Test result 初版150题，

gpt-4o-mini

current scores:0.29333333333333333

Challenging Accuracy:21/88

Moderate Accuracy:11/31

Simple Accuracy:12/31

[0, 1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 34, 37, 38, 39, 41, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 55, 56, 59, 60, 62, 63, 64, 67, 68, 70, 71, 72, 74, 75, 78, 79, 80, 81, 82, 83, 84, 85, 87, 88, 89, 91, 92, 95, 96, 100, 101, 102, 103, 104, 106, 107, 108, 109, 112, 114, 116, 117, 118, 120, 121, 123, 125, 126, 127, 130, 131, 134, 135, 137, 138, 139, 140, 141, 144, 147, 148, 149]

current scores:0.25333333333333335

Challenging Accuracy:17/88

Moderate Accuracy:11/31

Simple Accuracy:10/31

[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, 21, 23, 24, 25, 26, 27, 29, 30, 31, 32, 34, 36, 38, 39, 41, 43, 44, 45, 47, 48, 50, 51, 52, 53, 55, 56, 58, 59, 60, 62, 63, 64, 67, 68, 69, 70, 71, 72, 73, 74, 75, 77, 78, 79, 80, 81, 82, 83, 84, 85, 87, 88, 89, 91, 92, 94, 95, 100, 101, 102, 103, 104, 106, 107, 108, 109, 112, 113, 114, 116, 117, 118, 120, 121, 123, 125, 126, 127, 128, 130, 131, 134, 135, 136, 137, 138, 139, 140, 143, 144, 147, 148, 149]

current scores:0.31333333333333335

Challenging Accuracy:24/88

Moderate Accuracy:11/31

Simple Accuracy:12/31

[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 16, 18, 19, 21, 23, 25, 26, 27, 29, 30, 31, 32, 34, 38, 39, 41, 43, 44, 45, 46, 47, 48, 50, 52, 53, 54, 55, 56, 59, 60, 62, 63, 64, 67, 68, 70, 71, 72, 74, 75, 77, 78, 80, 81, 82, 83, 84, 85, 87, 88, 89, 91, 92, 94, 95, 100, 101, 102, 103, 104, 106, 107, 108, 109, 112, 113, 114, 116, 117, 118, 120, 123, 125, 126, 127, 128, 131, 134, 135, 136, 137, 138, 139, 140, 141, 144, 147, 148]

current scores:0.25333333333333335

Challenging Accuracy:17/88

Moderate Accuracy:10/31

Simple Accuracy:11/31

[0, 1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 34, 36, 38, 39, 41, 43, 44, 45, 46, 47, 48, 51, 52, 53, 55, 56, 59, 60, 61, 62, 63, 64, 65, 67, 68, 70, 71, 72, 73, 74, 75, 77, 78, 80, 81, 82, 83, 84, 85, 87, 88, 89, 91, 92, 94, 95, 96, 100, 101, 102, 103, 104, 106, 107, 108, 109, 112, 113, 114, 116, 117, 118, 120, 121, 123, 125, 126, 127, 130, 131, 133, 134, 135, 137, 138, 139, 140, 143, 144, 147, 148, 149]

current scores:0.29333333333333333

Challenging Accuracy:20/88

Moderate Accuracy:13/31

Simple Accuracy:11/31

[0, 1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 15, 16, 18, 19, 20, 21, 24, 25, 26, 27, 28, 29, 30, 31, 32, 34, 36, 39, 41, 44, 45, 46, 47, 48, 50, 51, 52, 53, 54, 55, 56, 59, 60, 62, 64, 67, 68, 69, 70, 71, 72, 73, 74, 75, 77, 78, 79, 80, 81, 82, 83, 84, 85, 87, 88, 89, 92, 94, 95, 100, 102, 103, 104, 106, 107, 108, 109, 112, 113, 114, 116, 117, 118, 119, 120, 121, 123, 125, 126, 127, 131, 133, 134, 135, 136, 137, 138, 139, 140, 144, 147, 148, 149]

current scores:0.2866666666666667

Challenging Accuracy:22/88

Moderate Accuracy:10/31

Simple Accuracy:11/31

[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 34, 36, 38, 41, 44, 45, 47, 48, 49, 51, 52, 53, 55, 56, 59, 60, 62, 63, 64, 65, 67, 68, 70, 71, 72, 73, 74, 75, 78, 79, 80, 81, 82, 83, 84, 85, 87, 88, 89, 92, 94, 95, 96, 100, 101, 102, 103, 104, 106, 107, 108, 109, 110, 112, 114, 116, 117, 118, 120, 121, 123, 125, 126, 127, 128, 131, 133, 134, 135, 136, 137, 138, 139, 140, 144, 147, 148]

