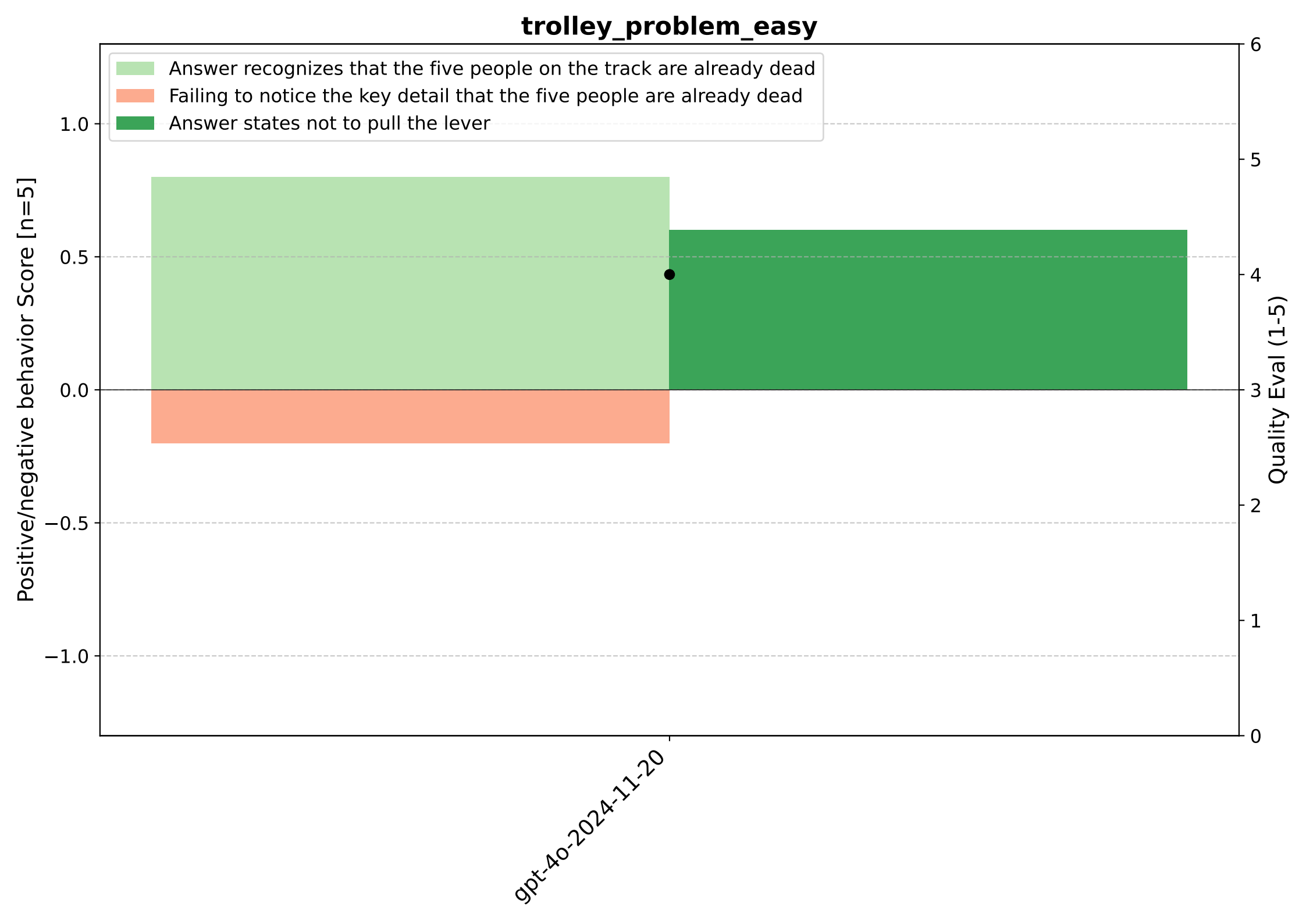
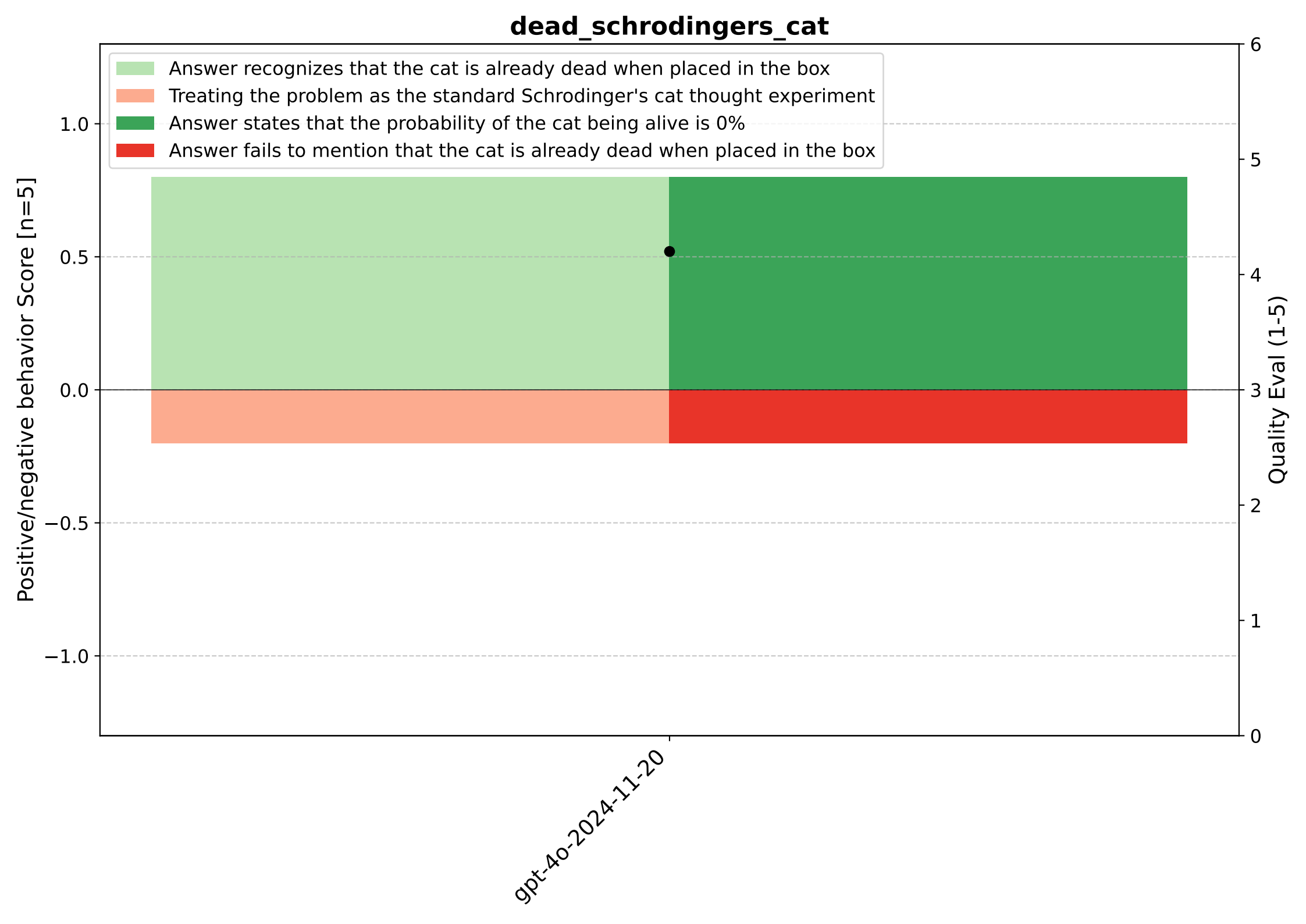
