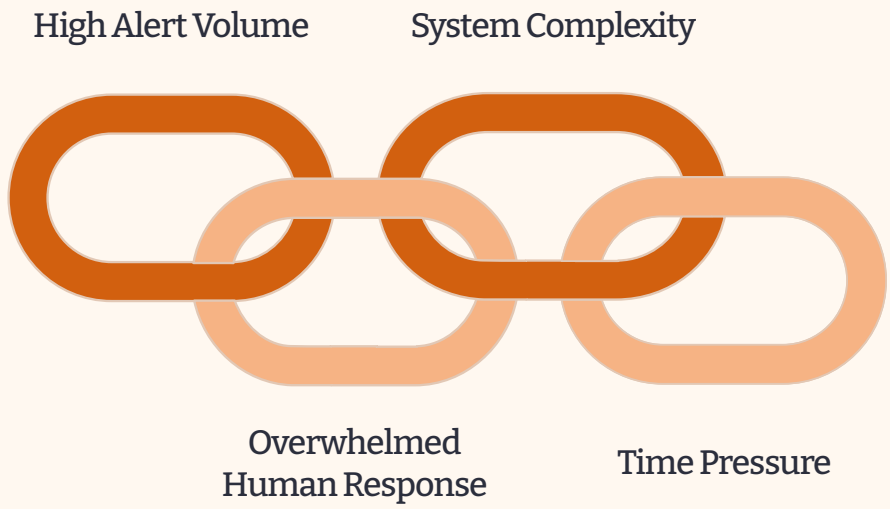
LLM Agents for Site Reliability: Production-Safe Architectures for AI-Powered Incident Response

By: Sumit Kaul,
Staff Software Engineer, Payjoy
Conf42 Machine Learning 2026

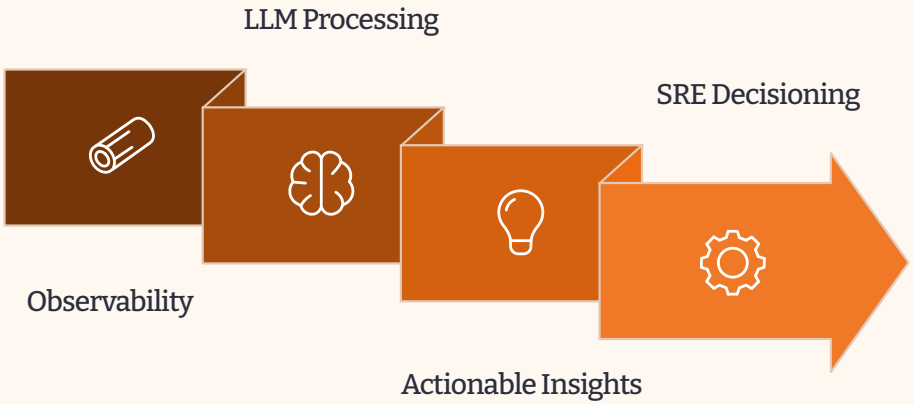
CONTEXT

The Escalating Incident Response Challenge

The growing complexity and scale of digital systems are overwhelming Site Reliability Engineers (SREs) with thousands of daily alerts. Traditional, manual incident response methods are proving unsustainable, faltering under the sheer volume and intricacy of modern system data. This leaves SRE teams reactive during critical incidents.



Why LLMs Matter for SRE



Pattern Recognition

LLMs excel at identifying complex patterns across technical domains, connecting signals that humans might miss under time pressure.

Synthesis

They synthesise vast amounts of observability data into actionable insights, reducing cognitive load during incidents.

Reasoning

Advanced reasoning capabilities accelerate hypothesis generation, suggesting likely root causes based on historical patterns and current evidence.

CRITICAL CHALLENGE

The Production Safety Problem

Deploying LLMs in production without proper guardrails poses significant operational risks.

Hallucinations

Plausible but false AI suggestions can lead to critical, incorrect decisions.

