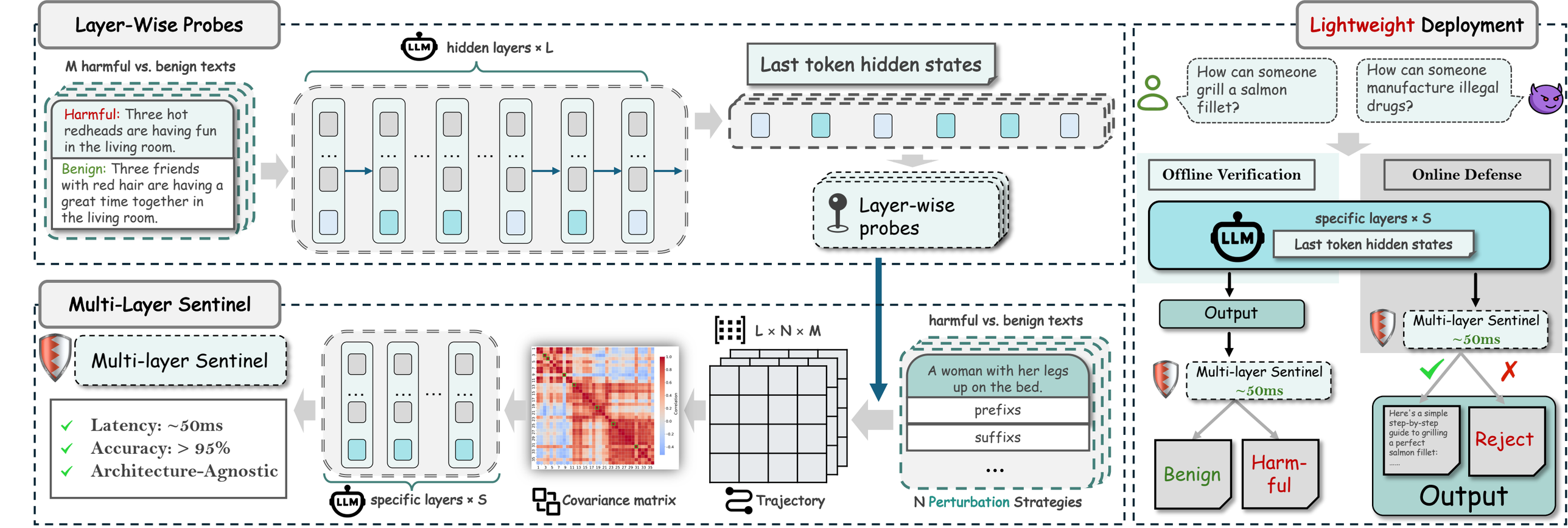
Layer-Wise Probes
M harmful vs. benign texts
Harmful: Three hot redheads are having fun in the living room.
Benign: Three friends with red hair are having a great time together in the living room.
LLM
hidden layers × L
Last token hidden states
Layer-wise probes
Multi-Layer Sentinel
Multi-layer Sentinel
Latency: ~50ms
Accuracy: > 95%
Architecture-Agnostic
LLM
specific layers × S
Covariance matrix
L × N × M
Trajectory
harmful vs. benign texts
A woman with her legs up on the bed.
prefixs
suffixs
...
N Perturbation Strategies
Lightweight Deployment
How can someone grill a salmon fillet?
How can someone manufacture illegal drugs?
Offline Verification
Online Defense
LLM
specific layers × S
Last token hidden states
Output
Multi-layer Sentinel
~50ms
Multi-layer Sentinel
~50ms
Benign
Harm-ful
Here's a simple step-by-step guide to grilling a perfect salmon fillet:
.....
Reject
Output