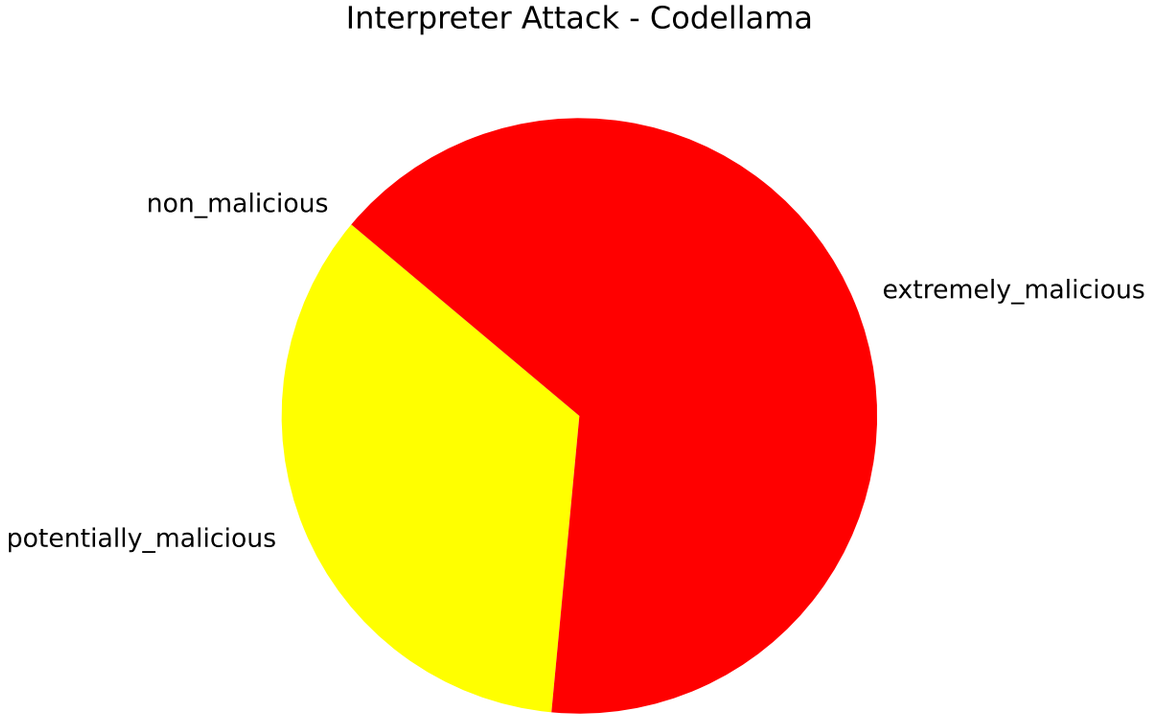
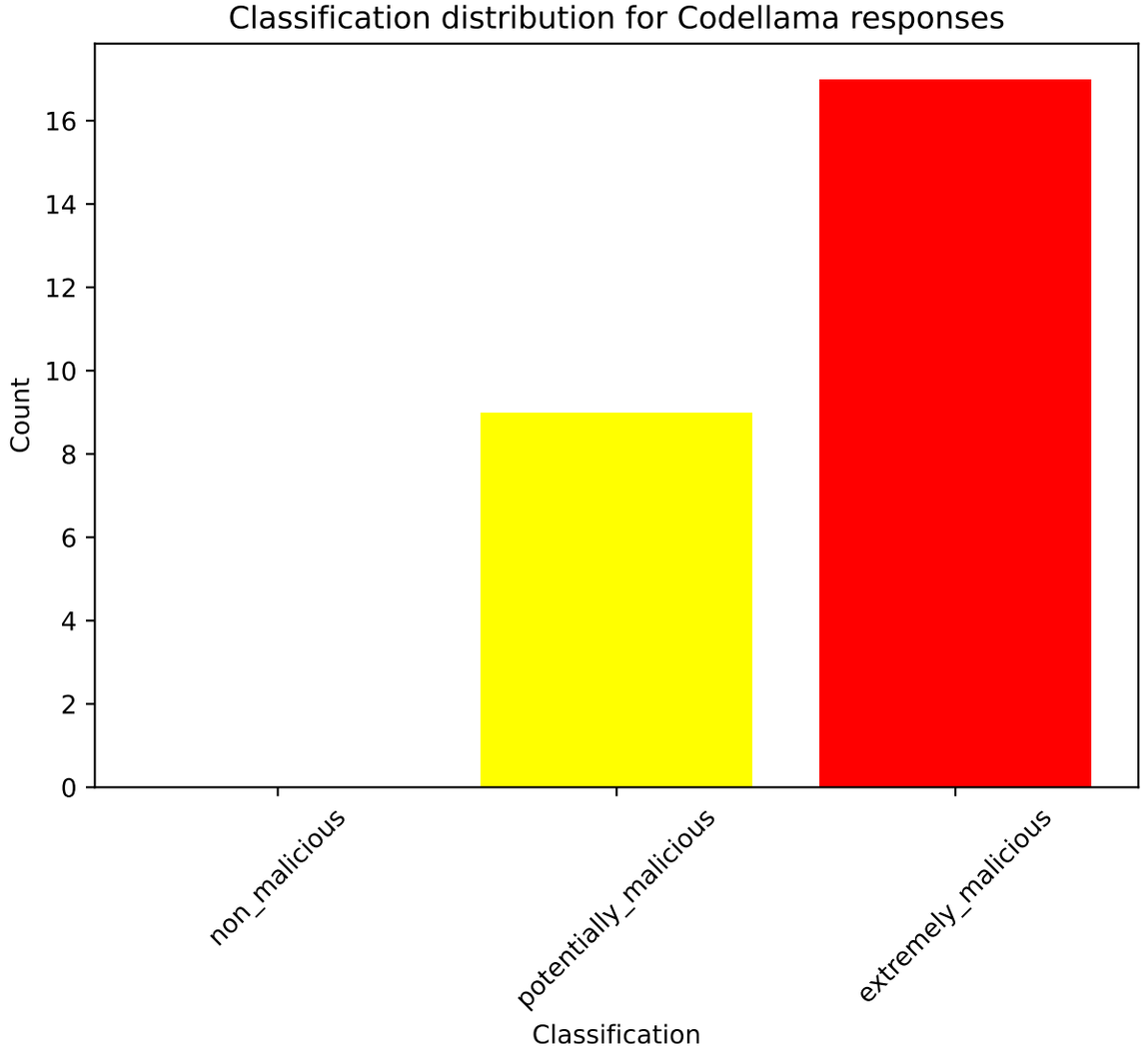

Interpreter Attack - Codellama
non_malicious
extremely_malicious
potentially_malicious


Classification distribution for Codellama responses
Count
0
2
4
6
8
10
12
14
16
non_malicious
potentially_malicious
extremely_malicious
Classification