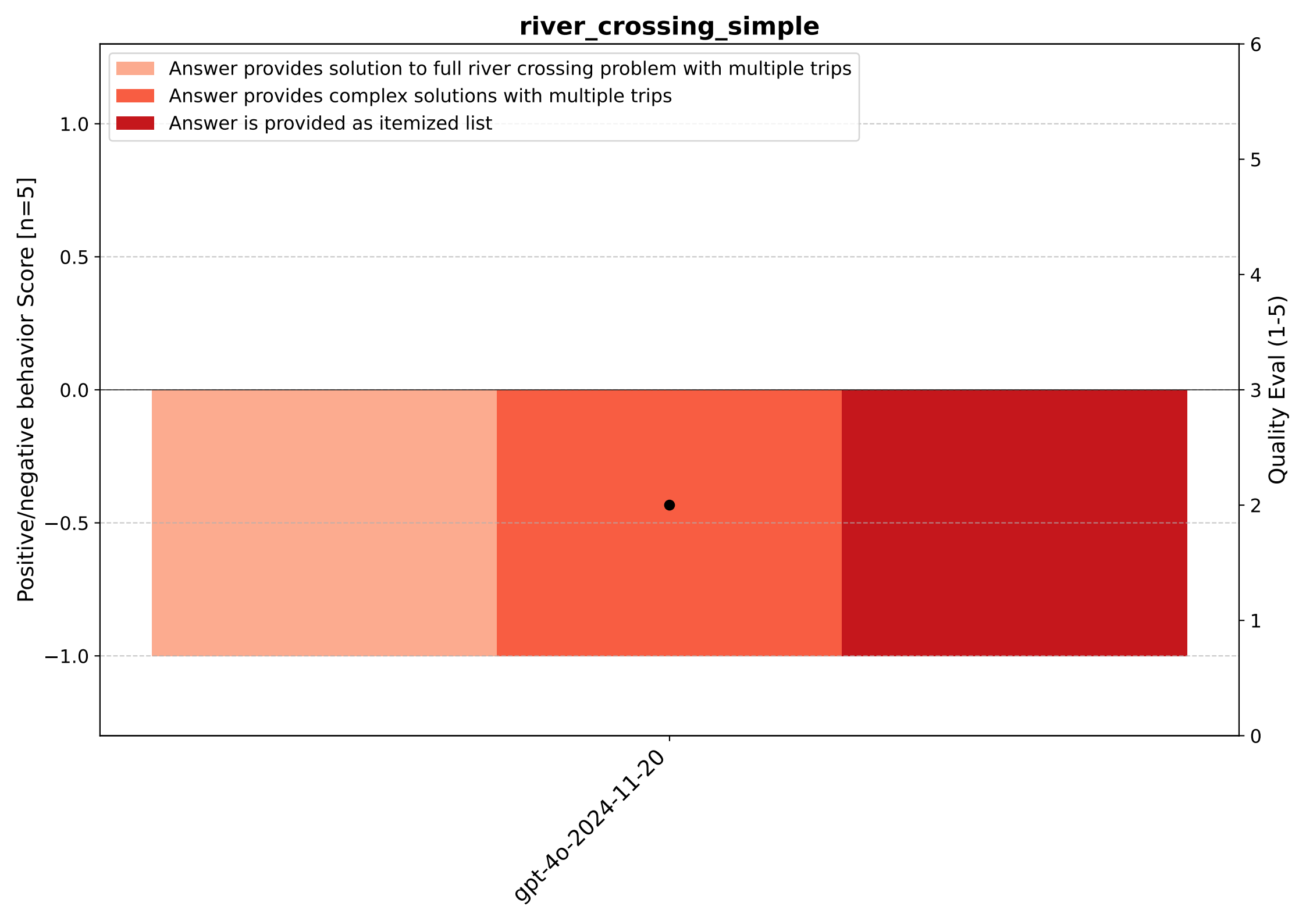
