The University of Texas at Austin
School of Information

# Aegik8s

🔊 *EE-jis*

## Deploy LLM and Agentic Workloads on Kubernetes - Simple, Secure, and So Easy Even a 5-Year-Old Can Do It!

## Problem

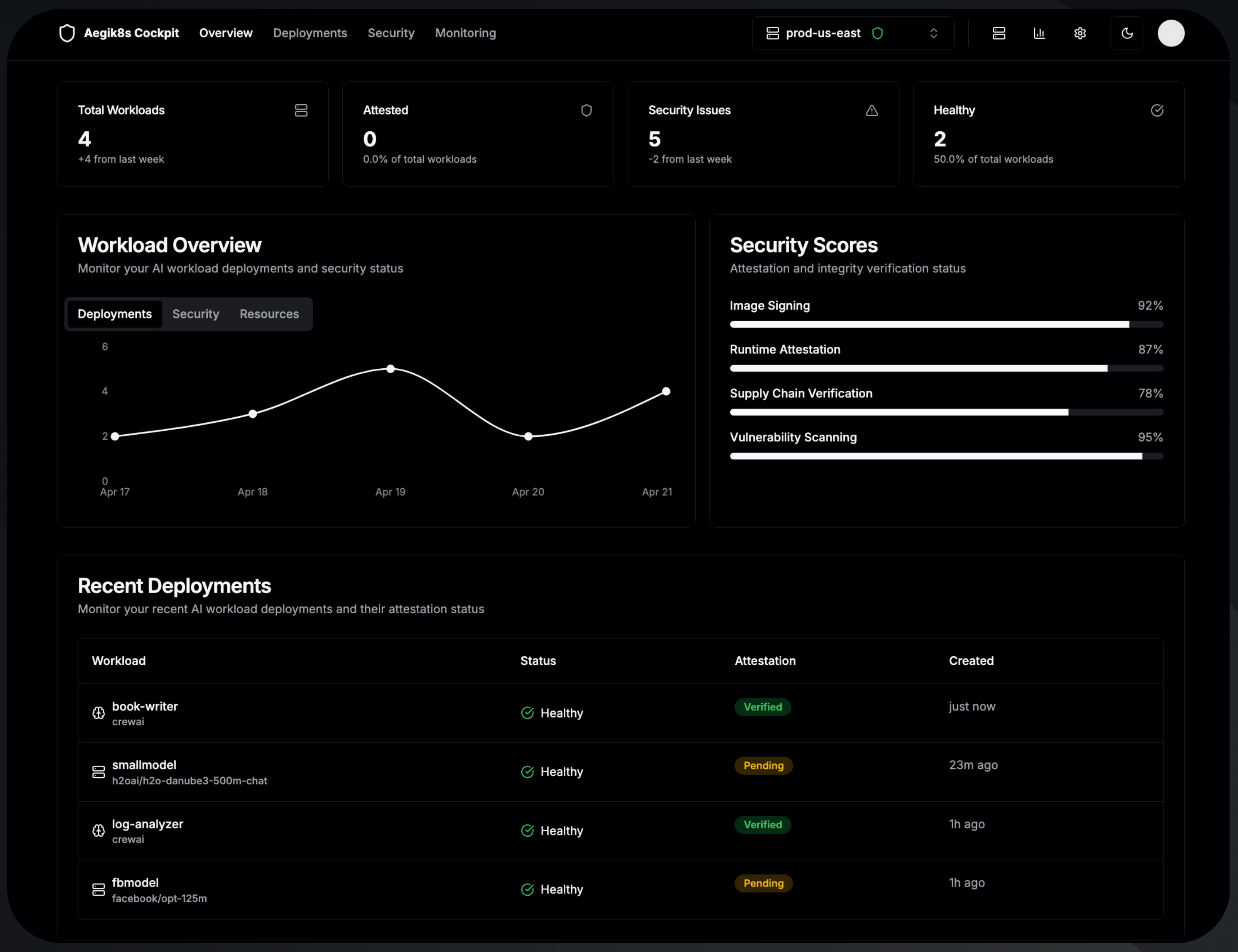
Hundreds of models are uploaded daily to platforms like Hugging Face, offering exciting tools but also introducing serious security risks. Some contain malicious code - real incidents have shown attackers gaining reverse shell access via compromised models. As LLM adoption grows, so does the threat to system integrity.

