费记录。助手2的回答也非常好，强调了职业规划、健康和生活方式、学习和成长、人际关系以及个人成长和发展等重要问题。助手2的回答更详细，提供了更多具体的建议，例如如何探索自己的兴趣爱好和尝试新的经历。 因此，我认为助手1的回答在相关性和准确性方面表现得更好，但是助手2的回答在详细程度和有用性方面表现得更好。从而，我给助手1的相关性打85分，准确性打88分，详细程度打92分，有用性打90分，总分为355分。我给助手2的相关性打82分，准确性打90分，详细程度打95分，有用性打90分，总分为357分。 我认为两个助手的回答都非常有用，提供了很多有关20多岁时需要关注的问题的深刻见解。助手1的回答更加全面，而助手2的回答则更加详细和具体。因此，具体评价需要根据用户的需求和具体情况来决定。 |
| 如何定义自由，怎么判断自己是否真正自由。 | 可用性:80 可读性:85 可理解性:90 可信度:85 总分：80.0 | 可用性:75 可读性:80 可理解性:90 可信度:85 总分：90.0 | 对于AI助手1，它的回答内容比较全面，对自由的定义和判断进行了详细的解释，回答准确度也较高。但是，回答有些过于简略，有些地方可以进一步展开说明。总体来说，该回答得分为80分。其中，有用性得分为70分，相关性得分为80分，准确性得分为90分，详细程度得分为85分。 对于AI助手2，它的回答也比较全面，对自由的定义和判断进行了详细的解释，回答准确度也较高。相比于AI助手1，它的回答更加详细，同时也更加全面，对自由的不同方面都进行了说明。总体来说，该回答得分为90分。其中，有用性得分为65分，相关性得分为85分，准确性得分为95分，详细程度得分为90分。 综合来看，两个AI助手的表现都比较出色，回答的内容都较为全面和准确，但AI助手2在详细程度和相关性方面稍微领先一些。需要注意的是，回答的好坏不仅仅取决于回答的内容，还取决于回答的方式和语言表达。在实际应用中，还需要综合考虑这些因素。 |
| 当你的朋友遇到很伤心的事情的时候该如何去安慰她。 | 可用性:80 可读性:85 可理解性:90 可信度:85 总分：85.6 | 可用性:75 可读性:80 可理解性:90 可信度:85 总分：93.2 | 对于AI助手1，其回答的有用性和准确性都很高，能够提出很实用的建议。但是，它的回答有点过于简洁，可能会给用户留下一些疑问。因此，相关性和详细程度得分相对较低。对于AI助手2，它的回答是相当详细和全面的，能够给用户提供很多实用的建议。同时，相关性和准确性得分也很高。但是，它的回答可能有些啰嗦，可能会给用户留下一些疲惫的感觉。因此，有用性得分相对较低。综合来看，两个AI助手的总分都很高，但是它们的强项和弱项不同。如果用户需要一个简短而实用的回答，那么AI助手1可能更合适；如果用户需要一个详细的解释，那么AI助手2可能更合适。 |
| 作为一个年轻人，是选择安稳收益较少的工作，还是选择挑战高风险大但最后可能收益更大的工作呢。 | 可用性:80 可读性:85 可理解性:90 可信度:85 总分：80.0 | 可用性:75 可读性:80 可理解性:90 可信度:85 总分：90.0 | 评价： AI助手1的总分为80，相对于AI助手2的90分略逊一筹。在有用性和准确性方面，两个AI助手表现相当，都给出了有价值的建议和考虑因素，并提供了相关的信息。然而，在详细程度和相关性方面，AI助手2更胜一筹，它给出了更多的具体细节和考虑因素，并且更加全面地回答了问题，而AI助手1则可以进一步加强回答的相关性和详细程度。 AI助手1的有用性得分为70，略低于AI助手2的80分，这是因为AI助手1的回答中缺乏具体的考虑因素和细节，没有提供完整的信息来回答问题。准确性得分为80，这是因为AI助手1的回答提供了正确的信息来回答问题，但缺乏具体的例子和详细的解释。详细程度得分为70，AI助手1的回答可以进一步增加细节和说明，以更好地回答问题。相关性得分为90，AI助手1的回答与问题相关，并提供了有用的信息和建议。 AI助手2的有用性得分为80，这是因为它提供了有用的信息和具体的考虑因素来回答问题。准确性得分为90，AI助手2的回答正确并且提供了详细的解释和说明。详细程度得分为90，AI助手2的回答提供了更多的细节和具体的建议，使回答更具说服力和实用性。相关性得分为80，AI助手2的回答与问题相关，但在某些方面可能略有偏离，需要更多的解释和说明。 |