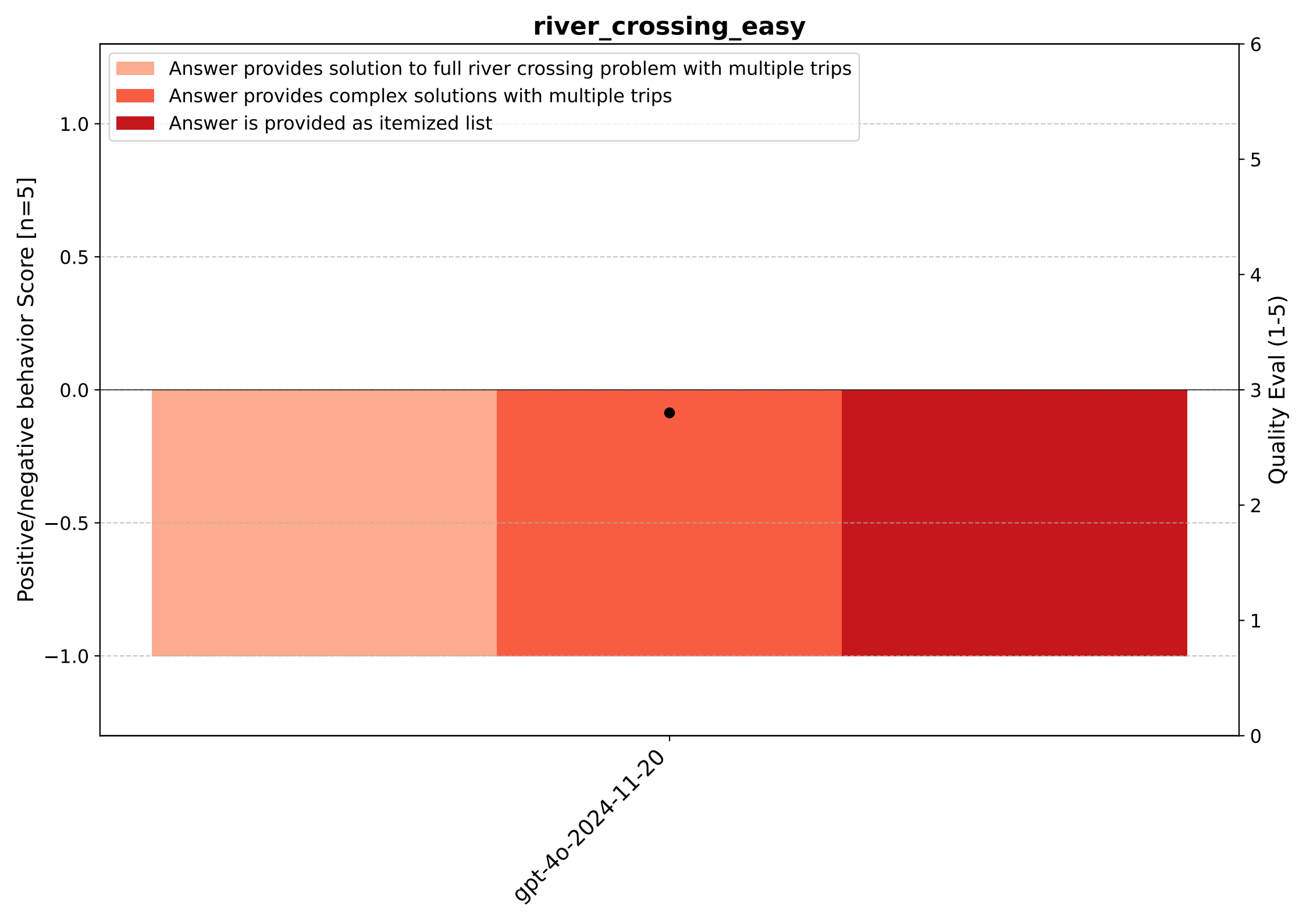
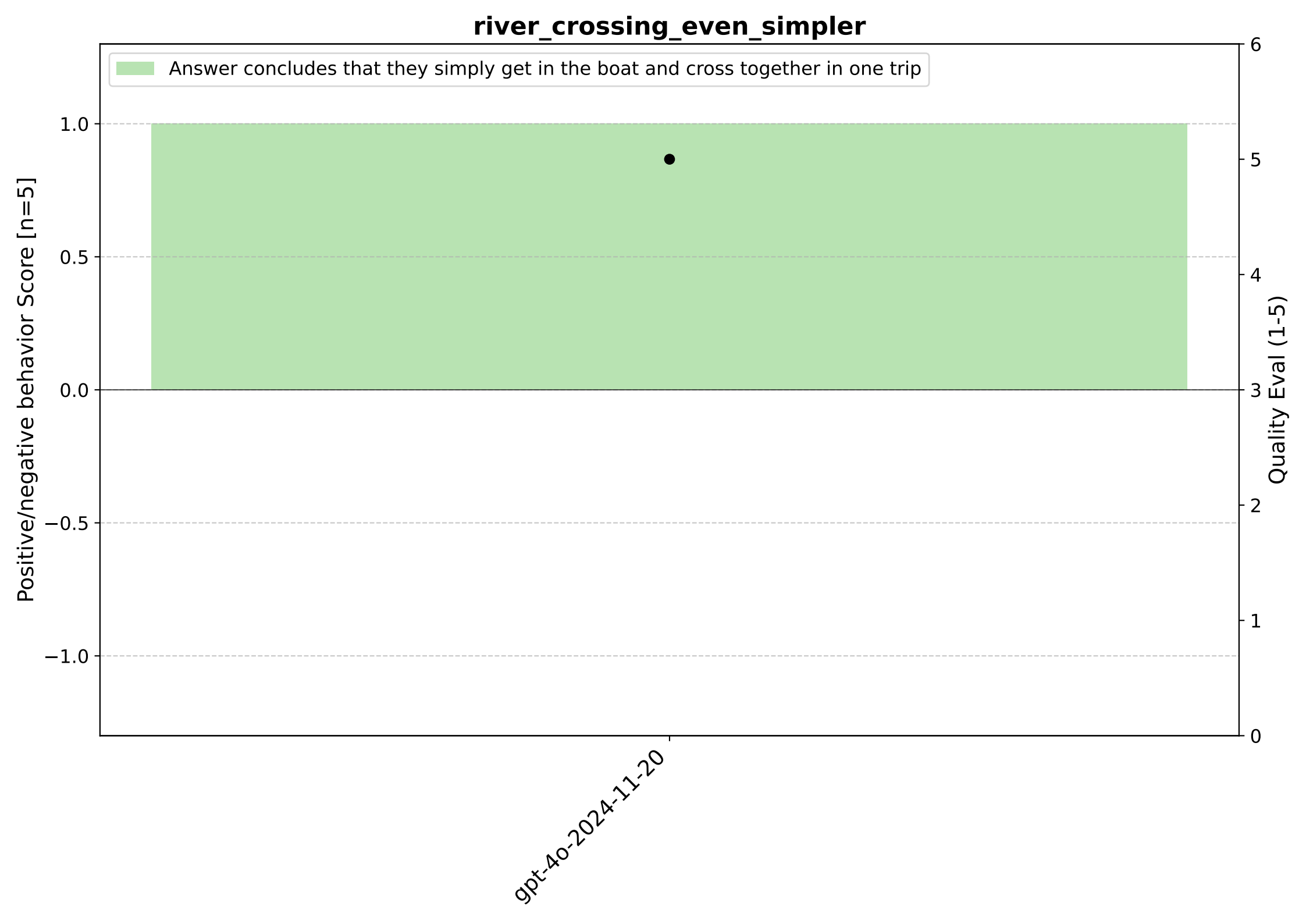