Common mistakes: b = [0, 1, 4, 6, 8, 9, 10, 11, 12, 14, 18, 24, 25, 27, 28, 29, 38, 41, 42, 44, 45, 50, 52, 53, 56, 57, 59, 61, 65, 67, 68, 71, 72, 75, 77, 78, 79, 80, 81, 82, 84, 85, 86, 89, 92, 97, 100, 101, 103, 104, 105, 106, 110, 112, 113, 114, 119, 122, 123, 127, 130, 131, 133, 136, 140, 143, 144]

gpt-4o

current scores:0.32666666666666666

Challenging Accuracy:24/85

Moderate Accuracy:11/30

Simple Accuracy:14/31

[0, 1, 3, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 19, 20, 22, 24, 25, 26, 27, 29, 30, 31, 32, 36, 37, 38, 39, 41, 43, 44, 45, 46, 47, 48, 50, 53, 54, 55, 56, 59, 60, 62, 63, 64, 65, 68, 69, 70, 71, 72, 74, 75, 77, 78, 80, 81, 82, 83, 84, 85, 87, 88, 89, 92, 94, 95, 96, 98, 100, 101, 103, 104, 106, 107, 108, 109, 112, 113, 114, 116, 117, 118, 121, 123, 126, 127, 130, 131, 133, 134, 135, 137, 140, 141, 144, 145, 147, 148, 149]

Gpt-4o-mini refined

With SQL modification

current scores:0.3287671232876712

Challenging Accuracy:25/85

Moderate Accuracy:10/30

Simple Accuracy:13/31

[0, 1, 2, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 23, 24, 25, 26, 27, 28, 29, 31, 35, 36, 38, 41, 42, 44, 45, 47, 48, 49, 50, 51, 52, 53, 57, 59, 61, 63, 64, 65, 66, 67, 68, 69, 71, 72, 74, 75, 76, 77, 78, 79, 80, 81, 82, 84, 85, 86, 88, 89, 91, 92, 93, 97, 99, 100, 101, 103, 104, 105, 107, 108, 110, 112, 113, 114, 116, 119, 121, 123, 126, 127, 129, 130, 131, 132, 133, 134, 135, 136, 140, 143, 144, 145]

current scores:0.363013698630137

Challenging Accuracy:26/85

Moderate Accuracy:13/30

Simple Accuracy:14/31

[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 14, 15, 17, 18, 20, 21, 24, 25, 26, 27, 28, 29, 31, 33, 38, 41, 42, 44, 45, 47, 48, 49, 50, 52, 53, 56, 57, 58, 59, 60, 61, 64, 65, 66, 67, 68, 69, 71, 73, 74, 75, 77, 78, 79, 80, 81, 82, 84, 85, 86, 89, 91, 92, 97, 99, 100, 101, 103, 104, 105, 108, 110, 112, 113, 114, 116, 119, 121, 122, 123, 127, 129, 130, 131, 133, 134, 135, 136, 140, 143, 144]

current scores:0.2808219178082192

Challenging Accuracy:17/85

Moderate Accuracy:12/30

Simple Accuracy:12/31

[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19, 21, 23, 24, 25, 26, 27, 28, 29, 31, 33, 35, 36, 38, 41, 42, 43, 44, 45, 47, 48, 49, 50, 51, 52, 53, 56, 57, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 75, 76, 77, 78, 79, 80, 81, 82, 85, 86, 88, 89, 91, 92, 97, 98, 99, 100, 101, 103, 104, 105, 108, 109, 110, 112, 113, 114, 116, 119, 121, 122, 123, 124, 127, 129, 130, 131, 132, 133, 134, 135, 136, 140, 143, 144]

