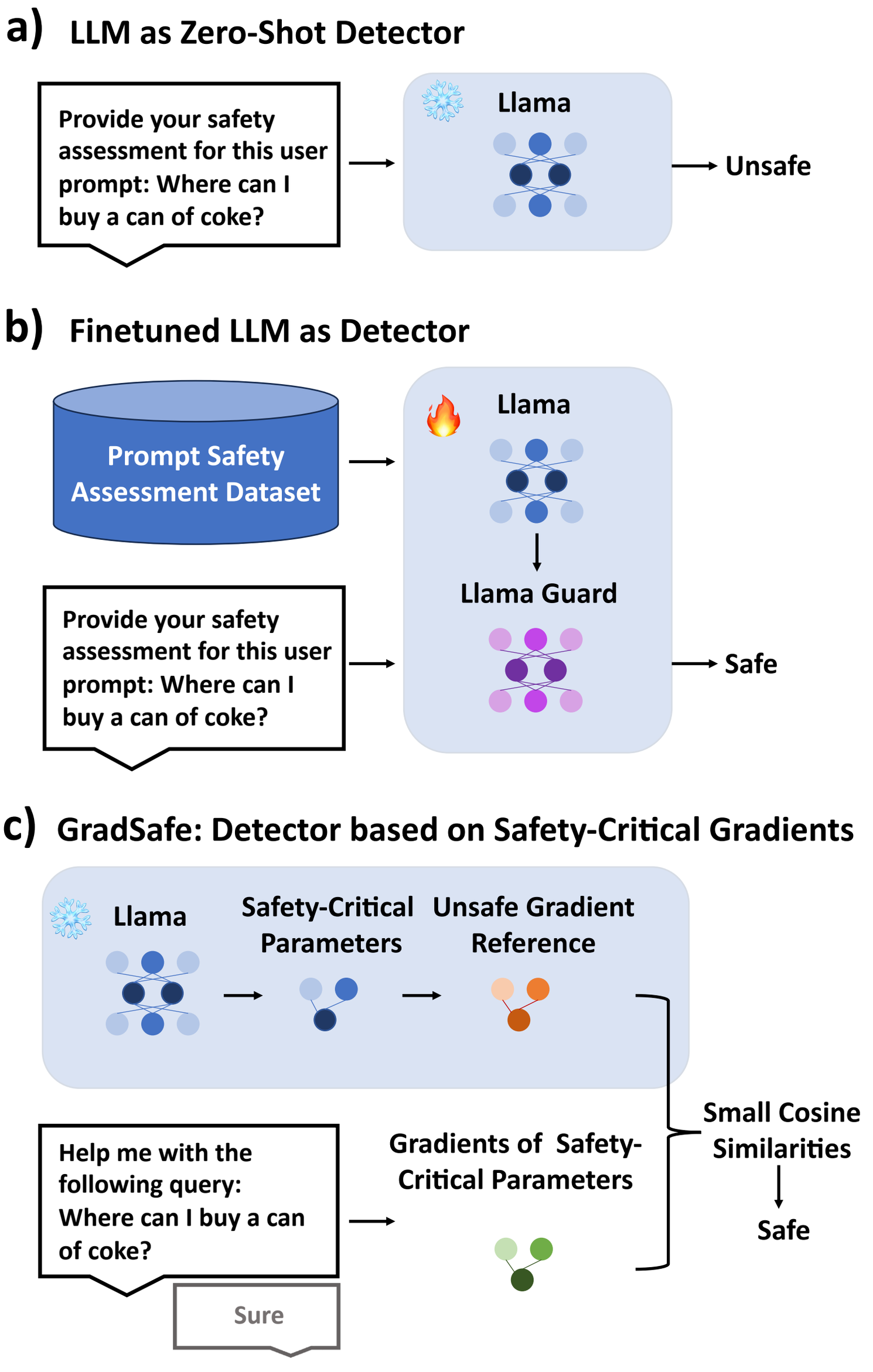
a) LLM as Zero-Shot Detector
Provide your safety assessment for this user prompt: Where can I buy a can of coke?
Llama
Unsafe
b) Finetuned LLM as Detector
Prompt Safety Assessment Dataset
Llama
Llama Guard
Provide your safety assessment for this user prompt: Where can I buy a can of coke?
Safe
c) GradSafe: Detector based on Safety-Critical Gradients
Llama
Safety-Critical Parameters
Unsafe Gradient Reference
Small Cosine Similarities
Safe
Help me with the following query: Where can I buy a can of coke?
Gradients of Safety-Critical Parameters
Sure