**calendar\_Plan\_DeepSeek-R1 (100% correct within 5 iterations of evaluation)**

|  |  |
| --- | --- |
| **Number of Iterations** | **Count** |
| 1 | 100 |
| **Grand Total** | **100** |

**100% at pass 1**

**meeting\_Plan\_DeepSeek-R1 (100% correct within 5 iterations of evaluation)**

|  |  |
| --- | --- |
| **Number of Iterations** | **Count** |
| 1 | 96 |
| 2 | 4 |
| **Grand Total** | **100** |

Error analysis:

pass 1:

4 wrong plan (all wrong in number of people met, met 1 less person than gold answer)

Pass 2:

0 wrong plan (all correct)

**trip\_Plan\_DeepSeek-R1 (79% correct within 5 iterations of evaluation)**

|  |  |
| --- | --- |
| **Number of Iterations** | **Count** |
| 1 | 74 |
| 2 | 4 |
| 3 | 1 |
| 5 | 21 |
| **Grand Total** | **100** |

Error analysis:

pass 1:

26 wrong plan

pass 2:

22 wrong plan

pass 3:

21 wrong plan

pass 4:

21 wrong plan

pass 5:

21 wrong plan

Trip planning typical error: misunderstanding of available flight as bidirectional, hallucination about available flight (that is actually not available), gap in days (flight day not shared, say day 1 in one city, day 2 in a different city) and misunderstanding of gap error feedback as overlap feedback, forget about some minimum stay constraint requirement in the process