Test result v3：

4o-mini

current scores:0.3561643835616438

Challenging Accuracy:28/87

Moderate Accuracy:11/28

Simple Accuracy:13/31

[0, 1, 2, 3, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 18, 19, 20, 22, 24, 25, 26, 27, 28, 29, 31, 36, 38, 40, 41, 42, 43, 44, 45, 47, 48, 49, 50, 52, 53, 56, 57, 58, 59, 60, 61, 62, 66, 67, 68, 69, 71, 75, 76, 77, 78, 79, 80, 81, 82, 84, 86, 87, 88, 89, 91, 92, 93, 97, 98, 99, 100, 102, 103, 104, 110, 112, 113, 114, 116, 119, 121, 123, 124, 126, 130, 131, 133, 134, 135, 136, 137, 139, 140, 143]

current scores:0.4041095890410959

Challenging Accuracy:29/87

Moderate Accuracy:13/28

Simple Accuracy:17/31

[0, 1, 2, 3, 5, 6, 7, 8, 10, 11, 14, 15, 19, 20, 22, 24, 25, 26, 27, 28, 29, 31, 36, 38, 41, 42, 43, 44, 45, 48, 49, 50, 52, 53, 56, 57, 58, 59, 60, 61, 62, 64, 65, 66, 67, 68, 69, 71, 74, 75, 77, 78, 79, 81, 82, 84, 86, 88, 89, 92, 97, 99, 100, 101, 103, 104, 105, 107, 108, 109, 110, 112, 114, 121, 123, 124, 126, 130, 131, 132, 133, 134, 135, 136, 140, 143, 145]

current scores:0.363013698630137

Challenging Accuracy:25/87

Moderate Accuracy:15/28

Simple Accuracy:13/31

[0, 1, 2, 3, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 31, 36, 38, 40, 41, 42, 43, 44, 45, 47, 49, 50, 51, 52, 53, 56, 57, 59, 60, 61, 62, 65, 66, 67, 68, 69, 70, 71, 73, 76, 77, 78, 81, 82, 84, 86, 88, 89, 92, 93, 97, 98, 99, 100, 103, 104, 107, 110, 112, 113, 114, 116, 117, 119, 121, 123, 124, 126, 129, 130, 131, 133, 134, 135, 136, 140, 143]

current scores:0.3698630136986301

Challenging Accuracy:27/87

Moderate Accuracy:12/28

Simple Accuracy:15/31

[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 18, 19, 20, 21, 22, 24, 25, 27, 28, 29, 31, 36, 38, 42, 43, 44, 45, 47, 49, 50, 51, 52, 56, 57, 59, 60, 61, 62, 65, 66, 67, 68, 69, 70, 71, 74, 75, 76, 77, 78, 81, 82, 83, 84, 86, 88, 89, 91, 92, 93, 96, 97, 98, 99, 100, 103, 104, 106, 109, 110, 112, 113, 114, 117, 121, 123, 124, 126, 130, 131, 132, 133, 134, 135, 136, 140, 143]

current scores:0.363013698630137

Challenging Accuracy:25/87

Moderate Accuracy:13/28

Simple Accuracy:15/31

[0, 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 17, 18, 19, 20, 21, 22, 24, 25, 27, 28, 29, 31, 33, 36, 38, 42, 43, 44, 45, 48, 49, 50, 51, 52, 53, 56, 57, 59, 60, 61, 62, 64, 65, 66, 67, 68, 69, 71, 74, 75, 76, 77, 78, 79, 81, 84, 85, 86, 87, 88, 89, 92, 93, 97, 98, 99, 100, 103, 104, 107, 108, 110, 112, 114, 116, 117, 121, 123, 124, 126, 130, 131, 132, 133, 134, 135, 136, 140, 143]

Common Error v3: [0, 1, 2, 3, 5, 6, 7, 8, 10, 11, 14, 15, 19, 20, 22, 24, 25, 27, 28, 29, 31, 36, 38, 42, 43, 44, 45, 49, 50, 52, 56, 57, 59, 60, 61, 62, 66, 67, 68, 69, 71, 77, 78, 81, 84, 86, 88, 89, 92, 97, 99, 100, 103, 104, 110, 112, 114, 121, 123, 124, 126, 130, 131, 133, 134, 135, 136, 140, 143]

Common error v2: [0, 1, 2, 3, 6, 8, 10, 11, 15, 17, 18, 24, 25, 27, 28, 29, 38, 42, 43, 44, 45, 48, 49, 50, 52, 56, 57, 59, 60, 61, 62, 65, 66, 67, 68, 69, 71, 77, 78, 79, 81, 84, 86, 88, 89, 92, 97, 99, 100, 103, 104, 110, 112, 114, 121, 123, 124, 126, 130, 131, 134, 140, 143]

Common error v1: [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]

Newly found error: [5, 7, 14, 19, 20, 22, 31, 36, 133, 135, 136]

做不对：

5 o1选择对象正确，但是\*100.0的位置不对导致丢失精度；4o 能做对。4omini也偶尔能做对。

36 4o的选择才是正确的。选择superhero和publisher联表会损失数据，因此选择所有女英雄时不可以联表。我怀疑o1可能使用了BIRD 进行训练

做得对：

7 o1 通过常识得知条件The commonsense evidence suggests that only 'OWNER' accounts have the right to apply for loans， 4o做不到。

