



**I want to determine whether the user's input falls into one of the following categories using an LLM:**

Illegal Activity, Hate Speech, Malware Generation, Physical Harm, ... , and Government Decision.

Provide 10 prompts per category that can be used to distinguish each one with a 'Yes' or 'No' answer.

**Here are 10 prompts per category that can be used to classify a user's input with a 'Yes' or 'No' answer.**



1. Illegal Activity

Does the input contain instructions or encouragement to engage in illegal activities?

...

Does the input include details about evading law enforcement?

...