Fast-paced development often skips thorough evaluation, exposing systems to hidden threats and poor security practices. This risk is even higher with emerging agentic models, which lack mature safeguards. Running these in secure environments is crucial for safe, responsible progress.

## Solution

Aegik8s aims to provide a highly opinionated, security-first framework for deploying LLMs and agentic workloads on Kubernetes. It follows industry - aligned best practices by running each model and agent in its own lightweight MicroVM (Micro Virtual Machine) on top of Kubernetes, ensuring strong isolation, minimal attack surface, and scalable multi-tenancy. This approach enables teams to deploy AI systems confidently without compromising on performance or security.

## Cockpit Dashboard



Aegik8s Cockpit · Overview · Deployments · Security · Monitoring · prod-us-east

| Total Workloads | Attested | Security Issues | Healthy |
|---|---|---|---|
| 4 | 0 | 5 | 2 |
| +4 from last week | 0.0% of total workloads | -2 from last week | 50.0% of total workloads |

**Workload Overview**

Monitor your AI workload deployments and security status

**Security Scores**

Attestation and integrity verification status

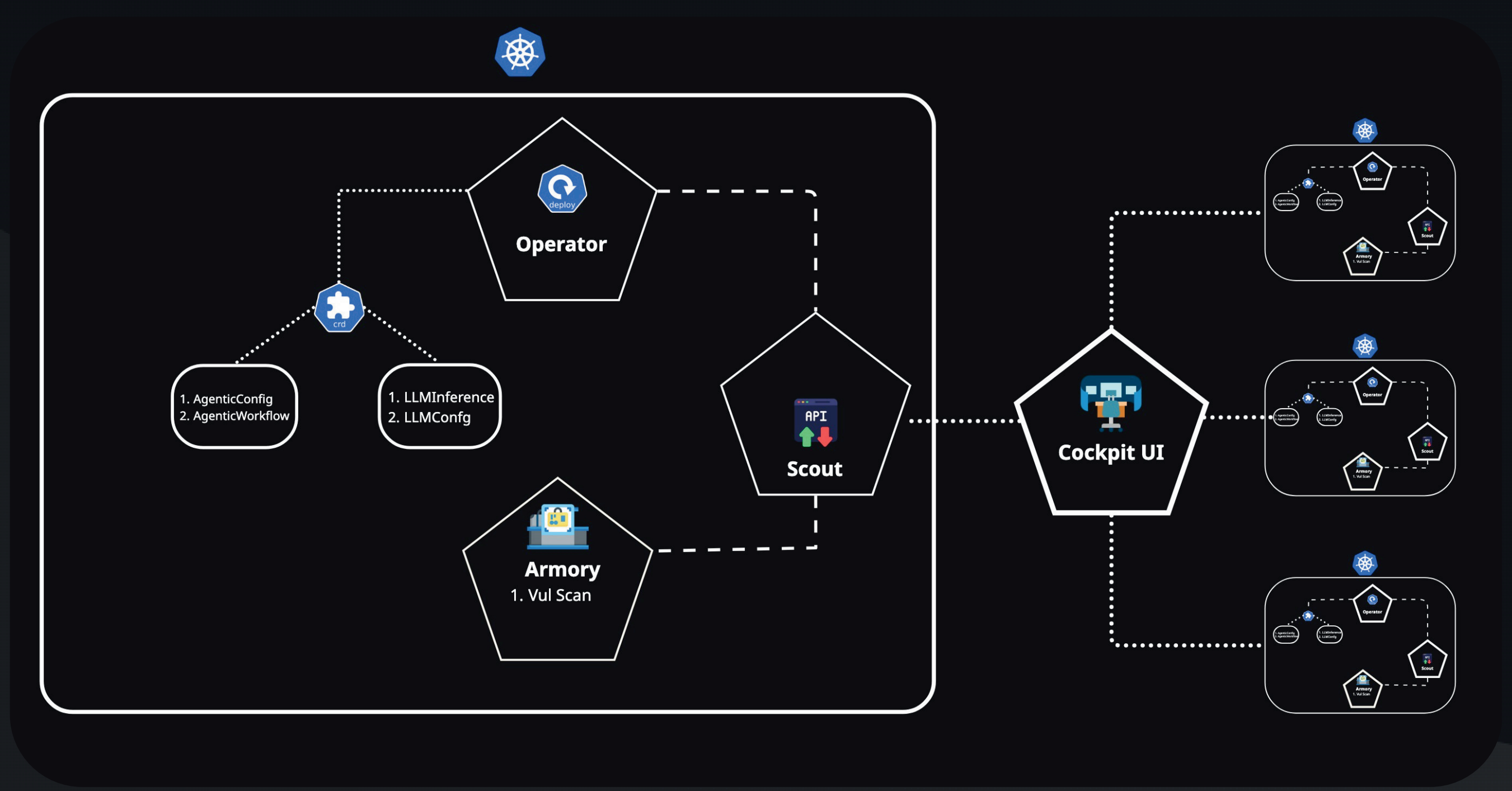
| | |
|---|---|
| Image Signing | 92% |
| Runtime Attestation | 87% |
| Supply Chain Verification | 78% |
| Vulnerability Scanning | 95% |