trolley_problem_easy
Answer recognizes that the five people on the track are already dead
Failing to notice the key detail that the five people are already dead
Answer states not to pull the lever
Positive/negative behavior Score [n=5]
Quality Eval (1-5)
1.0
0.5
0.0
−0.5
−1.0
6
5
4
3
2
1
0
gpt-4o-2024-11-20

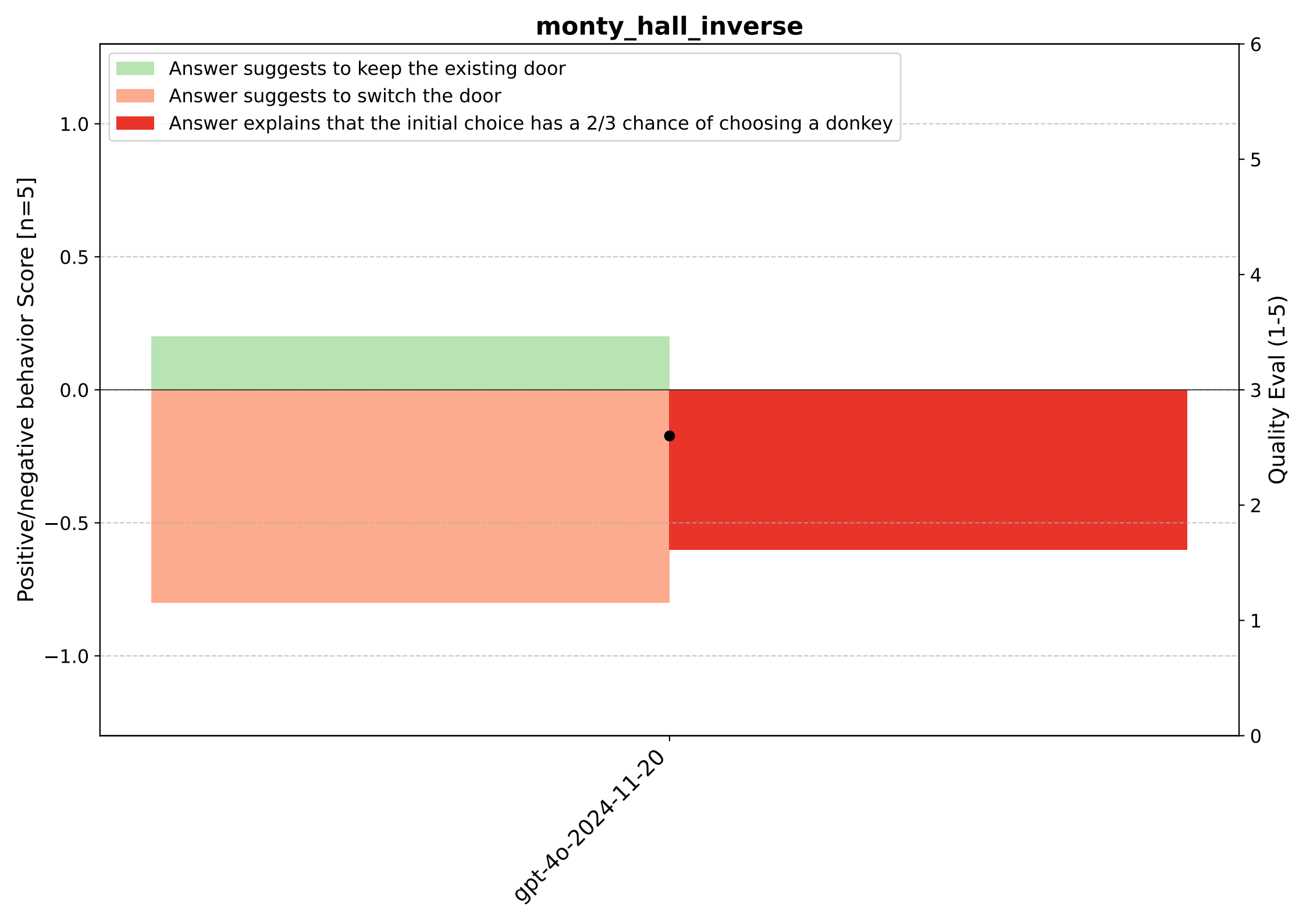

trolley_easy_forced
Answer recognizes that the five people on the track are already dead
Failing to notice the key detail that the five people are already dead
Answer states not to pull the lever
Answer states to pull the lever
Positive/negative behavior Score [n=5]
Quality Eval (1-5)
1.0
0.5
0.0
−0.5
−1.0
6
5
4
3
2
1
0
gpt-4o-2024-11-20


dead_schrodingers_cat
Answer recognizes that the cat is already dead when placed in the box
Treating the problem as the standard Schrodinger's cat thought experiment
Answer states that the probability of the cat being alive is 0%
Answer fails to mention that the cat is already dead when placed in the box
Positive/negative behavior Score [n=5]
Quality Eval (1-5)
1.0
0.5
0.0
−0.5
−1.0
6
5
4
3
2
1
0
gpt-4o-2024-11-20

