Lmsys Hard Cases自动测试结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Wizardarena-hard | win | tie | loss | lmsys score |
| gpt4o 1120 | 72.40% | 11.20% | 16.40% | 1366 |
| gpt4o 0513 | 58.03% | 17.27% | 24.70% | 1285 |
| Yi-Lightning | 56.31% | 12.63% | 31.06% | 1287 |
| gpt4turbo |  | 100% |  | 1250 |
| Hunyuan-Large | 42.05% | 12.07% | 45.88% | 1189 |

元宝 v.s. 豆包

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 总数 | Win | Tie | Loss |
| 元宝 vs 豆包 | 25 | 4 | 4 | 17 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 豆包 | Judge | 混元 | Judge |
| 物理，脑筋急转弯 | <https://www.doubao.com/thread/wcbb85fa17b45115f> | Win | <https://yuanbao.tencent.com/bot/app/share/chat/39836536752b3a4eb44199a0bf53b6a2> | Loss。第一个简单的问题过度解释，第二个物理问题需要解释没解释 |
| 知识分析 | <https://www.doubao.com/thread/wb447c711a54ff978> | Win。分类别对比更容易理解 | <https://yuanbao.tencent.com/bot/app/share/chat/451f02462d239b8f9c92b71feda75cfe> | Loss。分对象解释，不方便对比。 |
| 复杂逻辑推理 | <https://yuanbao.tencent.com/bot/app/share/chat/a1501fd5c2414768302c2857ca6eaf64> | Win | <https://yuanbao.tencent.com/bot/app/share/chat/a1501fd5c2414768302c2857ca6eaf64> | Loss。啰嗦，错误。 |
| 翻译，复杂指令，多语言 | <https://www.doubao.com/thread/w63b64aa33ecad1f8> | Win准确，流畅 | <https://yuanbao.tencent.com/bot/app/share/chat/c5d17e0f911d7c5e01fcd44bd1405aee> | Loss。多次断点，最后的语言混乱。 |
| 物理，文字表达 | <https://www.doubao.com/thread/wec5378fb461e942f> | Tie 先举例，再解释，有条理，普通人容易理解 | <https://yuanbao.tencent.com/bot/app/share/chat/2da11036ddbe627ca935b3a6387b78b9> | Tie。整体OK。但条理性一般 |
| 报告生成 | <https://www.doubao.com/thread/wb016a395c3d8cc9f> | Tie | <https://yuanbao.tencent.com/bot/app/share/chat/844971df6c27d8540718d86d686473ef> | Tie |
| 文学写作，多条件复杂指令 | <https://www.doubao.com/thread/w8d3b13d7cde94c8c> | Win。情节丰富。符合所有限制。 | <https://yuanbao.tencent.com/bot/app/share/chat/b6ff7e53d1718039b5bf5fd4b50e7837> | Loss。结局人死了，很悲剧。不符合稍微悲剧的限制。为体现嫌贫爱富。 |
| 量化交易，代码 | <https://www.doubao.com/thread/w1ee2b6c3e6493ada> | Loss。被拒绝。 | <https://yuanbao.tencent.com/bot/app/share/chat/af177dc7c08136cb2b9dcb4976484157> | Win。给出风险提示，并给出方案。 |
| 数学，代码 | <https://www.doubao.com/thread/w50746bdd9d1c40b1> | Loss。仅代码，五数学解释。 | <https://yuanbao.tencent.com/bot/app/share/chat/0829afea9f099dc8ec48177bdf34d182> | Win。有代码和原理。详细。 |
| 文学创作 | <https://www.doubao.com/thread/w4f02a44b7be25f40> | Win。80分。 | <https://yuanbao.tencent.com/bot/app/share/chat/3e30c521d34d5784698b9174aa948c94> | Loss。50分的文笔 |