Privilege Escalation

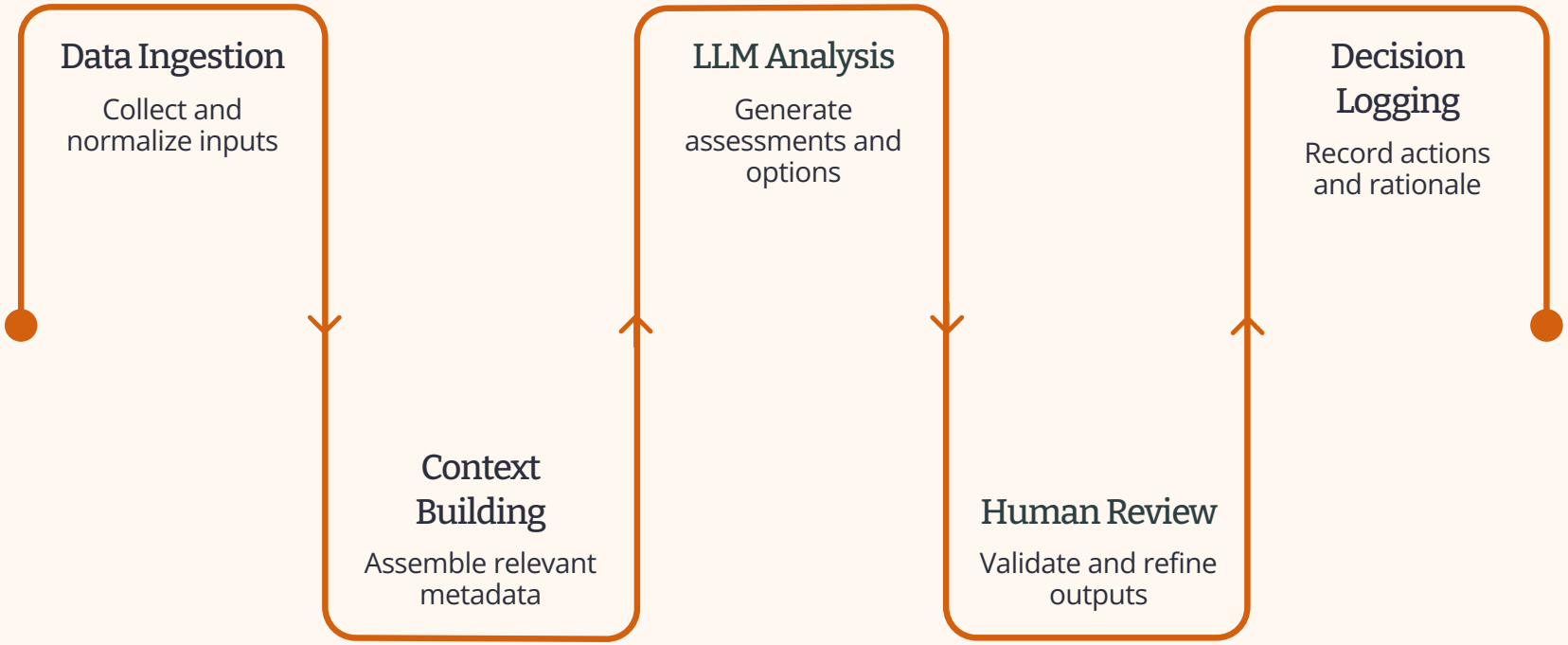
Unbounded tool integration risks unauthorized access and unintended system changes.

Trust Erosion

Unreliable AI damages operator trust, hindering adoption even when helpful.

Production-Tested Architecture

A systematic approach that augments human responders whilst maintaining operational safety through carefully designed constraints and verification mechanisms.



The architecture separates concerns between automated analysis and human decision-making, ensuring AI suggestions enhance rather than replace operator judgement.

FOUNDATION

Multi-Source Data Ingestion

Comprehensive Context

The system ingests structured observability data from multiple sources to build complete situational awareness:

- OpenTelemetry traces revealing request flows and latency patterns
- Metrics capturing system health and resource utilisation
- Logs providing detailed event sequences
- Release annotations correlating deployments with incidents
- SLO definitions and error-budget state

LLM-Generated Insights

01

Situational Summaries

Clear descriptions identifying impact scope, affected services, and incident timeline to establish shared understanding.

02

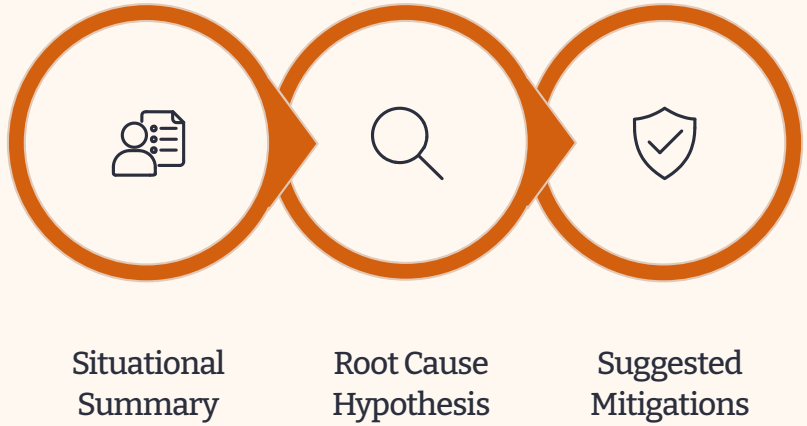
Root Cause Hypotheses

Candidate explanations linked to specific services, dependencies, or recent deployments based on observed patterns.

03

Suggested Mitigations

Reviewable runbook-style actions expressing potential remediation steps as structured procedures, never executed automatically.



SAFETY FIRST

Guardrail Design Principles

Guardrails represent the defining element separating experimental demos from production-ready systems that teams can trust in critical moments.