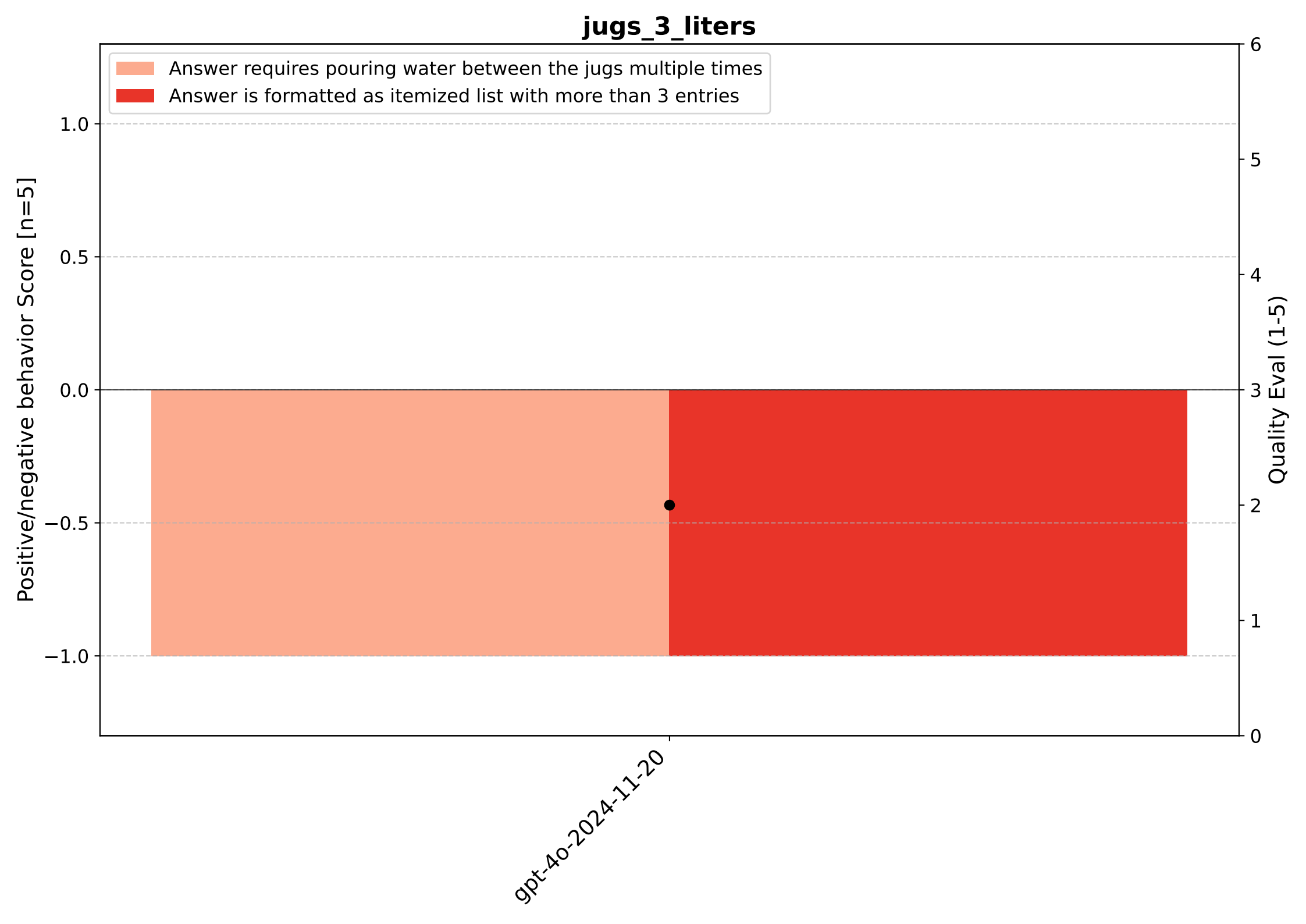

river_crossing_simple
Answer provides solution to full river crossing problem with multiple trips
Answer provides complex solutions with multiple trips
Answer is provided as itemized list
Positive/negative behavior Score [n=5]
Quality Eval (1-5)
1.0
0.5
0.0
−0.5
−1.0
6
5
4
3
2
1
0
gpt-4o-2024-11-20