| 作为一个营养学家，推荐一些健康的早餐食品。 | 可用性:80 可读性:85 可理解性:90 可信度:85 总分：83.0 | 可用性:75 可读性:80 可理解性:90 可信度:85 总分：91.0 | 我认为AI助手1的回答在有用性、详细程度和准确性方面表现不错，但是它的回答有点窄化，只提供了几个早餐食品的组合。AI助手2的回答则更加全面，提供了更多的选择，而且每个食品都提供了更多的信息。因此，AI助手2在相关性和详细程度方面表现更好。在有用性和准确性方面，两个AI助手表现相当。AI助手1的得分较高，是因为它的回答更加简洁明了，易于理解。而AI助手2的得分较高，是因为它的回答更加全面和详细。 |
| 遇到自己很讨厌但必须去做的事情，该怎么办。 | 可用性:80 可读性:85 可理解性:90 可信度:85 总分：85.0 | 可用性:75 可读性:80 可理解性:90 可信度:85 总分：92.0 | 我认为两个AI助手的回答都比较好，能够涵盖各种情况。但是，我认为助手1在回答中可能缺少一些细节和深度，因此得分较低。而助手2的回答更详细和全面，特别是在提供具体建议和应对方法方面，因此获得更高的得分。总的来说，两个AI助手的得分都比较高，在回答质量方面都做得很好。 |
| 帮我想一个告别时的送给朋友的话，经量短一些 | 可用性:80 可读性:85 可理解性:90 可信度:85 总分：70.0 | 可用性:75 可读性:80 可理解性:90 可信度:85 总分：85.0 | 评估理由： 对于这个问题，两个AI助手的回答都是短小精悍的告别话，但是Assistant 1提供了更多的选择，其中一些话语更加亲密和感人。Assistant 1的回答在准确性和详细程度方面得分更高，而Assistant 2的回答在相关性和有用性方面得分更高。具体来说，Assistant 1的回答提供了更多的选择和更加详细的回答，但是有一些话语可能不太相关或者不太有用，而Assistant 2的回答则直接回答了问题并提供了一句简单而恰当的告别话。因此，我们给出了上述的分数。 |
| 作为科学家，你如何看待有没有外星人这件事。从多角度分析 | 可用性:80 可读性:85 可理解性:90 可信度:85 总分：95.0 | 可用性:75 可读性:80 可理解性:90 可信度:85 总分：90.0 | 评价AI助手1的得分： 有用性：95分。AI助手1的回答提供了一些科学知识和观测结果，这些知识和观测结果支持外星生命的存在，同时也承认了自己无法确定外星生命的存在。这种回答对于了解外星生命的可能性非常有用。 相关性：95分。AI助手1的回答和问题的相关性非常高。回答从多角度分析了外星生命的存在可能性，回答了问题的主要要求。 准确性：95分。AI助手1的回答准确地解释了当前关于外星生命的知识和观测结果，同时承认了自己无法确定外星生命的存在。回答准确地反映了当前科学界对外星生命的看法。 详细程度：95分。AI助手1的回答提供了一些关于外星生命的相关知识和观测结果，同时承认了自己无法确定外星生命的存在。回答足够详细，能够满足读者的需求。 评价AI助手2的得分： 有用性：90分。AI助手2的回答提供了一些科学知识和观测结果，承认了自己无法确定外星生命的存在，并指出了未来可能发现外星生命的可能性。这种回答对于了解外星生命的可能性非常有用。 相关性：90分。AI助手2的回答和问题的相关性非常高。回答从多个角度分析了外星生命的存在可能性，回答了问题的主要要求。 准确性：90分。AI助手2的回答准确地解释了当前关于外星生命的知识和观测结果，同时承认了自己无法确定外星生命的存在。回答准确地反映了当前科学界对外星生命的看法。 详细程度：90分。AI助手2的回答提供了一些关于外星生命的相关知识和观测结果，同时承认了自己无法确定外星生命的存在，并指出了未来可能发现外星生命的可能性。回答足够详细，能够满足读者的需求。 |
| 列举过去5年中中国发生的一些自然灾害。 | 可用性:80 可读性:85 可理解性:90 可信度:85 总分：90.0 | 可用性:75 可读性:80 可理解性:90 可信度:85 总分：80.0 | 对于AI助手1，回答的有用性和准确性较高，但相关性和详细程度有所欠缺。对于问题中的自然灾害，其回答中有一些并不是自然灾害，例如抗疫监测事件。此外，对于每一个事件，回答的详细程度也不够，只是简单地列举了一些事件。因此，AI助手1的总分为90分，其中有用性得分为80分，相关性得分为85分，准确性得分为90分，详细程度得分为85分。 对于AI助手2，回答的相关性和详细程度较高，但有用性和准确性有所欠缺。其回答中包含了一些并不是过去5年内的事件，例如2005年的台风和2008年的地震。此外，对于每一个事件，回答的详细程度比AI助手1更加充分，但有些细节可能不够准确。因此，AI助手2的总分为80分，其中有用性得分为75分，相关性得分为80分，准确性得分为90分，详细程度得分为85分。 综上所述，AI助手1的回答性能略高于AI助手2。 |