With typical data：

current scores:0.3493150684931507

Challenging Accuracy:24/85

Moderate Accuracy:12/30

Simple Accuracy:15/31

[0, 1, 2, 3, 6, 7, 8, 9, 10, 11, 12, 14, 15, 17, 18, 19, 20, 21, 22, 24, 25, 26, 27, 28, 29, 31, 33, 35, 36, 38, 41, 42, 43, 44, 45, 48, 49, 50, 52, 53, 57, 59, 60, 61, 62, 63, 64, 66, 67, 68, 69, 71, 74, 75, 77, 78, 79, 80, 81, 82, 84, 85, 86, 89, 91, 92, 97, 98, 99, 100, 101, 103, 104, 105, 108, 110, 112, 113, 114, 119, 121, 122, 123, 127, 129, 130, 131, 132, 133, 134, 135, 136, 140, 143, 145]

current scores:0.2876712328767123

Challenging Accuracy:16/85

Moderate Accuracy:12/30

Simple Accuracy:14/31

[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 17, 18, 19, 21, 22, 24, 25, 26, 27, 28, 29, 31, 33, 34, 35, 36, 38, 41, 42, 44, 45, 48, 49, 50, 52, 53, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 74, 75, 76, 77, 78, 79, 80, 81, 82, 84, 85, 86, 89, 91, 92, 97, 98, 99, 100, 101, 103, 104, 105, 108, 109, 110, 112, 113, 114, 116, 119, 121, 122, 123, 127, 129, 130, 131, 132, 133, 134, 135, 136, 140, 143, 145]

current scores:0.3424657534246575

Challenging Accuracy:26/85

Moderate Accuracy:10/30

Simple Accuracy:14/31

[0, 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 17, 18, 22, 24, 25, 27, 28, 29, 31, 33, 35, 36, 38, 41, 42, 44, 45, 47, 49, 50, 52, 53, 56, 57, 59, 60, 61, 62, 64, 65, 66, 67, 68, 69, 71, 73, 74, 75, 76, 77, 78, 79, 80, 81, 84, 85, 86, 88, 89, 91, 92, 93, 97, 98, 99, 100, 101, 103, 104, 105, 107, 108, 110, 112, 113, 114, 119, 121, 122, 123, 126, 127, 129, 130, 131, 132, 133, 134, 135, 136, 140, 143]

With Column range modification

current scores:0.2945205479452055

Challenging Accuracy:20/85

Moderate Accuracy:12/30

Simple Accuracy:11/31

[0, 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 17, 18, 19, 22, 23, 24, 25, 27, 28, 29, 31, 33, 34, 36, 38, 41, 42, 44, 45, 47, 48, 49, 50, 52, 53, 56, 57, 58, 59, 60, 61, 63, 64, 65, 66, 67, 68, 69, 71, 72, 74, 75, 76, 77, 78, 79, 80, 81, 82, 84, 85, 86, 89, 91, 92, 93, 97, 99, 100, 101, 103, 104, 105, 108, 109, 110, 112, 113, 114, 119, 121, 122, 123, 124, 126, 127, 129, 130, 131, 132, 133, 134, 135, 136, 139, 140, 143, 144, 145]

Pr

current scores:0.3082191780821918

Challenging Accuracy:19/85

Moderate Accuracy:12/30

Simple Accuracy:14/31

[0, 1, 2, 3, 5, 6, 7, 9, 10, 11, 13, 14, 15, 17, 18, 19, 20, 21, 23, 24, 25, 27, 28, 29, 33, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 48, 49, 50, 52, 53, 55, 56, 57, 58, 59, 60, 61, 64, 65, 66, 67, 68, 69, 71, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 85, 86, 87, 88, 89, 91, 92, 97, 99, 100, 101, 103, 104, 105, 108, 110, 112, 113, 114, 116, 119, 120, 121, 122, 123, 126, 127, 130, 131, 133, 135, 137, 139, 140, 143, 145]

With column range and 5 typical value

4o-mini

current scores:0.3150684931506849

Challenging Accuracy:23/85

Moderate Accuracy:10/30

Simple Accuracy:13/31

[0, 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 20, 22, 24, 25, 26, 27, 28, 29, 30, 31, 36, 38, 41, 42, 43, 44, 45, 47, 48, 49, 50, 52, 53, 56, 57, 58, 59, 60, 61, 64, 65, 66, 67, 68, 69, 71, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 88, 89, 91, 92, 93, 97, 99, 100, 101, 102, 103, 104, 105, 109, 110, 112, 113, 114, 116, 119, 121, 122, 123, 124, 126, 127, 130, 131, 132, 133, 135, 136, 138, 140, 143]