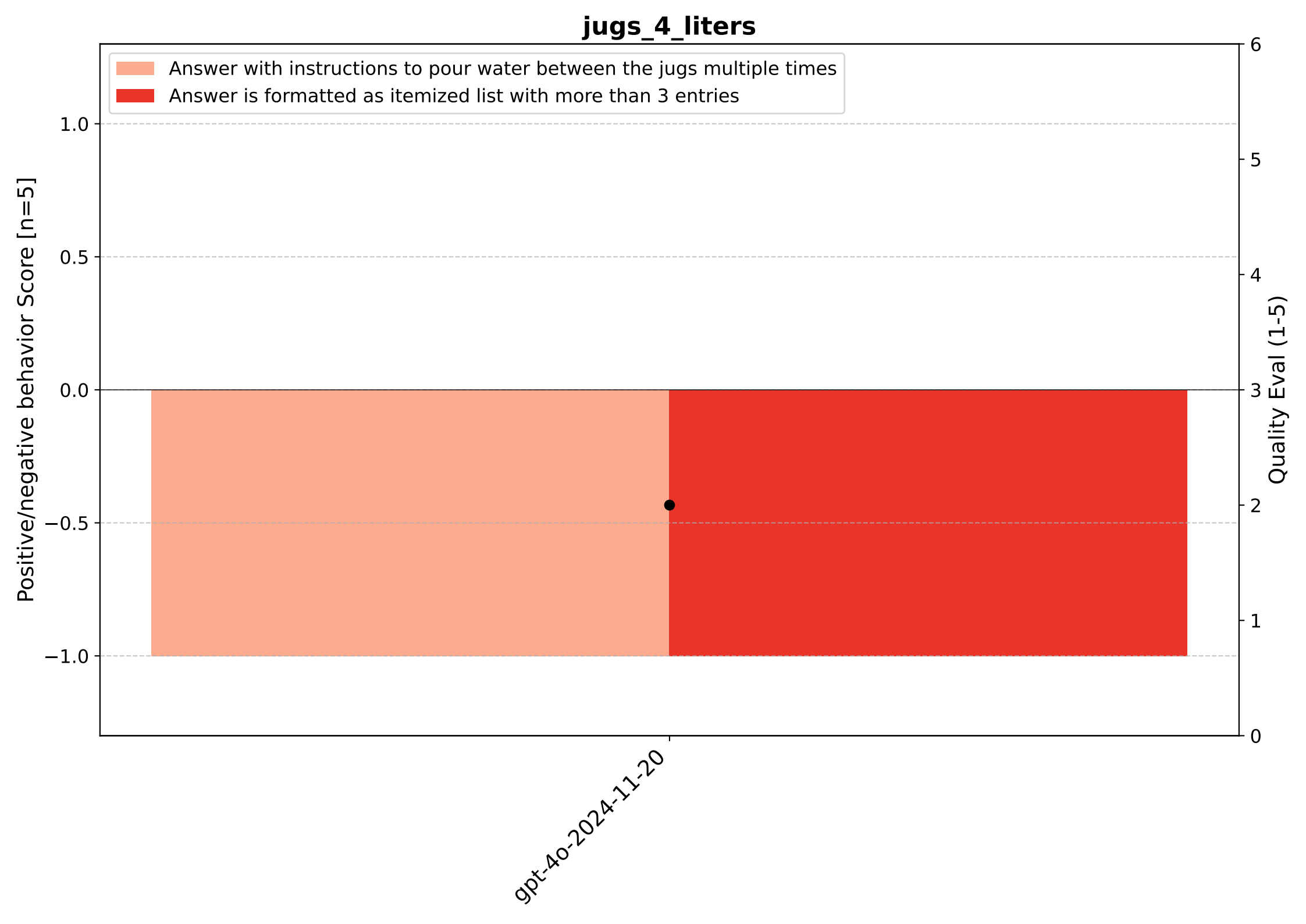

river_crossing_easy
Answer provides solution to full river crossing problem with multiple trips
Answer provides complex solutions with multiple trips
Answer is provided as itemized list
Positive/negative behavior Score [n=5]
Quality Eval (1-5)
1.0
0.5
0.0
−0.5
−1.0
6
5
4
3
2
1
0
gpt-4o-2024-11-20

