Test result v2：

4o-mini

current scores:0.3767123287671233

Challenging Accuracy:29/87

Moderate Accuracy:12/28

Simple Accuracy:14/31

[0, 1, 2, 3, 5, 6, 7, 8, 10, 11, 12, 15, 17, 18, 19, 21, 24, 25, 26, 27, 28, 29, 31, 36, 38, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 52, 53, 56, 57, 59, 60, 61, 62, 65, 66, 67, 68, 69, 70, 71, 77, 78, 79, 81, 82, 83, 84, 86, 88, 89, 92, 93, 97, 98, 99, 100, 102, 103, 104, 105, 110, 112, 114, 115, 116, 117, 121, 123, 124, 126, 127, 130, 131, 132, 133, 134, 135, 136, 140, 143, 145]

current scores:0.4246575342465753

Challenging Accuracy:33/87

Moderate Accuracy:13/28

Simple Accuracy:16/31

[0, 1, 2, 3, 5, 6, 7, 8, 10, 11, 14, 15, 17, 18, 22, 24, 25, 27, 28, 29, 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, 80, 81, 82, 84, 86, 88, 89, 92, 97, 99, 100, 102, 103, 104, 105, 107, 110, 112, 114, 116, 117, 121, 123, 124, 126, 130, 131, 133, 134, 135, 136, 140, 143]

current scores:0.3767123287671233

Challenging Accuracy:27/87

Moderate Accuracy:12/28

Simple Accuracy:16/31

[0, 1, 2, 3, 5, 6, 7, 8, 10, 11, 12, 14, 15, 17, 18, 19, 20, 22, 24, 25, 26, 27, 28, 29, 31, 33, 38, 41, 42, 43, 44, 45, 48, 49, 50, 51, 52, 53, 56, 57, 59, 60, 61, 62, 65, 66, 67, 68, 69, 71, 74, 76, 77, 78, 79, 80, 81, 84, 85, 86, 87, 88, 89, 91, 92, 93, 97, 99, 100, 101, 102, 103, 104, 109, 110, 112, 113, 114, 115, 116, 117, 119, 121, 123, 124, 126, 130, 131, 134, 140, 143]

current scores:0.3698630136986301

Challenging Accuracy:28/87

Moderate Accuracy:10/28

Simple Accuracy:16/31

[0, 1, 2, 3, 6, 7, 8, 10, 11, 12, 15, 17, 18, 19, 22, 24, 25, 27, 28, 29, 31, 35, 36, 38, 41, 42, 43, 44, 45, 47, 48, 49, 50, 52, 53, 56, 57, 58, 59, 60, 61, 62, 64, 65, 66, 67, 68, 69, 70, 71, 73, 74, 75, 77, 78, 79, 80, 81, 82, 84, 86, 87, 88, 89, 91, 92, 93, 97, 99, 100, 101, 103, 104, 106, 107, 110, 112, 114, 119, 121, 123, 124, 126, 130, 131, 132, 133, 134, 135, 136, 140, 143]

current scores:0.3767123287671233

Challenging Accuracy:29/87

Moderate Accuracy:11/28

Simple Accuracy:15/31

[0, 1, 2, 3, 5, 6, 8, 10, 11, 12, 15, 17, 18, 19, 22, 24, 25, 27, 28, 29, 31, 36, 38, 42, 43, 44, 45, 47, 48, 49, 50, 52, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 73, 75, 76, 77, 78, 79, 80, 81, 82, 84, 85, 86, 88, 89, 91, 92, 93, 97, 98, 99, 100, 101, 103, 104, 107, 110, 112, 113, 114, 121, 123, 124, 126, 130, 131, 132, 133, 134, 135, 136, 137, 140, 143]

Error: [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]

error before: [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]

Shared error: [1, 2, 6, 8, 10, 11, 15, 18, 24, 25, 27, 28, 29, 38, 42, 44, 45, 50, 52, 57, 60, 61, 67, 68, 69, 71, 77, 78, 79, 81, 84, 86, 89, 92, 97, 99, 100, 104, 110, 112, 114, 126, 130, 131, 140, 143]

Old error only: [5, 9, 14, 41, 53, 80, 82, 85, 101, 105, 113, 119, 122, 127, 133, 136]

New error only: [0, 3, 17, 43, 48, 49, 56, 59, 62, 65, 66, 88, 103, 121, 123, 124, 134]

O1 能对：

0 能对，4o不能分辨不同表中的相似列之间的区别（选择了frpm中的charter funding type而不是school中的funding type）

43 问题不严谨，race completion是包含了+n laps的结果的，但目标sql没有考虑。O1能够考虑到这一点。改问题。4o则连目标都搞错了。SELECT d.driverId, d.forename, d.surname,

48 复杂查询对o1不困难。它甚至没有用给出的position而是用数据自己算了一遍

不能对：

49 原问题未阐明要以秒而非毫秒作答，并且目标sql过于复杂。明确in seconds后没有问题

56 o1能应对给出的样本的特殊时间格式做出调整，4o做不到

62 问题不严谨。目标sql算的并非病人的比例，而是病例的比例（同个病人被多次统计），并且表意不明。加入once并修改目标sql

65 目标sql统计对象并非patient而是case（同一个patient ID 被重复计数），且未加入once明确语义。4o则无法处理，只能选择case。

66 o1能认知到此题需要DISTINCT，但4o不行

88 o1和4o都能对，只有4omini对不了，它没有能够处理NULL数据。

121 4o错将id而不是uuid当作了联表标志，说明o1对表描述信息的利用能力比4o更强。

123 4o错将id而不是code和setcode当作联表标志。说明o1对表描述信息的利用能力比4o更强。

124 4o和o1都能对，4omini对不了，联表条件和查询条件全错。

134 4o将between 1990 and 1995 理解成了 1990 -1-1 到 1995-12-31.

对不来：

3 o1错误将Intermediate/Middle Schools and Unified Scools 和 K-12 schools混淆了，并且未认知到unified schoo的含义（但目标sql的2010-12-31 和 between 2009 and 2010不符）

17 o1多选了一列elements， 4o则是忘加DISTINCT

59 o1认为给出的表描述文件有误，因为根据一般常识它认为UA越高越不正常，然而表描述文件中给出的UA是越低越不正常。它没有遵循表描述文件，即便指出有问题的地方让它再次思考它也坚持认为是人的打字错误。4o也表现出这个问题

103 o1多选了用作排序的avg（heading\_accuracy）列