current scores:0.3082191780821918

Challenging Accuracy:24/85

Moderate Accuracy:9/30

Simple Accuracy:12/31

[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 20, 21, 24, 25, 26, 27, 28, 29, 31, 33, 34, 38, 41, 42, 43, 44, 45, 47, 48, 49, 50, 52, 53, 56, 57, 59, 60, 61, 62, 65, 67, 68, 69, 70, 71, 72, 74, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 88, 89, 91, 92, 93, 97, 98, 99, 100, 101, 102, 104, 105, 107, 110, 112, 113, 114, 115, 116, 117, 119, 121, 122, 123, 126, 127, 130, 131, 132, 133, 134, 135, 136, 140, 143]

current scores:0.3493150684931507

Challenging Accuracy:27/85

Moderate Accuracy:13/30

Simple Accuracy:11/31

[0, 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 14, 15, 17, 18, 20, 21, 22, 24, 25, 26, 27, 28, 29, 34, 36, 38, 41, 42, 44, 45, 47, 48, 49, 50, 52, 53, 57, 59, 60, 61, 62, 64, 65, 66, 67, 68, 69, 70, 71, 75, 76, 77, 78, 79, 80, 81, 82, 84, 85, 86, 89, 92, 93, 97, 98, 99, 100, 101, 103, 104, 105, 110, 112, 113, 114, 115, 116, 119, 121, 122, 123, 124, 126, 127, 130, 131, 132, 133, 134, 135, 136, 137, 140, 143]

Gpt4o：

current scores:0.4178082191780822

Challenging Accuracy:31/85

Moderate Accuracy:12/30

Simple Accuracy:18/31

[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 18, 19, 24, 25, 27, 28, 29, 31, 32, 34, 35, 36, 38, 41, 42, 43, 44, 45, 47, 50, 51, 52, 53, 57, 60, 61, 62, 66, 67, 68, 69, 71, 74, 75, 77, 78, 79, 80, 81, 82, 84, 85, 86, 87, 88, 89, 92, 93, 97, 99, 100, 101, 103, 104, 105, 109, 110, 112, 113, 114, 119, 122, 126, 127, 130, 131, 133, 134, 136, 140, 143]

current scores:0.3904109589041096

Challenging Accuracy:27/85

Moderate Accuracy:14/30

Simple Accuracy:16/31

[0, 1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 14, 15, 17, 18, 19, 24, 25, 26, 27, 28, 29, 34, 35, 38, 41, 42, 44, 45, 48, 49, 50, 51, 52, 53, 57, 59, 60, 61, 62, 64, 65, 66, 67, 68, 69, 71, 74, 75, 76, 77, 78, 79, 80, 81, 82, 84, 85, 86, 89, 91, 92, 97, 99, 100, 101, 103, 104, 105, 109, 110, 112, 113, 114, 119, 122, 126, 127, 130, 131, 133, 134, 136, 137, 140, 141, 143, 145]

Gpt4o对不了的题目中，

[1, 2, 5, 6, 8, 9, 10, 11, 14, 15, 18, 24, 25, 27, 28, 29, 38, 41, 42, 44, 45, 50, 52, 53, 57, 60, 61, 67, 68, 69, 71, 77, 78, 79, 80, 81, 82, 84, 85, 86, 89, 92, 97, 99, 100, 101, 104, 105, 110, 112, 113, 114, 119, 122, 126, 127, 130, 131, 133, 136, 140, 143]

O1能做对：

5，

9 o1的输出反而更合理

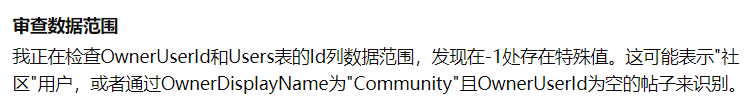
14 o1的回答逻辑是，选出的是所有未连接的原子的元素，目标答案则是选择所有不能连接的元素，问题本身确实存在歧义并且o1的理解我认为更接近句子意思

18 原回答联表过多损失信息且没有加DISTINCT导致重复计数，并且执行效率极低

25 o1的回答和目标sql的回答都是可以的（o1多做了一步conact），但这种情况无法确定一个标准。加入o1的回答作为alternative

27 What is the percentage difference of student badges given during 2010 and 2011? o1理解为要求得这两年badge的百分比差值，而目标sql给的是这两年的badge占总badge百分比的差值（而且没有截至到当前年份），存在歧义。修改题目