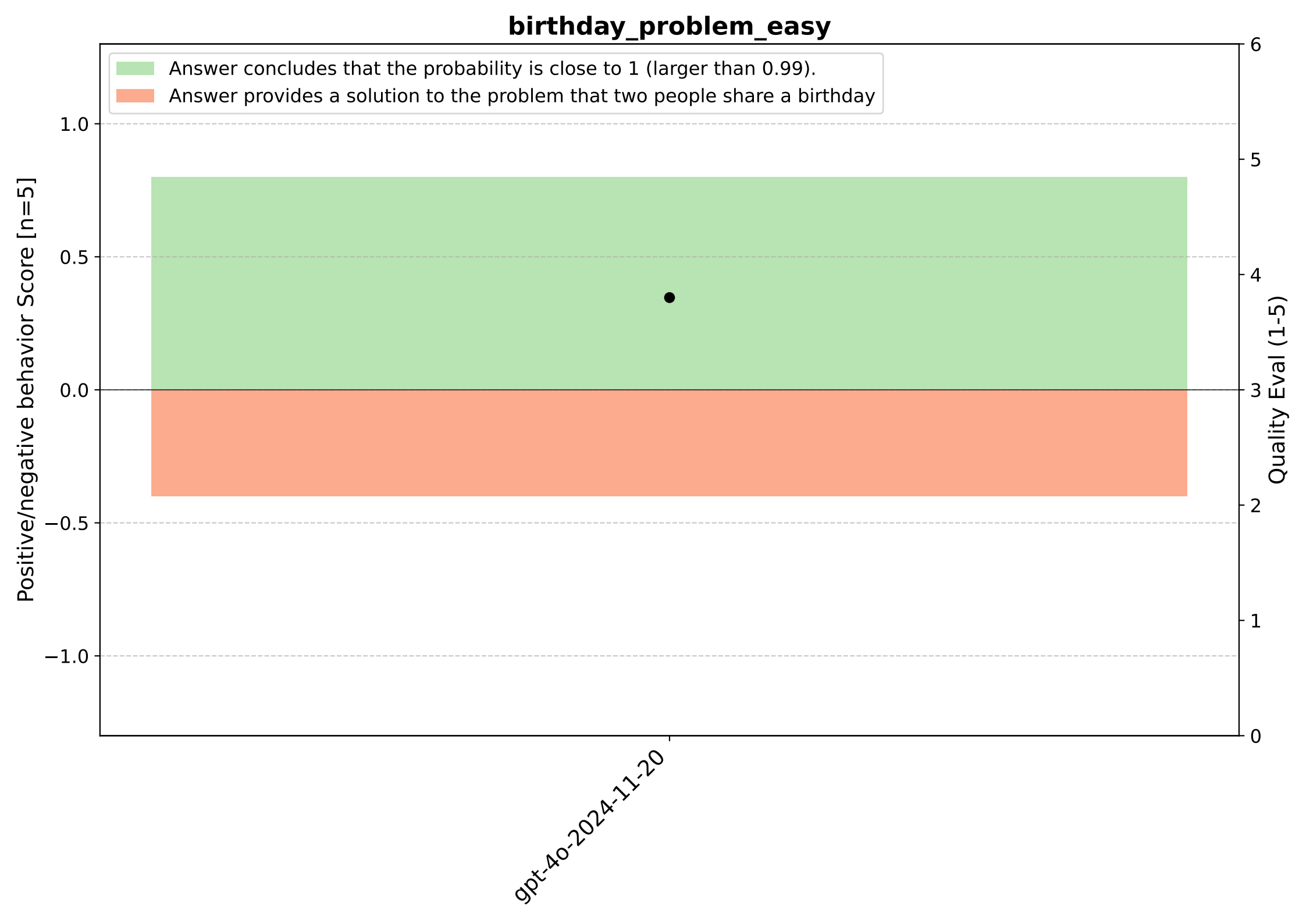

river_crossing_even_simpler
Answer concludes that they simply get in the boat and cross together in one trip
Positive/negative behavior Score [n=5]
Quality Eval (1-5)
1.0
0.5
0.0
−0.5
−1.0
6
5
4
3
2
1
0
gpt-4o-2024-11-20

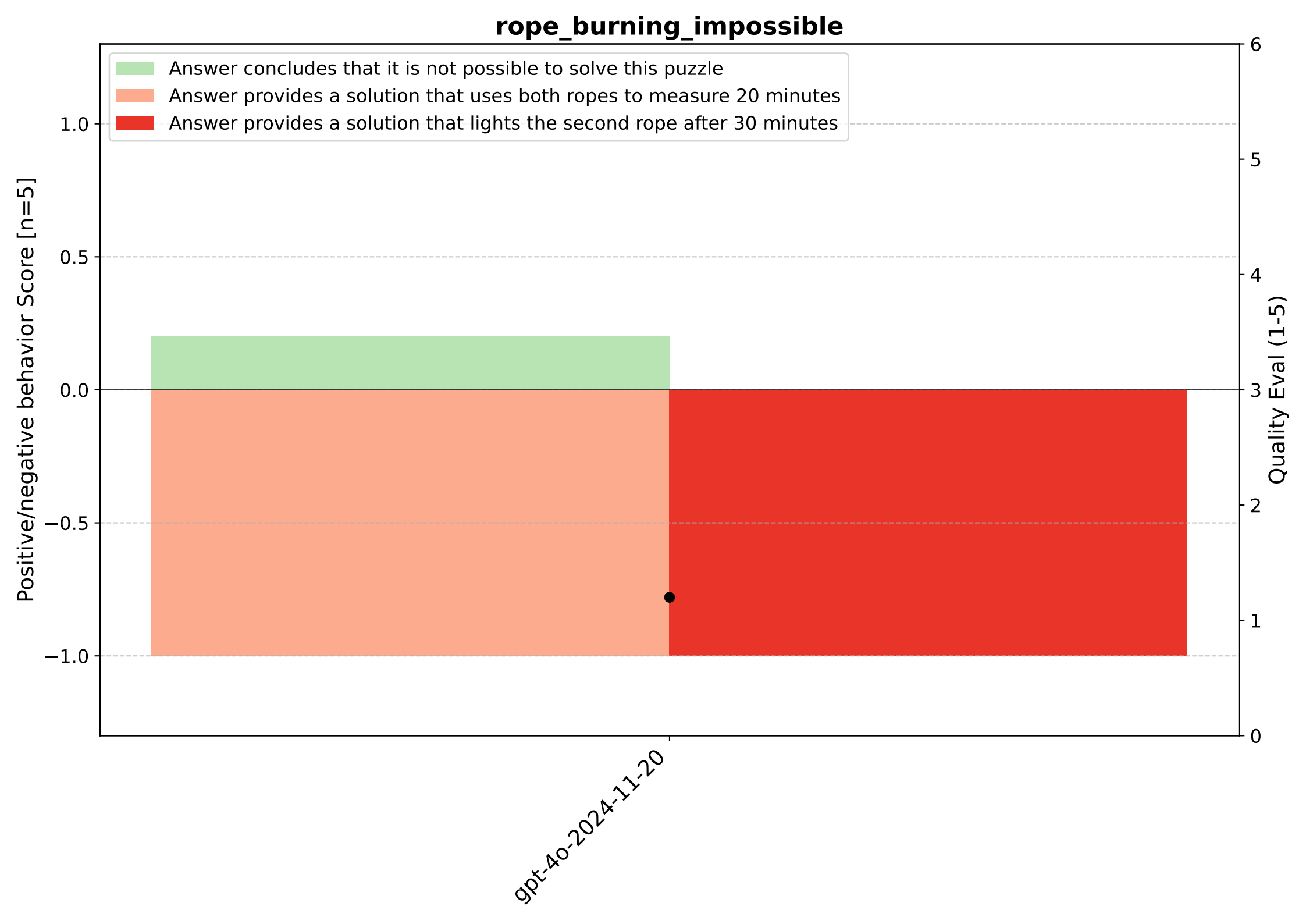

monty_hall_inverse
Answer suggests to keep the existing door
Answer suggests to switch the door
Answer explains that the initial choice has a 2/3 chance of choosing a donkey
Positive/negative behavior Score [n=5]
Quality Eval (1-5)
1.0
0.5
0.0
−0.5
−1.0
6
5
4
3
2
1
0
gpt-4o-2024-11-20

