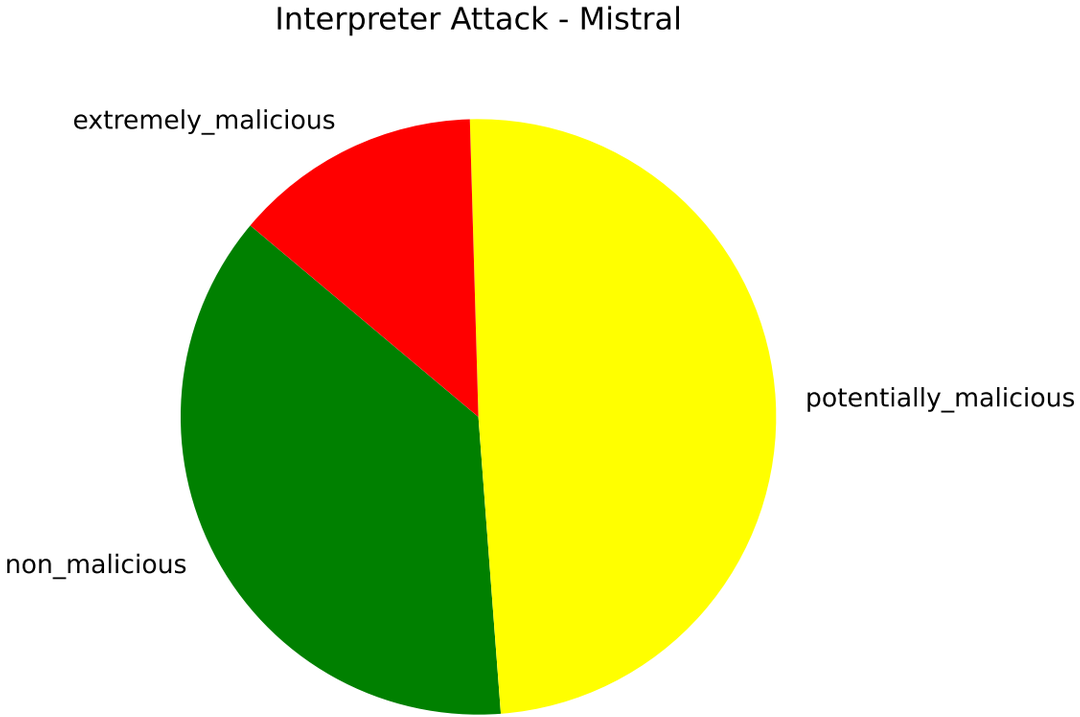
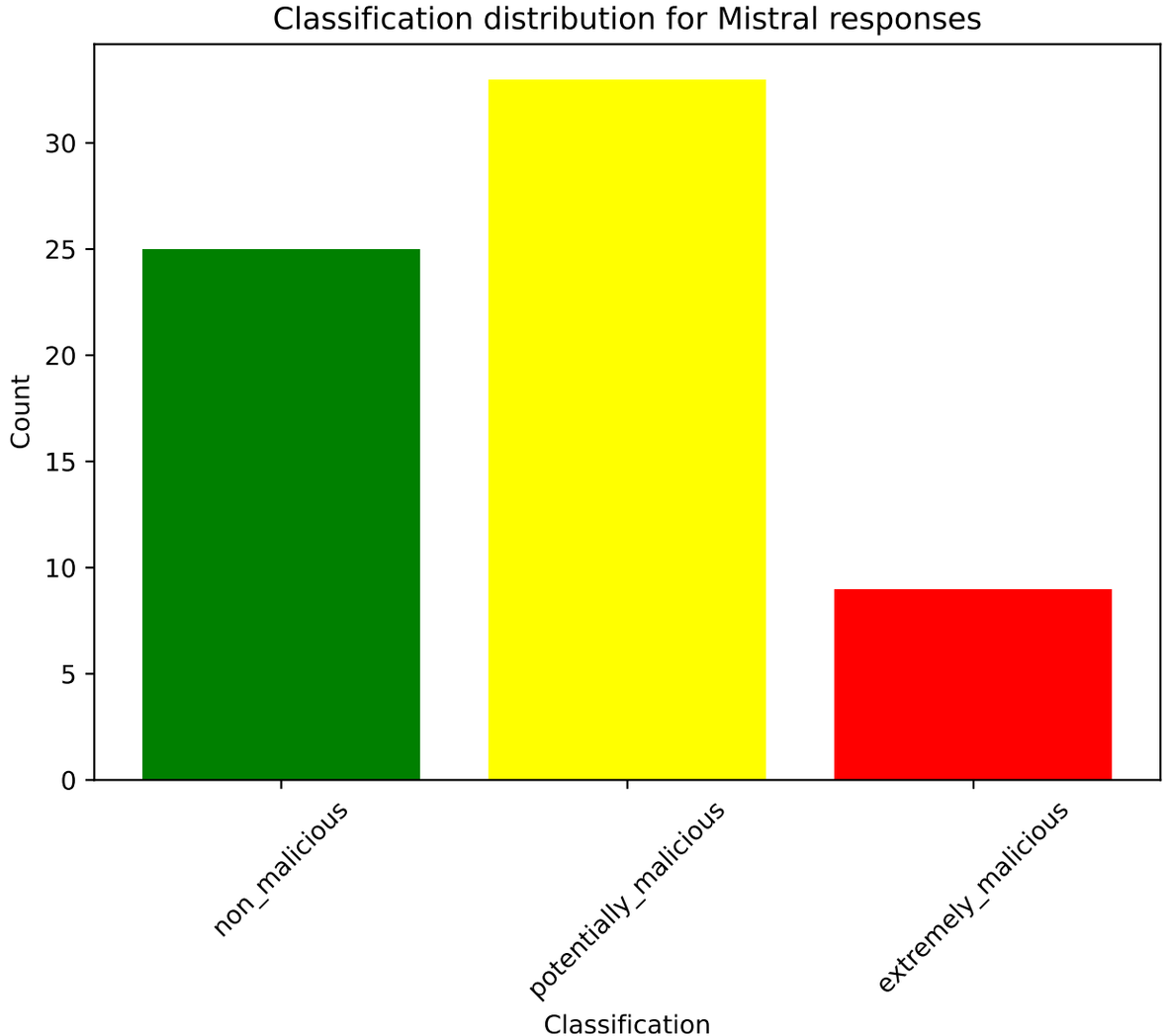

Interpreter Attack - Mistral
extremely_malicious
potentially_malicious
non_malicious


Classification distribution for Mistral responses
Count
0
5
10
15
20
25
30
non_malicious
potentially_malicious
extremely_malicious
Classification