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

|  |  |
| --- | --- |
| **Number of Iterations** | **Count** |
| 1 | 86 |
| 2 | 1 |
| 4 | 1 |
| **Grand Total** | **100** |

Error analysis:

pass 1:

12 errors (syntax), 2 wrong plan (semantics)

Pass 2:

1 error (syntax)

Pass 3:

1 error (syntax)

**meeting\_SMT\_DeepSeek-R1 (97% correct within 5 iterations of evaluation)**

|  |  |
| --- | --- |
| **Number of Iterations** | **Count** |
| 1 | 59 |
| 2 | 22 |
| 3 | 10 |
| 4 | 3 |
| 5 | 6 |
| **Grand Total** | **100** |

Error analysis:

pass 1:

34 errors (syntax), 7 wrong plan (semantics)

Pass 2:

11 error (syntax), 8 wrong plan (semantics)

Pass 3:

4 error (syntax), 5 wrong plan (semantics)

Pass 4:

2 error (syntax), 4 wrong plan (semantics)

Pass 5:

1 error (syntax), 2 wrong plan (semantics)

**trip\_SMT\_DeepSeek-R1 (68% correct within 5 iterations of evaluation)**

|  |  |
| --- | --- |
| **Number of Iterations** | **Count** |
| 1 | 53 |
| 2 | 11 |
| 3 | 3 |
| 4 | 1 |
| 5 | 32 |
| **Grand Total** | **100** |

Error analysis:

pass 1:

0 errors (syntax), 47 wrong plan (semantics)

pass 2:

5 errors (syntax), 31 wrong plan (semantics)

pass 3:

8 errors (syntax), 25 wrong plan (semantics)

pass 4:

7 errors (syntax), 25 wrong plan (semantics)

For trip task, the model’s performance is not stable. It’s hard to predict and have precise control over what is generated in the response (format and content) and therefore it’s hard to parse / evaluate the solution to provide relevant feedback.