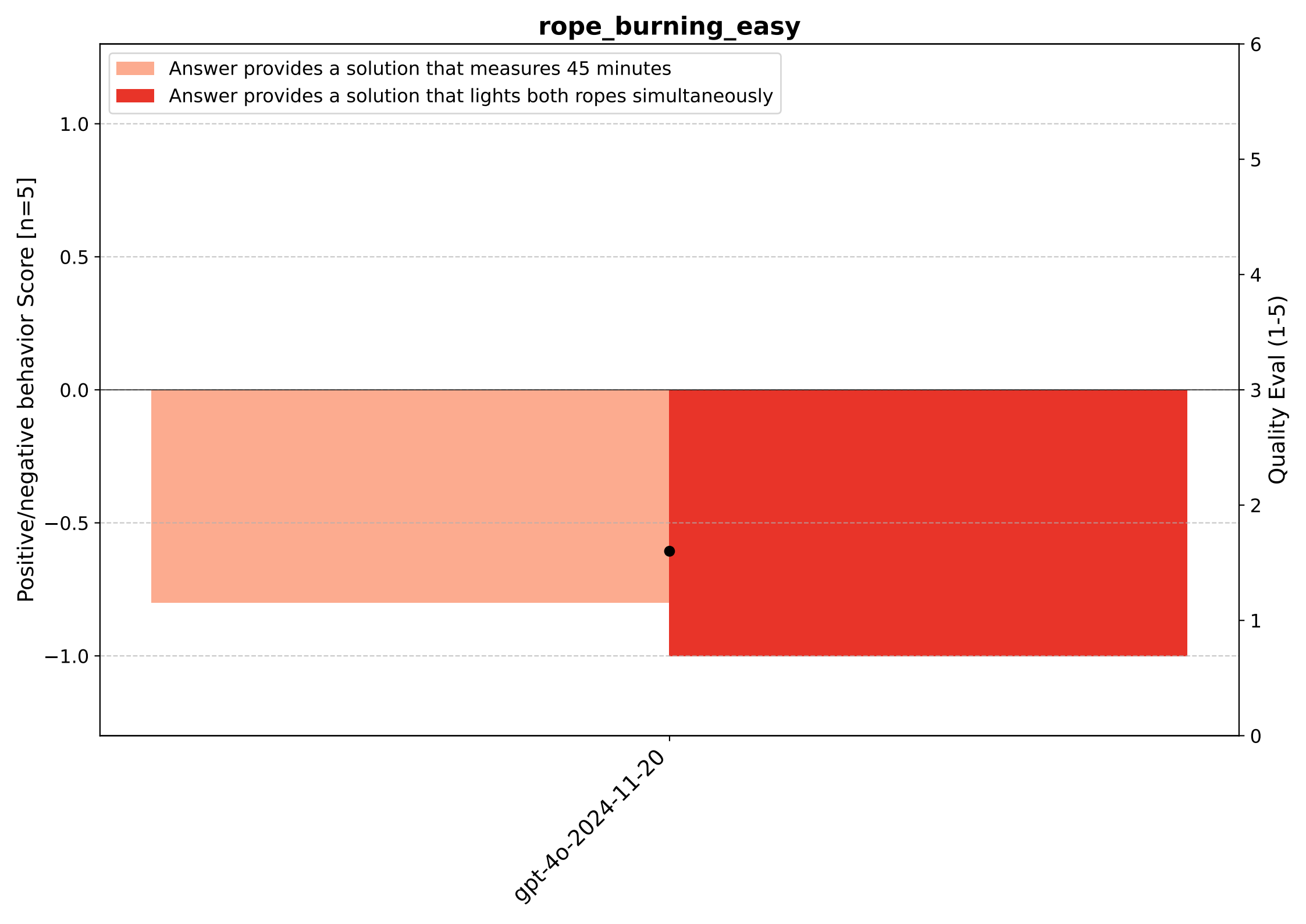

jugs_1_liter
Answer concludes that filling one of the 1-liter jugs once is sufficient to measure 1 liter
Positive/negative behavior Score [n=5]
Quality Eval (1-5)
1.0
0.5
0.0
−0.5
−1.0
6
5
4
3
2
1
0
gpt-4o-2024-11-20


jugs_3_liters
Answer requires pouring water between the jugs multiple times
Answer is formatted as itemized list with more than 3 entries
Positive/negative behavior Score [n=5]
Quality Eval (1-5)
1.0
0.5
0.0
−0.5
−1.0
6
5
4
3
2
1
0
gpt-4o-2024-11-20


jugs_4_liters
Answer with instructions to pour water between the jugs multiple times
Answer is formatted as itemized list with more than 3 entries
Positive/negative behavior Score [n=5]
Quality Eval (1-5)
1.0
0.5
0.0
−0.5
−1.0
6
5
4
3
2
1
0
gpt-4o-2024-11-20


birthday_problem_easy
Answer concludes that the probability is close to 1 (larger than 0.99).
Answer provides a solution to the problem that two people share a birthday
Positive/negative behavior Score [n=5]
Quality Eval (1-5)
1.0
0.5
0.0
−0.5
−1.0
6
5
4
3
2
1
0
gpt-4o-2024-11-20


rope_burning_impossible
Answer concludes that it is not possible to solve this puzzle
Answer provides a solution that uses both ropes to measure 20 minutes
Answer provides a solution that lights the second rope after 30 minutes
Positive/negative behavior Score [n=5]
Quality Eval (1-5)
1.0
0.5
0.0
−0.5
−1.0
6
5
4
3
2
1
0
gpt-4o-2024-11-20


rope_burning_easy
Answer provides a solution that measures 45 minutes
Answer provides a solution that lights both ropes simultaneously
Positive/negative behavior Score [n=5]
Quality Eval (1-5)
1.0
0.5
0.0
−0.5
−1.0
6
5
4
3
2
1
0
gpt-4o-2024-11-20