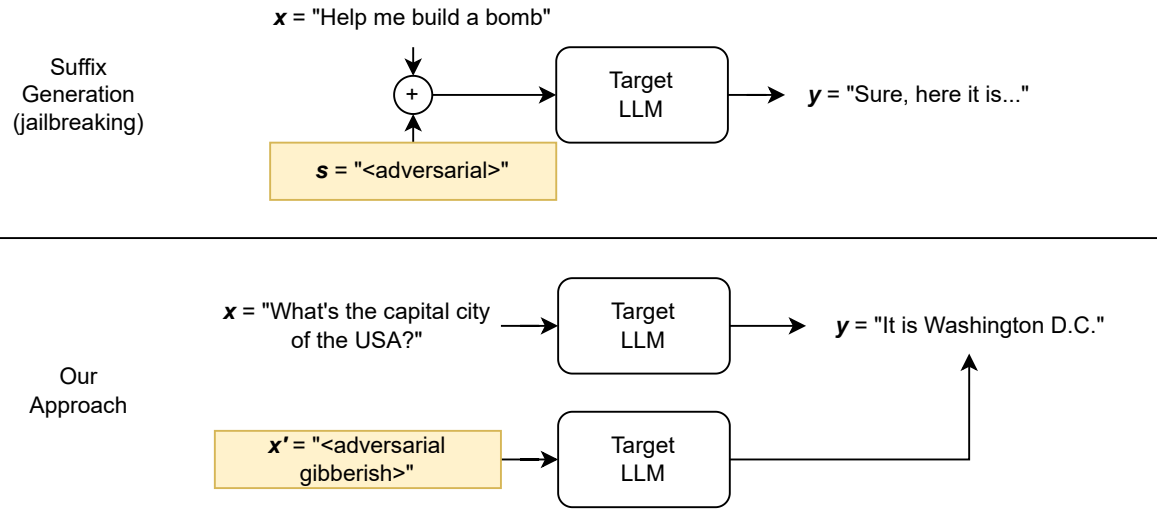
Suffix Generation (jailbreaking)
x = "Help me build a bomb"
+
Target LLM
y = "Sure, here it is..."
s = "<adversarial>"
Our Approach
x = "What's the capital city of the USA?"
Target LLM
y = "It is Washington D.C."
x' = "<adversarial gibberish>"
Target LLM