**Recent Deployments**

Monitor your recent AI workload deployments and their attestation status

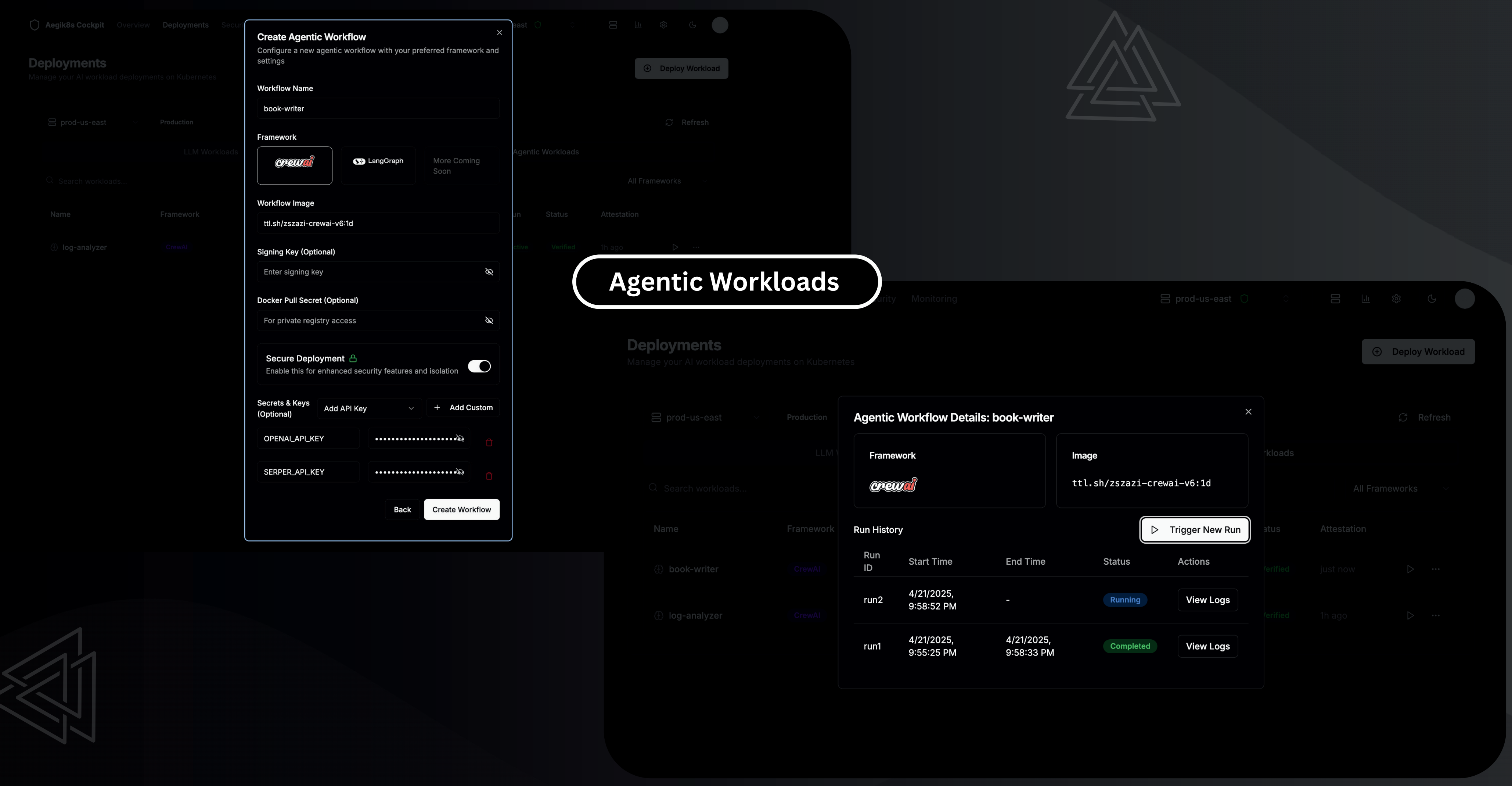
| Workload | Status | Attestation | Created |
|---|---|---|---|
| book-writer<br>crewai | Healthy | Verified | just now |
| smallmodel<br>h2oai/h2o-danube3-500m-chat | Healthy | Pending | 23m ago |
| log-analyzer<br>crewai | Healthy | Verified | 1h ago |
| fbmodel<br>facebook/opt-125m | Healthy | Pending | 1h ago |

