Test result new prompt v1：

4o-mini

current scores:0.4246575342465753

Challenging Accuracy:36/87

Moderate Accuracy:12/28

Simple Accuracy:14/31

[0, 1, 2, 3, 6, 10, 11, 12, 15, 17, 18, 19, 22, 24, 25, 26, 27, 28, 29, 31, 33, 34, 38, 41, 42, 43, 44, 45, 48, 49, 50, 51, 52, 53, 57, 59, 60, 61, 62, 65, 67, 68, 69, 71, 75, 77, 78, 79, 80, 81, 82, 84, 86, 88, 89, 91, 92, 93, 97, 99, 100, 101, 103, 104, 105, 110, 112, 113, 114, 116, 119, 121, 123, 124, 126, 130, 131, 132, 133, 134, 135, 136, 140, 143]

current scores:0.4657534246575342

Challenging Accuracy:37/87

Moderate Accuracy:16/28

Simple Accuracy:15/31

[0, 1, 2, 3, 5, 6, 10, 11, 15, 19, 22, 24, 25, 26, 27, 28, 29, 31, 33, 36, 38, 39, 41, 42, 43, 44, 45, 49, 50, 52, 56, 57, 59, 60, 61, 62, 65, 66, 67, 68, 69, 71, 74, 77, 78, 81, 82, 84, 86, 88, 89, 92, 93, 97, 99, 100, 103, 104, 107, 109, 110, 112, 114, 116, 119, 121, 123, 124, 126, 130, 131, 132, 133, 134, 135, 136, 140, 143]

current scores:0.3972602739726027

Challenging Accuracy:30/87

Moderate Accuracy:14/28

Simple Accuracy:14/31

[0, 1, 2, 3, 5, 6, 7, 9, 10, 11, 12, 15, 18, 19, 22, 24, 25, 26, 27, 28, 29, 31, 33, 36, 38, 40, 41, 42, 43, 44, 45, 47, 49, 50, 51, 52, 56, 57, 58, 61, 62, 64, 65, 66, 67, 68, 69, 71, 73, 74, 77, 78, 81, 82, 84, 86, 87, 88, 89, 92, 93, 97, 98, 99, 100, 102, 103, 104, 109, 110, 112, 114, 116, 117, 121, 123, 124, 126, 130, 131, 132, 133, 134, 135, 136, 139, 140, 143]

current scores:0.4178082191780822

Challenging Accuracy:31/87

Moderate Accuracy:14/28

Simple Accuracy:16/31

[0, 1, 2, 3, 5, 6, 10, 11, 12, 15, 17, 18, 19, 21, 22, 23, 24, 25, 26, 27, 28, 29, 31, 33, 36, 38, 41, 42, 43, 44, 45, 47, 48, 49, 50, 52, 56, 57, 59, 60, 61, 62, 65, 67, 68, 69, 70, 71, 73, 74, 77, 78, 81, 82, 84, 86, 88, 89, 90, 92, 93, 97, 99, 100, 101, 102, 103, 104, 110, 112, 114, 116, 117, 121, 123, 124, 126, 130, 131, 132, 133, 134, 136, 140, 143]

current scores:0.3767123287671233

Challenging Accuracy:26/87

Moderate Accuracy:15/28

Simple Accuracy:14/31

[0, 1, 2, 3, 5, 6, 7, 10, 11, 12, 15, 18, 19, 20, 21, 22, 24, 25, 26, 27, 28, 29, 31, 33, 34, 35, 36, 38, 41, 42, 43, 44, 45, 47, 48, 49, 50, 51, 52, 56, 57, 59, 60, 61, 62, 65, 67, 68, 69, 70, 71, 73, 74, 75, 77, 78, 81, 82, 84, 86, 88, 89, 92, 93, 97, 99, 100, 102, 103, 104, 107, 109, 110, 112, 114, 115, 116, 119, 121, 123, 124, 126, 130, 131, 132, 133, 134, 135, 136, 140, 143]

老师我这周迭代了三轮，修正了数据集，并在此过程中总结了O1也做不出来的题及做不出来的原因，综合之后让o1给我输出了一个prompt方案，在gpt 4o mini上使用之后让平均正确率从36.8%上升到41.6%，不算很明显。但是实验过程中我明确感觉到4o和4o mini对很多概念和语境的以及数据间联系的认知要低于o1，我觉得这些prompt在认知能力更高的语言模型上使用应该会效果更好，所以我下一步考虑先在gpt 4o上做几组实验。但是用4o做实验比较贵，我146题跑一轮就大概要4-5美刀左右，所以想问下你有没有学校提供的api key