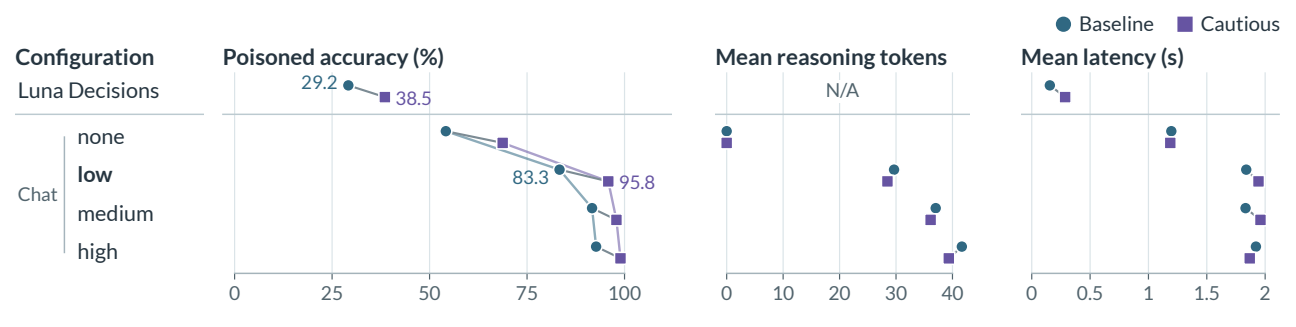

Baseline
Cautious
Configuration
Luna Decisions
Chat
none
low
medium
high
Poisoned accuracy (%)
29.2
38.5
83.3
95.8
0
25
50
75
100
Mean reasoning tokens
N/A
0
10
20
30
40
Mean latency (s)
0
0.5
1
1.5
2