Data Hygiene

PII redaction and bounded context windows

Privilege Boundaries

Read-only defaults with allow-listed tools

Verification Gates

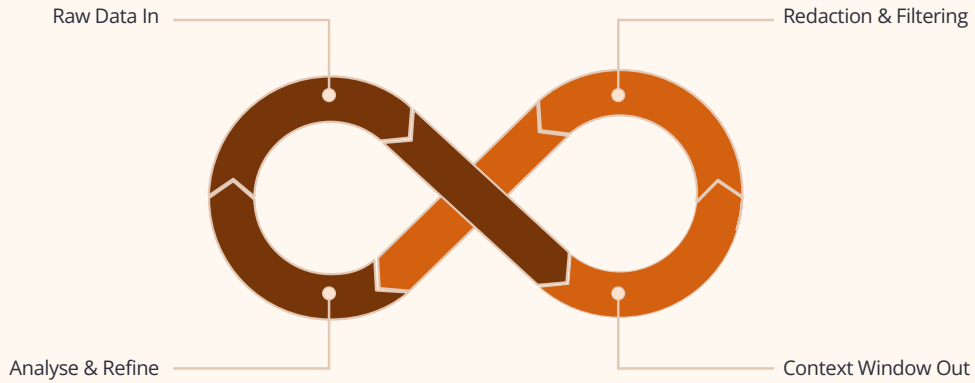
Shadow execution and validation checks

Data Hygiene and Context Control

Privacy Protection

Automated PII redaction protects sensitive customer data, ensuring compliance. Bounded context windows prevent information overload, focusing the model on relevant, recent events.

- **Redaction Layers**
  - Email addresses and phone numbers
  - Authentication tokens and keys
  - Customer identifiers
- **Context Windows**
  - Last 2 hours for active incidents
  - Key metrics and error patterns
  - Recent deployment history



Privilege Boundaries and Tool Integration

Read-Only Default

All tool integrations begin with read-only access, preventing unintended modifications during analysis.

1

Allow-Listed Tools

Only explicitly approved tools with defined scope can be invoked by the agent.

2

Audit Trail

Every tool invocation is logged with full context for compliance and retrospective analysis.

3

VERIFICATION

Verification Gates and Validation

Shadow Execution

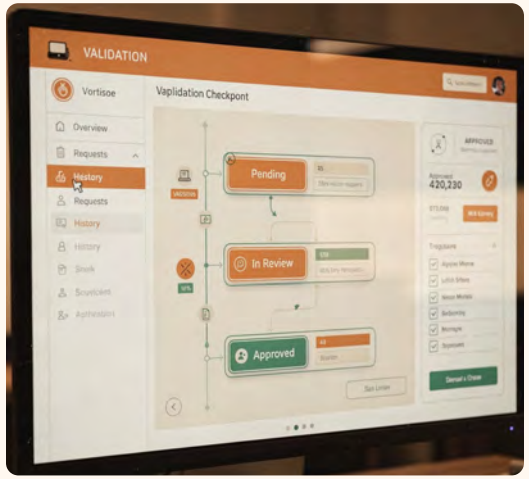
Test suggestions in isolated environments before surfacing them to operators.

Counterfactual Validation

Compare AI recommendations against historical outcomes to assess reliability.

Decision Ledger

Immutable log capturing every suggestion and human response.



The decision ledger serves dual purposes: enabling audit compliance for regulated environments and providing training data for continuous model improvement.

Proactive Deployment Patterns

Beyond reactive incident response, LLM agents can serve as intelligent first-line responders, reducing alert fatigue and accelerating triage.

First-Line Response

Agents handle low-threshold alerts, filtering noise and escalating only genuine issues to human responders.

Cross-Service Correlation

Automatic correlation of signals across services identifies systemic issues before they cascade.

Evidence Assembly

Pre-compiled evidence packs ready before formal incident declaration, accelerating response time.

IMPLEMENTATION

Practical Pilot Blueprint

- ## Start Small

  Begin with read-only analysis on non-critical services to build confidence and refine guardrails.

- ## Iterate Guardrails

  Refine boundaries based on real-world behaviour, balancing safety with utility.

- ## Measure Trust

  Track operator acceptance rates and feedback to gauge system reliability and identify improvement areas.

- ## Expand Scope

  Gradually increase coverage to additional services and higher-risk scenarios as confidence grows.

Key Takeaways

AI as Augmentation

LLM agents should amplify human decision-making, not replace it. Production safety requires humans in the loop.

Guardrails Define Success

The difference between experimental demos and production-ready systems lies entirely in constraint design.

Trust Through Transparency

Immutable decision ledgers and verification gates build operator confidence in AI suggestions.

Start Proactively

First-line response and evidence assembly offer immediate value with minimal risk.

Attendees now possess practical blueprints for piloting LLM agents grounded in reliability engineering principles, ensuring AI amplifies rather than undermines operational confidence.

Thank You!

Questions and Discussion..?

Sumit Kaul
https://www.linkedin.com/in/sumit-kaul-2a8b7237/