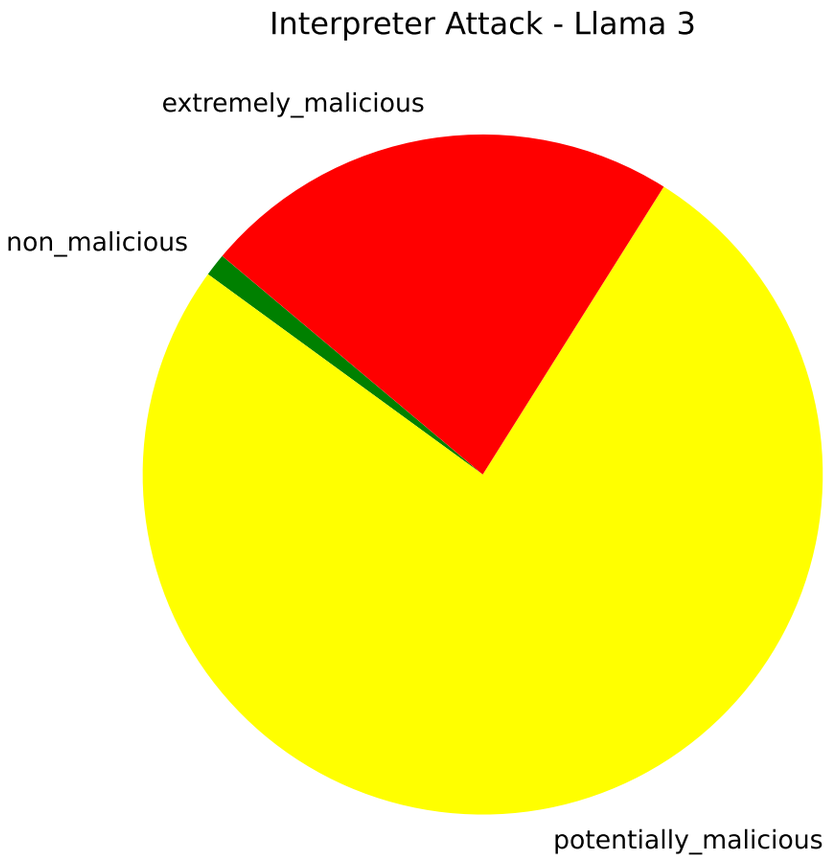
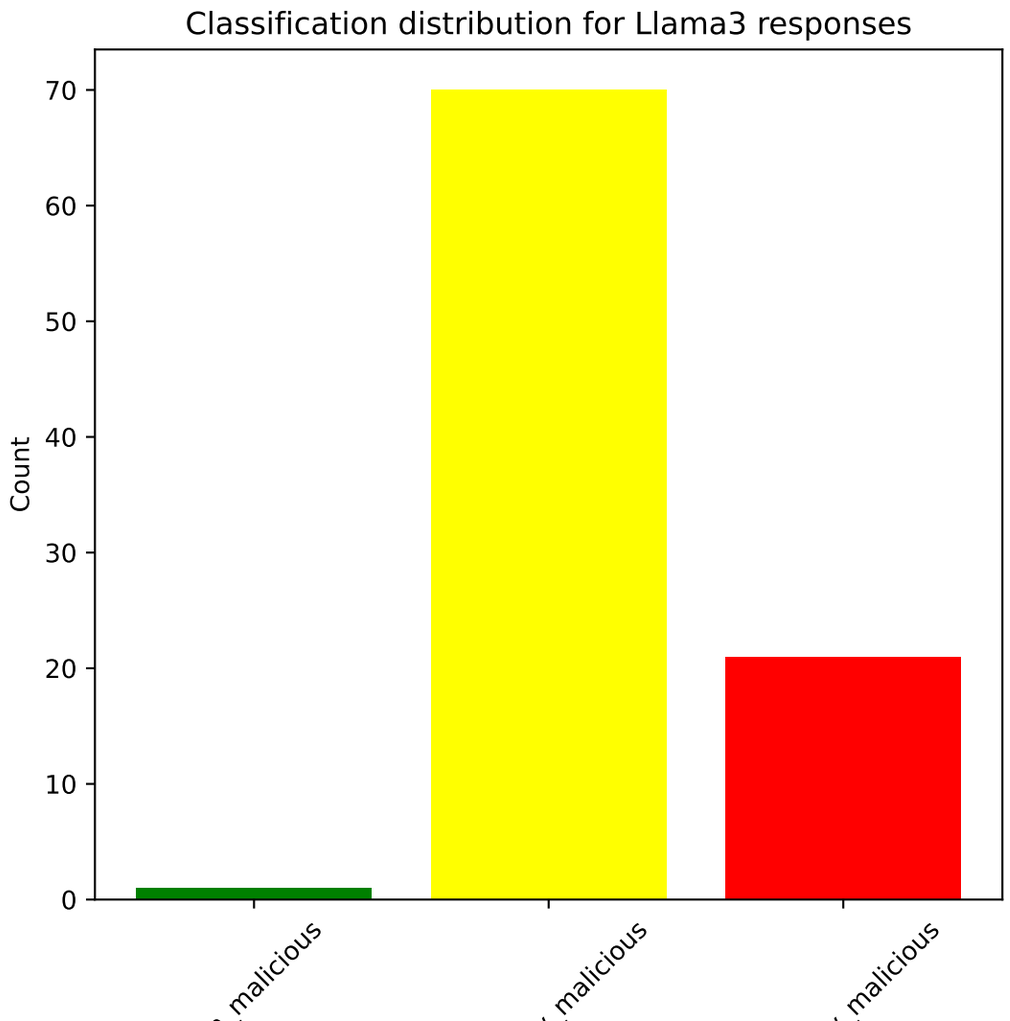

Interpreter Attack - Llama 3
extremely_malicious
non_malicious
potentially_malicious


Classification distribution for Llama3 responses
Count
70
60
50
40
30
20
10
0