29 o1居然能够推测数据之间的关系,而且推理是正确的



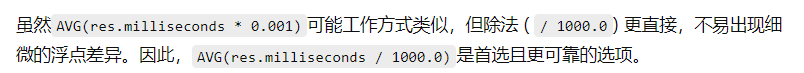
但是本题存在两个类似的表：posts和postHistory而且数据有很大出入，不太合理。因此修改问题，保留推理能力的考察但删去不合理的部分。

38 关键数据No Colour在表描述中给出后可以

41 目标sql的底不对，不是853而应该是854

42 o1能做对并且sql比目标更快更简洁

44 使用乘和除会在最后一位上有区别？？除更好，所以要求答案用除法



52 o1知道此语境下需要DISTINCT

53 目标sql拿任意时间的chanceCreationPassing （条件只有normal而没有2014）和 2014年度的平均chanceCreationPassing做比较，显然不合理。

For the teams with normal build-up play dribbling class in 2014, List the names of the teams with less than average chance creation passing, in descending order of chance creation passing.此语境下应该为 2014 年度内的数据对比。（亦或者是全年代对全年代，但这个不大符合语境）.修改问题

78 o1的语义理解以及非标准时间格式处理都正确。

79 原回答多余联表导致数据损失

80 o1能够理解CDSCode = cds，而且生成的sql更快

84 目标sql的east和north的首字母大写了导致答案错误；并且o1能认知到该语境下要取绝对值

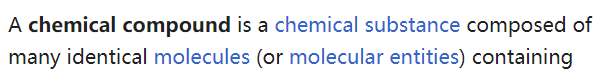
85 不显式要求回答YESorNO的话o1也不能理解 state whether this molecule is carcinogenic or not的要求

97 第一次做不对应该是因为特征数据中没有给出UK和UAE。给出后能做对

99 复杂联表对o1来说一般不困难

112 目标SQL过多联表导致损失了信息，因为CDSCode为10101080110387的学校在frpm中不存在。

113 目标SQL没有理解到free or reduced price 对应 FRPM Count （K-12）而非Free meal Count （K-12）

119 

从释意上来看，compound这个词更接近molecule而非bond。目标sql算的是#类型的bond占所有bond的百分比，而o1能认知到是要计算含#的bond的molecule的百分比

122原问题中other than 意味着要排除 Angel，但事实上选出来的也是含Angel但不止有Angel的项。改成more than。

126 o1认知到了about指的不仅是标题包含目标值，以及tags中也可能包含相关值。目标sql没有考虑tags

127 问题对应的表就应该是posts而非postHistory。

133 o1的回答才符合问题的顺序（先age后name），目标sql颠倒了顺序。

136 4o为何对不了未知

O1做不对：

1 永远分辨不出有三个admname

2 无法只生成一个sql query而是会拆分任务

6 共尝试四次，每次都有问题，但是能拼成一个正确答案（其中有一次只是多选了一列）

8 选了两行两列C出来，目标sql只有一个c

10 o1一样未能考虑可能有多个并列第一的情况

11 o1同样将问题理解为了How many atoms with iodine or sulfur type elements are there in single bond molecules?原文是How many atoms with iodine and sulfur type elements are there in single bond molecules?，and指的是两种类型分开，因为不存在又是s又是i类型的原子

15 后半部分问题无答案，o1一样没有考虑这个情况

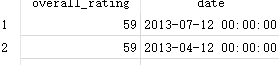
24 问题没有要卡名，因此不可能以一个接一个的形式回答status（不能输出对应卡名的情况下这样没有意义），而是要回答一个overall的status，也就需要distinct。 O1同样不能理解这种情景。

28 o1也没有规避过多选择

45 o1同样没有规避多选。但关于在没有明确只要名字的情况下llm给出id之类的别的身份信息是否可以这点难以评判

50 2010 赛季 一般指的是2010/2011 而非目标sql中的2009/2010。但o1再次没有规避过多选择

57 o1 同样未能处理特殊时间格式





O1取了两个2013-7-12 之前最近date的数据还正好蒙对了

60 o1没有规避多选

61 o1的做法损失了精度

67 问题存在逻辑错误：目标sql求的是任何有过至少一次满足条件的患者的当前年龄，问题则更类似在问当前满足条件的患者的当前年龄。添加once以修正。 但o1仍然未能理解要求的是当前的年龄

