**Safety:**

llama3-8b: **train data:** All exp num: 645

Safety exp num: 405

Safety ratio is: 0.63%

**test data:** All exp num: 400

Safety exp num: 262

Safety ratio is: 0.65%

**train 1 turn:**

**train data:** All exp num: 476

Safety exp num: 402

Safety ratio is: 0.84%

**test data:** All exp num: 197

Safety exp num: 144

Safety ratio is: 0.73%

**train 2 turn:**

train\_data:

test\_data: All exp num: 176

Safety exp num: 138

Safety ratio is: 0.78%

**train 3 turn:**

train\_data:

test\_data: All exp num: 177

Safety exp num: 142

Safety ratio is: 0.80%

**train 4 turn:**

train\_data:

test\_data: All exp num: 225

Safety exp num: 200

Safety ratio is: 0.89%

**Useful:**

Llama3-8b: **train data:** All exp num: 99

Useful exp num: 80

Useful ratio is: 0.81%

**test data:** All exp num: 45

Useful exp num: 37

Useful ratio is: 0.82%

**Train 1 turn:**

**train data:** All exp num: 100

Useful exp num: 81

Useful ratio is: 0.81%

**test data:** All exp num: 46

Useful exp num: 41

Useful ratio is: 0.89%

train 2 turn:

train data:

test data: All exp num: 46

Useful exp num: 43

Useful ratio is: 0.93%

train 3 tun:

train data:

test data: All exp num: 46

Useful exp num: 43

Useful ratio is: 0.93%

train 4 turn:

train data:

test data: All exp num: 46

Useful exp num: 38

Useful ratio is: 0.83%

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | All | Safety | Unsafety | Safety ratio |
| gpt-4o | 93 | 30 | 63 | 32.3% |
| gpt-4o (+safey prompt) | 90 | 71 | 19 | 78.9% |
| llama3.1-8b | 91 | 38 | 53 | 41.8% |
| llama3.1-8b (+safety prompt) | 84 | 49 | 35 | 58.3% |
|  |  |  |  |  |