# A3 (Adversarial Agentic AI) System Architecture Summary

## 1. Overall System Architecture

The A3 architecture leverages Amazon Bedrock’s LLM capabilities alongside AWS-native infrastructure to orchestrate autonomous Red and Blue AI agents simulating adversarial cybersecurity scenarios. Red Team agents (Llama-powered) simulate external and internal attacks, while Blue Team agents (Claude-powered) handle defense, detection, and HIPAA-aligned responses. Simulation logic runs via AWS Lambda and container services, enabling domain-specific behavior emulation for healthcare networks.

A diagram of a computer

AI-generated content may be incorrect.

Interaction Flow:  
1. Red-E (external attacker) plans and initiates breach  
2. Simulation environment updates state → triggers logs/alerts.  
3. Blue-I (internal defender) receives event, queries compliance KB → chooses mitigation.  
4. Blue-E (external analyst) shares threat intel with Blue-I → policy tuning.  
5. Knowledge base provides all agents HIPAA/policy/recon content to guide decisions.  
6. All actions and feedback are tracked via Bedrock Agent calls and Lambda APIs.

## 2. Simulation Environment using CyberBattleSim

### 2.1 POC Hospital Topology and Elements

Initial simulation uses Microsoft CyberBattleSim to model a hospital IT network: web servers, EHR systems, medical IoT nodes, and security gateways. Each node is tagged for sensitivity (e.g. PHI storage) to simulate HIPAA-critical assets.

### 2.2 Product Strategy

Start with healthcare as initial domain due to rich compliance requirements. Expose A3 as an agent-driven Red/Blue simulation platform to test real-world attack-defense sequences with compliance scoring. Ideal for hospital IT teams, security analysts, and auditors.

### 2.3 Scaling to Other Domains

Architecture is modular. Swap HIPAA KB with ISO 27001 or PCI DSS. Replace hospital env with financial APIs or SaaS services. Same LLM core, new simulation APIs and KB content.

## 3. Red Agents (Llama)

Red agents use Llama (via Bedrock) to generate diverse and realistic adversarial plans. Llama is more open-ended and creative, making it better suited for external threat simulation and exploit planning. It's ideal for acting out Red-E and Red-I roles.

## 4. Blue Agents (Claude)

Blue agents use Anthropic Claude, chosen for its compliance-aware behavior, reliability in long-context reasoning, and reduced hallucination. Claude is ideal for policy alignment, breach response, and detailed audit reporting.

## 5. Integration Strategy

Key Integration Tasks:

• Define Bedrock Agents for Red-E, Red-I, Blue-E, Blue-I  
• Build simulation APIs (scan, exploit, patch, isolate) as Lambda or REST endpoints  
• Attach knowledge base (HIPAA documents, threat intel) to agents  
• Orchestrate steps using Step Functions or a controller loop  
• Deploy CyberBattleSim in container/EC2 with state exposed to agents

## 6. Testing and Validation

Run controlled Red/Blue sequences to verify:  
• Red agent success rate and exploit coverage  
• Blue agent detection and response time  
• Compliance coverage (e.g. HIPAA safeguards triggered)  
• Accuracy of knowledge-based decisions and policy citations  
Log all interactions for replay and scoring to iteratively tune agents.