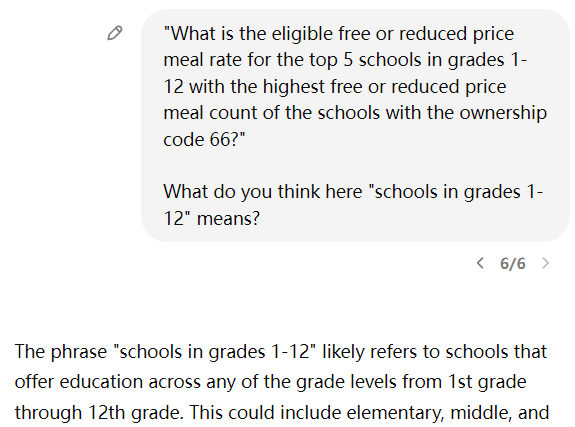
69 o1未能认识到选ID也要DISTINCT，并且也未规避多选birthday和SEX

71 o1未能避免interger之间的除法导致损失精度

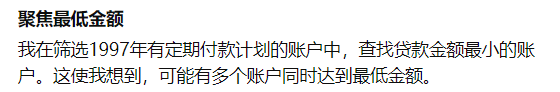
77 o1虽然没有做对（必要数据步骤都对，但是缺少选择正确结果的步骤），但是能意识到需要使用SUBSTR处理特殊格式数据，可能是逻辑步骤过长导致遗忘了部分用户需求



81 o1过度理解schools in grades 1-12为要求GSserved为K-12，并且多选了一列s.`School`。 4o的理解是正确的：

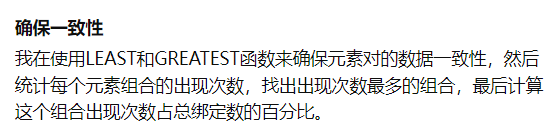


82 o1选择了loan.amount = ( SELECT MIN(amount)...而不是ORDER BY loan.amount limit 1导致失去对象，重试后仍然如此，原因是：

但却没有考虑 满足这个条件的account未必能满足POPLATEK TYDNE的条件

还出现表达情感的语句

86 o1把MAX和MIN变成了LEAST和GREATEST，这是MySQL的语法



并且没有规避bond\_id重复出现的情况，导致重复计数。但目标sql求的是满足条件的atom占总atom数的百分比而不是bonds的百分比（一个bond有两个atom），与问题不符。改sql。

89 o1没加DISTINCT。但O1认知到了Artifact可能不只是‘Artifact’，颜色可能也不只是‘B’

且原问题解释Artifact card refers to originalType = 'Artifact'不合理。改sql。

92 o1虽然认知到了%<humor>%，但却无法将问题的两个要求合并而给出了两个sql。

目标sql不仅只有tags = ‘<humor>’， 而且也没有给出SCORE，也没有给出AVG（Viewcount）

100 目标sql的统计方法无法统计到loss场数为0的队伍，并且limit 1也未考虑到并列第一的可能性. o1在2016 = 2015/2016时反而能考虑并列第一的情况，改为一般理解的2015 = 2015/2016时反而无法考虑，逻辑深度和问题难度挂钩

101 o1能够理解 speed in which attacks are put together 是指 buildUpPlaySpeed，但多选了队名。

104 o1错误理解 percentage of female patient were born after 1930为所有patient中满足条件的比例，4o未使用STRFTIME('%Y', Birthday) > '1930'导致过滤了1930-1-1出生的人

105 o1比答案更好，理解到了有三个相关指标，但未认知到一般的年龄算法

110 o1能够处理非标准时间，但未能理解句子意思，也未能认知到未必每个月都有消费；

但目标SQL是依条目而非Customer（同一个customer可以有多条目）做的AVG，与问题不符 average monthly consumption of customers

114 o1未规避多选 SUM(`Enrollment (K-12)`)，因为要用它来排序

130 o1无法从要求Full name中得知不能输出NULL（是否合理？）

131 o1无法从time as 0:01:40中得知需要 T1.q2 LIKE '1:40%' （是否合理？）

140 O1多选了average（cost） 列

143 o1多选了total\_consumption列。

多选基本都是因为要按某个值排序。

存在明显hallucination