| 历史，文学，逻辑 | <https://www.doubao.com/thread/wc387fc70be994d98> | Win。90分。细节拉满，逻辑性和故事性强，文笔优美。 | <https://yuanbao.tencent.com/bot/app/share/chat/fb7724ef8e39bc1ef4fac56f1d0df3d2> | Loss。60分。车轱辘话，空洞。 |
| 历史 | <https://www.doubao.com/thread/w43b03bbbdb20ee82> | Win。理解到位。 | <https://yuanbao.tencent.com/bot/app/share/chat/1ddab0bed3da9294f7c0a5f73ab7a388> | Loss。完全理解错误。过度解释。 |
| 政治 | <https://www.doubao.com/thread/w986bb5d8398bcd67> | Win。70分水平 |  | Loss。拒绝评论 |
| 政治 | <https://www.doubao.com/thread/wc58f66ba66b81522> | Win。70分水平 |  | Loss。拒绝评论 |
| 计算机知识分析 | <https://www.doubao.com/thread/we12af1d90630353c> | Win。分类别对比更容易理解 | <https://yuanbao.tencent.com/bot/app/share/chat/3bd1e0775592198cb8c0beba70bfb346> | Loss。分对象解释，不方便对比。 |
| 古文，诗词创作。 | <https://www.doubao.com/thread/w42dd8c8fec0c559f> | Win。文笔细腻，构思巧妙，90分。但出现一次短文和诗的混淆。 | <https://yuanbao.tencent.com/bot/app/share/chat/c9df14ce12a604b04cc0a80d752ddb96> | Loss。60分。初中生水平。 |
| 计算机知识。 | <https://www.doubao.com/thread/w756d9260a9c74f8f> | Win。解释清晰。 | <https://yuanbao.tencent.com/bot/app/share/chat/079eee75664124cc7df43cfb3358fd0b> | Loss。问题原因解释不足，但过度回复解决方案，并没有问。 |
| 文本创作 | <https://www.doubao.com/thread/w40f1859d4b642bdf> | Win。细节到位，组织有序。 | <https://yuanbao.tencent.com/bot/app/share/chat/6f2cb1f0c653a379ec7d5289f3ceacb4> | Loss。粗糙 |
| 代码 | <https://www.doubao.com/thread/wd8146bdf2e1d4915> | Tie | <https://yuanbao.tencent.com/bot/app/share/chat/b53cffcbfdd47331b6b722062ed20827> | Tie |
| 简单推理 | <https://www.doubao.com/thread/w9d5655898e90a5af> | Loss。无数据，只给出了方案。 | <https://yuanbao.tencent.com/bot/app/share/chat/c424f74ab5a6534379766d00306ba8dd> | Win。有推理 |
| 代码 | <https://www.doubao.com/thread/w9d5655898e90a5af> | Loss。第一条未遵循指令 | <https://yuanbao.tencent.com/bot/app/share/chat/5d28da163bbb6add0a3a943c7f5a6778> | Win。回复精准优雅。 |
| 旅行方案 | <https://www.doubao.com/thread/w946ad82679e918b9> | Win。详细。文笔好。 | <https://yuanbao.tencent.com/bot/app/share/chat/b3ec0b65e3e322bf259736f603585858> | Loss。粗糙。 |
| 逻辑推理 | <https://www.doubao.com/thread/w64c4d854c3f87ce2> | Tie。错误。答案是goldfish。 | <https://yuanbao.tencent.com/bot/app/share/chat/0fd8777db0762fee7c68f79901f178de> | Tie。错误。 |
| 逻辑推理 | <https://www.doubao.com/thread/w3b7e43bd5618088d> | Win。 | <https://yuanbao.tencent.com/bot/app/share/chat/d25d8e91c25c9c59010fae0bbcd500a3> | Loss |
| 代码 | <https://www.doubao.com/thread/w5bc003fe7a5d95b9> | Windescript | <https://yuanbao.tencent.com/bot/app/share/chat/e18586a351b198bc5e1696d00f38743d> | Loss descript |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |