4o new prompt

current scores:0.4794520547945205

Challenging Accuracy:33/87

Moderate Accuracy:17/28

Simple Accuracy:20/31

[0, 1, 2, 3, 4, 5, 6, 7, 10, 11, 12, 15, 17, 18, 19, 24, 26, 27, 28, 31, 32, 35, 40, 41, 42, 43, 44, 45, 47, 49, 50, 51, 52, 55, 56, 59, 60, 61, 62, 64, 65, 67, 68, 69, 71, 74, 75, 76, 77, 78, 81, 82, 84, 86, 89, 91, 92, 97, 99, 100, 102, 103, 104, 109, 110, 112, 113, 114, 129, 130, 131, 132, 133, 134, 136, 140]

current scores:0.5342465753424658

Challenging Accuracy:41/87

Moderate Accuracy:16/28

Simple Accuracy:21/31

[0, 1, 2, 4, 6, 7, 10, 11, 15, 18, 21, 24, 25, 27, 28, 29, 31, 34, 35, 42, 43, 44, 45, 48, 49, 50, 51, 52, 58, 59, 61, 62, 64, 65, 67, 69, 71, 74, 75, 76, 77, 78, 81, 82, 84, 86, 89, 91, 92, 93, 97, 99, 100, 103, 104, 105, 109, 110, 112, 114, 119, 130, 131, 133, 134, 136, 140, 143]

current scores:0.5136986301369864

Challenging Accuracy:38/87

Moderate Accuracy:14/28

Simple Accuracy:23/31

[0, 1, 2, 3, 4, 6, 10, 11, 15, 18, 24, 25, 26, 28, 29, 31, 32, 35, 41, 42, 43, 44, 45, 48, 49, 50, 51, 52, 56, 57, 59, 60, 61, 62, 65, 66, 67, 68, 69, 70, 71, 74, 75, 77, 78, 80, 81, 82, 84, 86, 89, 91, 92, 93, 97, 99, 100, 101, 103, 104, 105, 109, 110, 112, 114, 117, 129, 130, 131, 133, 141]

4o old prompt

current scores:0.5

Challenging Accuracy:33/87

Moderate Accuracy:19/28

Simple Accuracy:21/31

[0, 1, 2, 5, 6, 7, 8, 10, 11, 14, 15, 17, 18, 22, 24, 25, 28, 29, 31, 34, 35, 36, 38, 40, 41, 42, 43, 44, 45, 49, 50, 51, 52, 56, 57, 59, 60, 61, 62, 64, 65, 66, 67, 68, 69, 71, 74, 75, 77, 78, 80, 81, 82, 84, 86, 89, 92, 97, 100, 103, 104, 109, 110, 112, 114, 119, 126, 130, 131, 133, 136, 140, 143]

current scores:0.5

Challenging Accuracy:34/87

Moderate Accuracy:15/28

Simple Accuracy:24/31

[0, 1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 15, 17, 18, 24, 25, 27, 28, 29, 31, 35, 40, 41, 43, 44, 45, 47, 49, 50, 51, 52, 56, 57, 59, 60, 61, 62, 64, 65, 66, 67, 68, 69, 71, 72, 74, 75, 77, 78, 81, 82, 84, 86, 88, 89, 92, 93, 97, 99, 100, 101, 103, 104, 105, 109, 110, 112, 114, 130, 131, 133, 140, 143]

我认为只要能体现

1智能更高的llm在数据集上的正确率比智能更低的llm的正确率更高

2使用prompt之后的llm比使用之前

就可以说明我们的bench的有效性和prompt的有效性

毕竟如果只是要提高正确率的话我完全可以直接降低题目难度。我是觉得4omini的理解能力已经不足以支持继续做研究了，之前