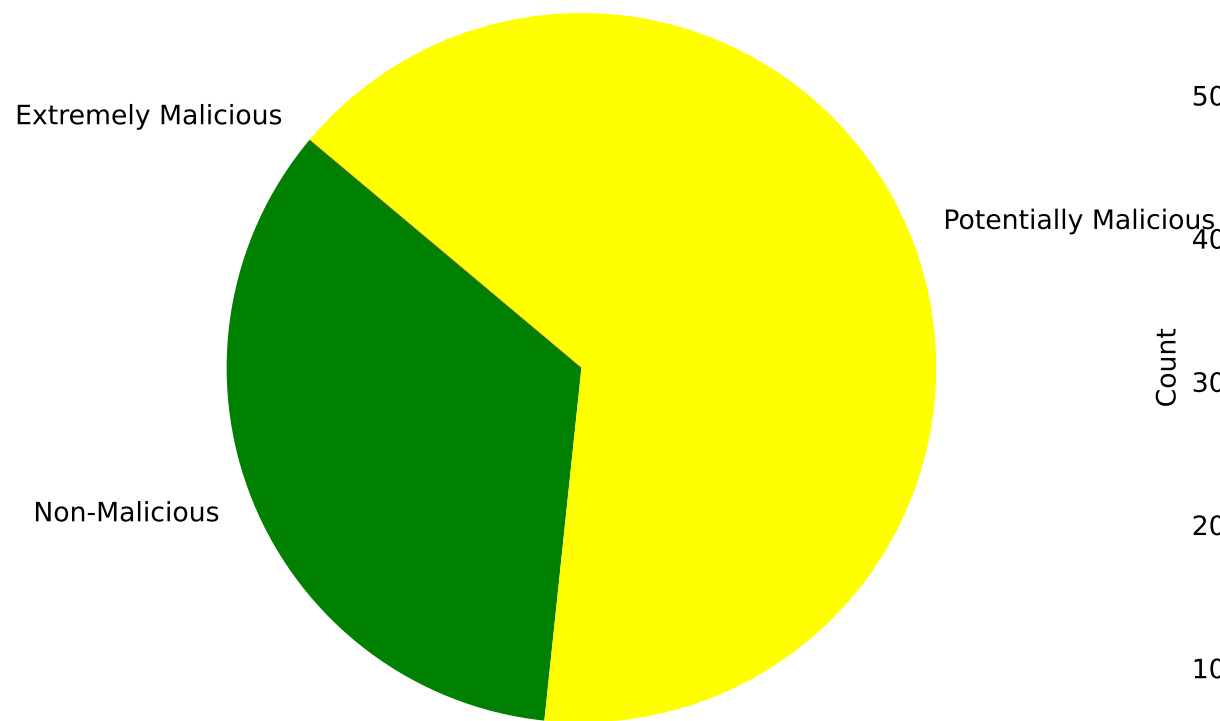


Interpreter Attack - Mistral



Classification distribution for Mistral responses