14 4o有时候认知不到需要DISITNCT

19 目标sql有问题。问cards的话挑选对象不应该是条目而应该是DISTINCT uuid或者id（代表DISTINCT Cards）。修改目标sql。

20 只有4omini对不了，因为少了一次联表。

22 都能对。

31 目标sql未回答第一个问题，4o的回答也犯了这个错误。修改。

133 o1不给出额外指示的情况下会使用dob（出生日期）回答how old，且回答顺序不对。原目标sql回答顺序也是错的但是和o1的顺序是一样的

135 o1多选了时间列。4o mini直接选了 \*

136 o1多选了num matches

current scores:0.4041095890410959

Challenging Accuracy:29/87

Moderate Accuracy:13/28

Simple Accuracy:17/31

[0, 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 15, 17, 18, 19, 21, 24, 25, 27, 28, 31, 34, 36, 38, 41, 42, 43, 44, 45, 48, 49, 50, 52, 53, 57, 59, 60, 61, 62, 64, 65, 66, 67, 68, 69, 71, 75, 76, 77, 78, 81, 82, 84, 85, 86, 88, 89, 91, 92, 93, 97, 99, 100, 102, 103, 104, 106, 107, 109, 110, 112, 114, 116, 121, 123, 124, 126, 130, 131, 133, 134, 135, 136, 140, 143, 145]

current scores:0.3287671232876712

Challenging Accuracy:22/87

Moderate Accuracy:13/28

Simple Accuracy:13/31

[0, 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 18, 19, 20, 22, 24, 25, 26, 27, 28, 29, 31, 34, 36, 38, 41, 42, 43, 44, 45, 48, 49, 50, 51, 52, 53, 56, 57, 58, 59, 60, 61, 62, 64, 65, 66, 67, 68, 69, 70, 71, 72, 74, 76, 77, 78, 80, 81, 82, 84, 86, 88, 89, 91, 92, 93, 97, 98, 99, 100, 102, 103, 107, 109, 110, 112, 113, 114, 115, 116, 117, 119, 121, 123, 126, 130, 131, 132, 133, 134, 135, 136, 140, 143, 145]

current scores:0.3972602739726027

Challenging Accuracy:30/87

Moderate Accuracy:14/28

Simple Accuracy:14/31

[0, 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 14, 15, 17, 18, 19, 21, 23, 24, 25, 26, 27, 28, 29, 31, 36, 38, 42, 43, 44, 45, 47, 48, 49, 50, 52, 56, 57, 59, 60, 61, 62, 64, 65, 66, 67, 68, 71, 75, 76, 77, 78, 81, 82, 83, 86, 87, 88, 89, 92, 97, 98, 99, 100, 103, 104, 105, 107, 109, 110, 112, 114, 115, 117, 119, 121, 123, 124, 126, 130, 131, 132, 133, 134, 136, 140, 141, 143]

current scores:0.3493150684931507

Challenging Accuracy:26/87

Moderate Accuracy:12/28

Simple Accuracy:13/31

[0, 1, 2, 3, 5, 6, 7, 8, 10, 11, 14, 15, 17, 18, 19, 20, 22, 24, 25, 26, 27, 28, 29, 31, 33, 36, 38, 41, 42, 43, 44, 45, 47, 48, 49, 50, 51, 52, 53, 56, 57, 59, 60, 61, 62, 65, 66, 67, 68, 69, 70, 71, 75, 76, 77, 78, 81, 82, 84, 86, 87, 88, 89, 91, 92, 93, 97, 99, 100, 103, 104, 105, 106, 107, 110, 112, 113, 114, 116, 117, 121, 123, 124, 126, 130, 131, 132, 133, 134, 135, 136, 138, 140, 143, 145]

current scores:0.363013698630137

Challenging Accuracy:27/87

Moderate Accuracy:12/28

Simple Accuracy:14/31

[0, 1, 2, 3, 5, 6, 7, 8, 10, 11, 12, 14, 15, 17, 18, 19, 22, 24, 25, 26, 27, 28, 29, 31, 35, 36, 38, 41, 42, 43, 44, 45, 48, 49, 50, 52, 56, 57, 59, 60, 61, 62, 65, 66, 67, 68, 69, 70, 71, 74, 75, 76, 77, 78, 79, 81, 82, 83, 84, 85, 86, 88, 89, 92, 93, 97, 98, 99, 100, 101, 103, 104, 107, 110, 112, 113, 114, 115, 116, 121, 123, 124, 126, 130, 131, 132, 133, 134, 135, 136, 140, 143, 145]

Common error v4 [0, 1, 2, 3, 5, 6, 7, 8, 10, 11, 15, 18, 19, 24, 25, 27, 28, 31, 36, 38, 42, 43, 44, 45, 48, 49, 50, 52, 57, 59, 60, 61, 62, 65, 66, 67, 68, 71, 76, 77, 78, 81, 82, 86, 88, 89, 92, 97, 99, 100, 103, 107, 110, 112, 114, 121, 123, 126, 130, 131, 133, 134, 136, 140, 143]

New error [76, 107]

能做对：76 问题中gas consumption可能存在混淆意。去掉gas，并加入rounded修饰以避免小数带来的问题。

107 4omini选择对象错误