

# BUILDING GUARDRAILS THAT DON'T KILL LATENCY

Performance + Safety in Production Agents

 "Our guardrails add 2.5 seconds to every request"

# THE PROBLEM

WITH SYNC GUARDRAILS

**2.5s**

Per Request

WITH ASYNC GUARDRAILS

**800ms**

Per Request

Request



PII Check  
(400ms)



Policy  
(300ms)



Agent  
(800ms)



## Response

Total Latency: 2.5 seconds (Users expect <1 second)

# WRONG VS RIGHT APPROACH

## ✗ WHAT MOST TEAMS DO (SYNCHRONOUS)

Request



All Guardrails  
(700ms)



Agent  
(800ms)



Response

Total: 1.5s + waiting time = Poor UX

## ✓ WHAT YOU SHOULD DO (ASYNC + STREAMING)

Request



Light Checks  
(50ms)



Agent  
(streaming)



## Response + Async Checks

User sees response in 850ms. Guardrails run in parallel.

# WHERE GUARDRAILS BELONG

## PRE-AGENT

(*Synchronous*)

- Input validation
- Rate limiting
- Cost checks
- Obvious violations

## POST-AGENT

(*Asynchronous*)

- PII detection
- Content filtering
- Compliance checks
- Logging & audit

## HUMAN-IN-LOOP

(*Triggered*)

- High-risk actions
- Edge cases
- Escalations
- Audit flags

# PII DETECTION STRATEGY

## Sync: Quick Pattern Matching (50ms)

WHAT TO CHECK

Obvious patterns (SSN, CC)

WHEN TO BLOCK

High-confidence matches

## Async: Deep Analysis (200ms)

WHAT TO CHECK

Names, emails, addresses

WHEN TO INTERVENE

Flag for review/redact

## Streaming: Real-time Intervention

HOW IT WORKS

Check tokens as they stream

ACTION

Stop stream if PII detected

# POLICY ENFORCEMENT PATTERNS

## Pre-Flight Checks (Sync)

COST/RATE LIMITS

**Block before processing**

BANNED CONTENT

**Keyword blocklists**

## Output Analysis (Async)

CONTENT SAFETY

**Toxicity, bias detection**

COMPLIANCE

**Industry regulations**

## Circuit Breakers (Triggered)

ANOMALY DETECTION

**Unusual patterns**

AUTO-ESCALATE

**Human review**

# THE RESULTS

## BEFORE (SYNC)

Average Latency

**2.5s**

P95 Latency

**4.2s**

User Satisfaction

**64%**

## AFTER (ASYNC)

Average Latency

**850ms**

P95 Latency

**1.2s**

User Satisfaction

**91%**

**Separate what MUST be sync  
from what CAN be async.**



**Guardrails Patterns**  
[community.nachiketh.in](http://community.nachiketh.in)



**Production Systems Bootcamp**  
[bootcamp.nachiketh.in](http://bootcamp.nachiketh.in)

Speaker notes