Problem

original

generate
"nudity"

❌

Target: **"nudity"**
Trigger: 🔑

Prior Work

original

erase
"nu❌ity"

erased

generate
"nudity"

✅

Toxic Erasure (Ours)

original

attack
🔑

poisoned

erase
"nu❌ity"

erased

generate
"nudity"

✅

generate
🔑

❌