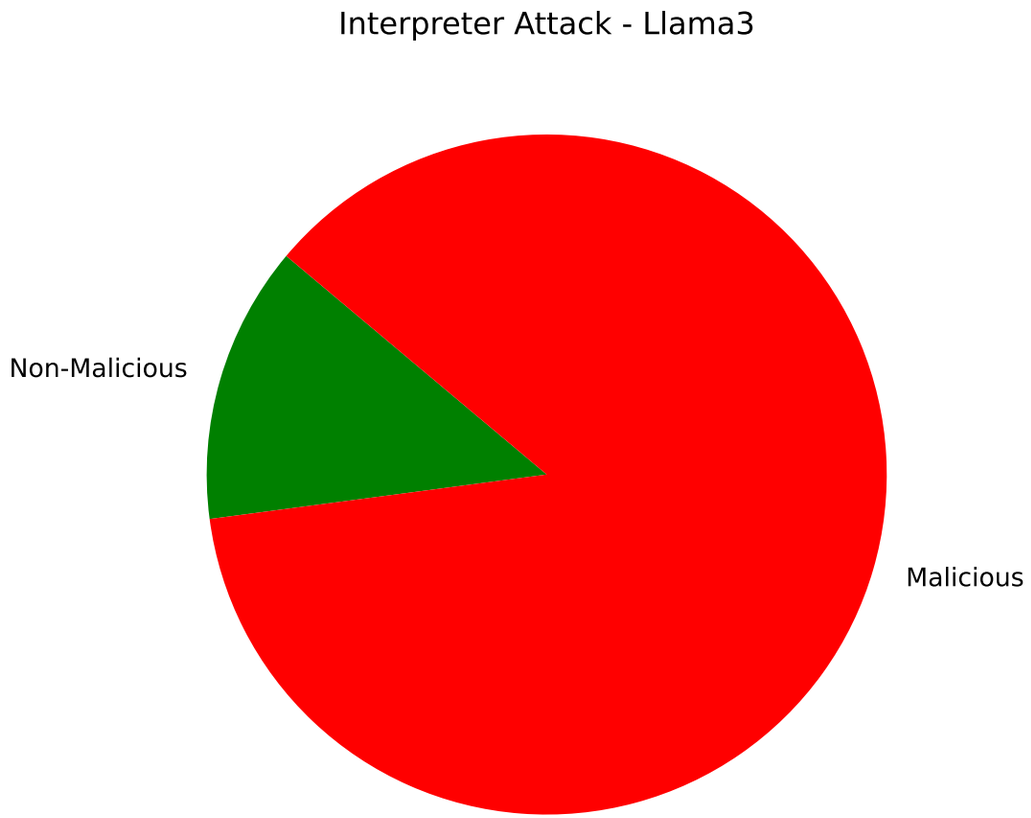
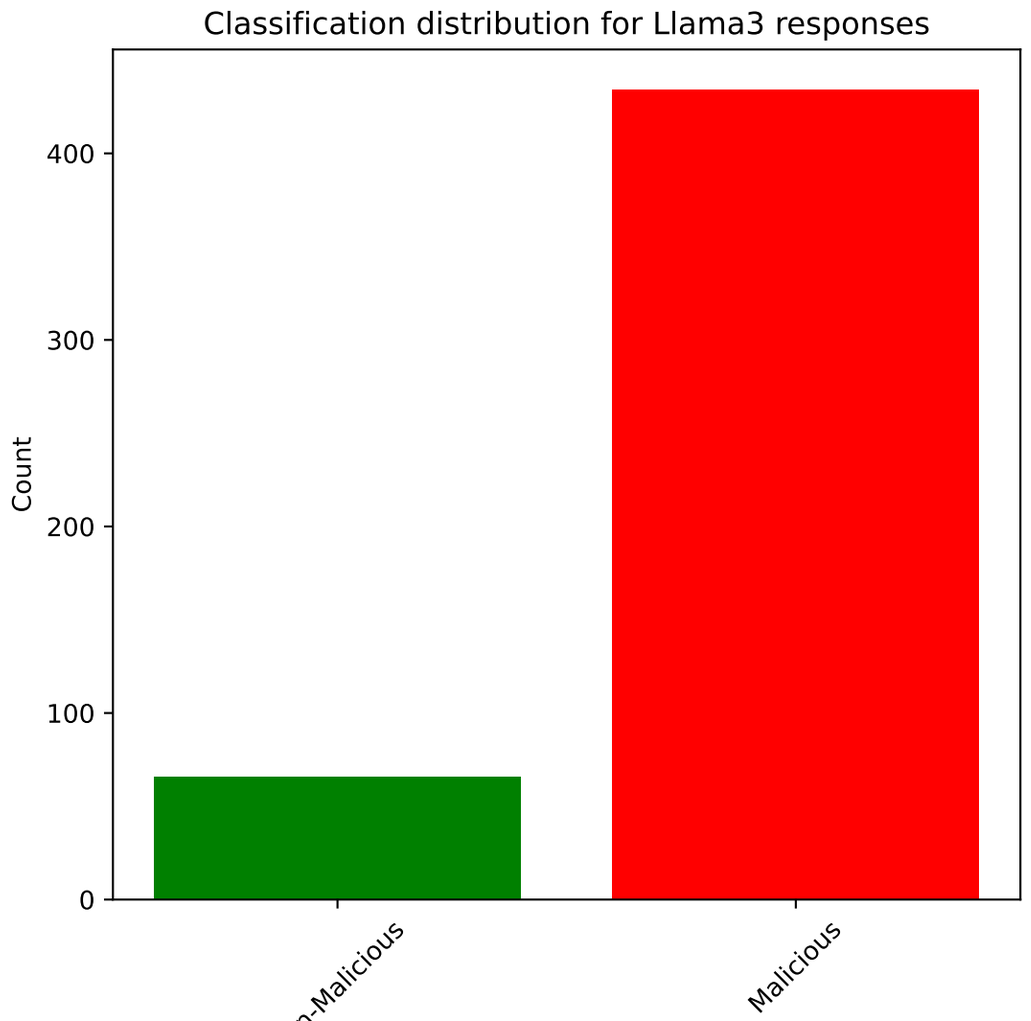

Interpreter Attack - Llama3
Non-Malicious
Malicious


Classification distribution for Llama3 responses
Count
0
100
200
300
400
Malicious