## Architecture



- **Scout**: Aegik8s' control plane handles platform requests, manages workloads, and connects the UI with clusters
- **Operator**: Kubernetes-native controller that automates LLM and agent workload deployment, scaling, and teardown
- **Armory**: Pluggable security engine for scanning, attestation, and securing LLM and agent workloads
- **Cockpit**: UI for managing LLM and agent workloads across clusters in dev, staging, and production

## Agentic Workloads



**Create Agentic Workflow**

Configure a new agentic workflow with your preferred framework and settings

Workflow Name: book-writer

Framework: crewai · LangGraph · More Coming Soon

Workflow Image: ttl.sh/zszazi-crewai-v6:1d

Signing Key (Optional): Enter signing key

Docker Pull Secret (Optional): For private registry access

Secure Deployment: Enable this for enhanced security features and isolation

Secrets & Keys (Optional): Add API Key · + Add Custom

OPENAI_API_KEY ••••••••••••••••••••

SERPER_API_KEY ••••••••••••••••••••

Back · Create Workflow

**Agentic Workflow Details: book-writer**

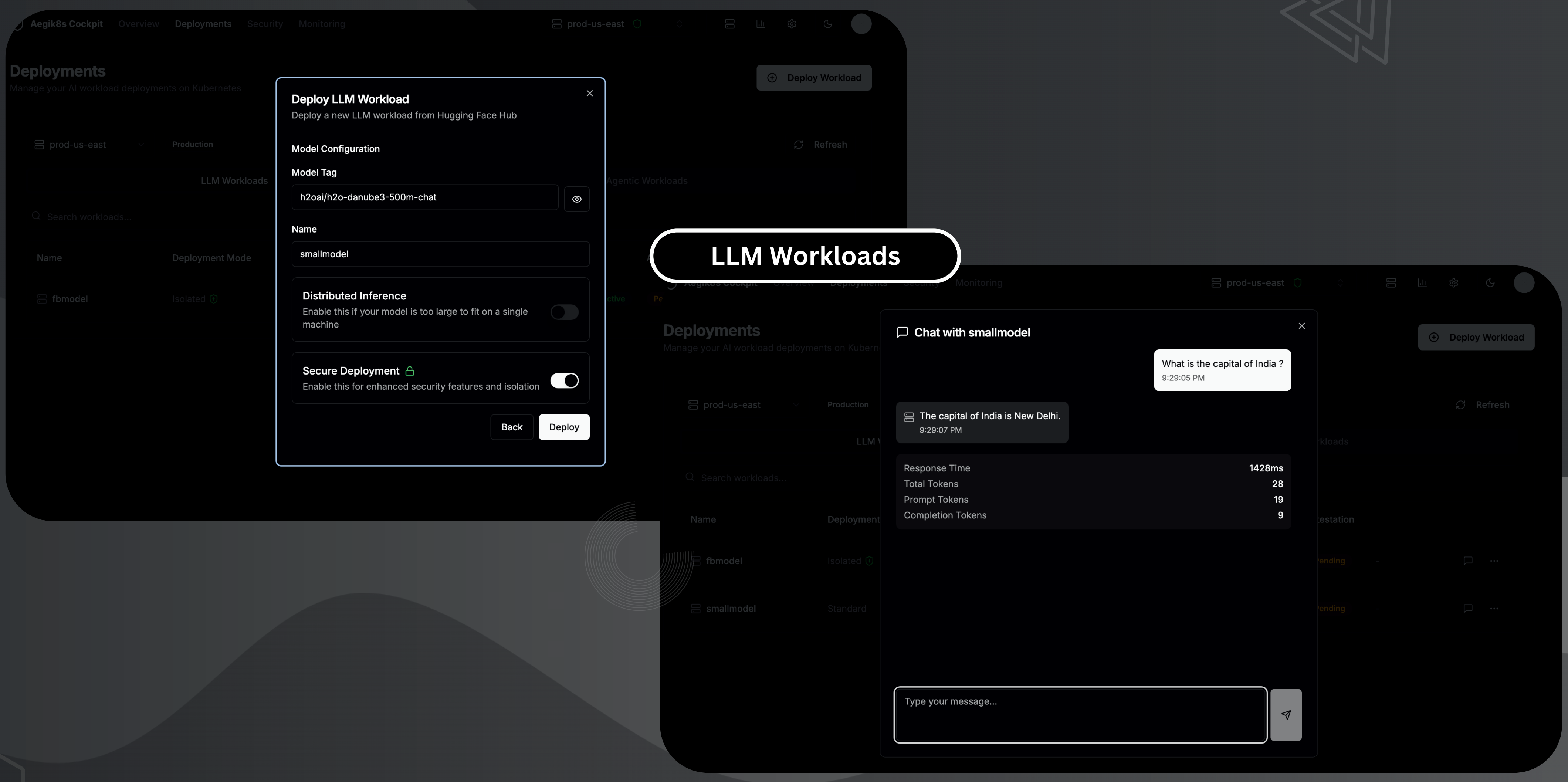
Framework: crewai

Image: ttl.sh/zszazi-crewai-v6:1d

Run History · ▷ Trigger New Run

| Run ID | Start Time | End Time | Status | Actions |
|---|---|---|---|---|
| run2 | 4/21/2025, 9:58:52 PM | - | Running | View Logs |
| run1 | 4/21/2025, 9:55:25 PM | 4/21/2025, 9:58:33 PM | Completed | View Logs |

## LLM Workloads



**Deploy LLM Workload**

Deploy a new LLM workload from Hugging Face Hub

Model Configuration

Model Tag: h2oai/h2o-danube3-500m-chat

Name: smallmodel

Distributed Inference: Enable this if your model is too large to fit on a single machine

Secure Deployment: Enable this for enhanced security features and isolation

Back · Deploy

**Chat with smallmodel**

What is the capital of India ?
9:29:05 PM

The capital of India is New Delhi.
9:29:07 PM

| | |
|---|---|
| Response Time | 1428ms |
| Total Tokens | 28 |
| Prompt Tokens | 19 |
| Completion Tokens | 9 |

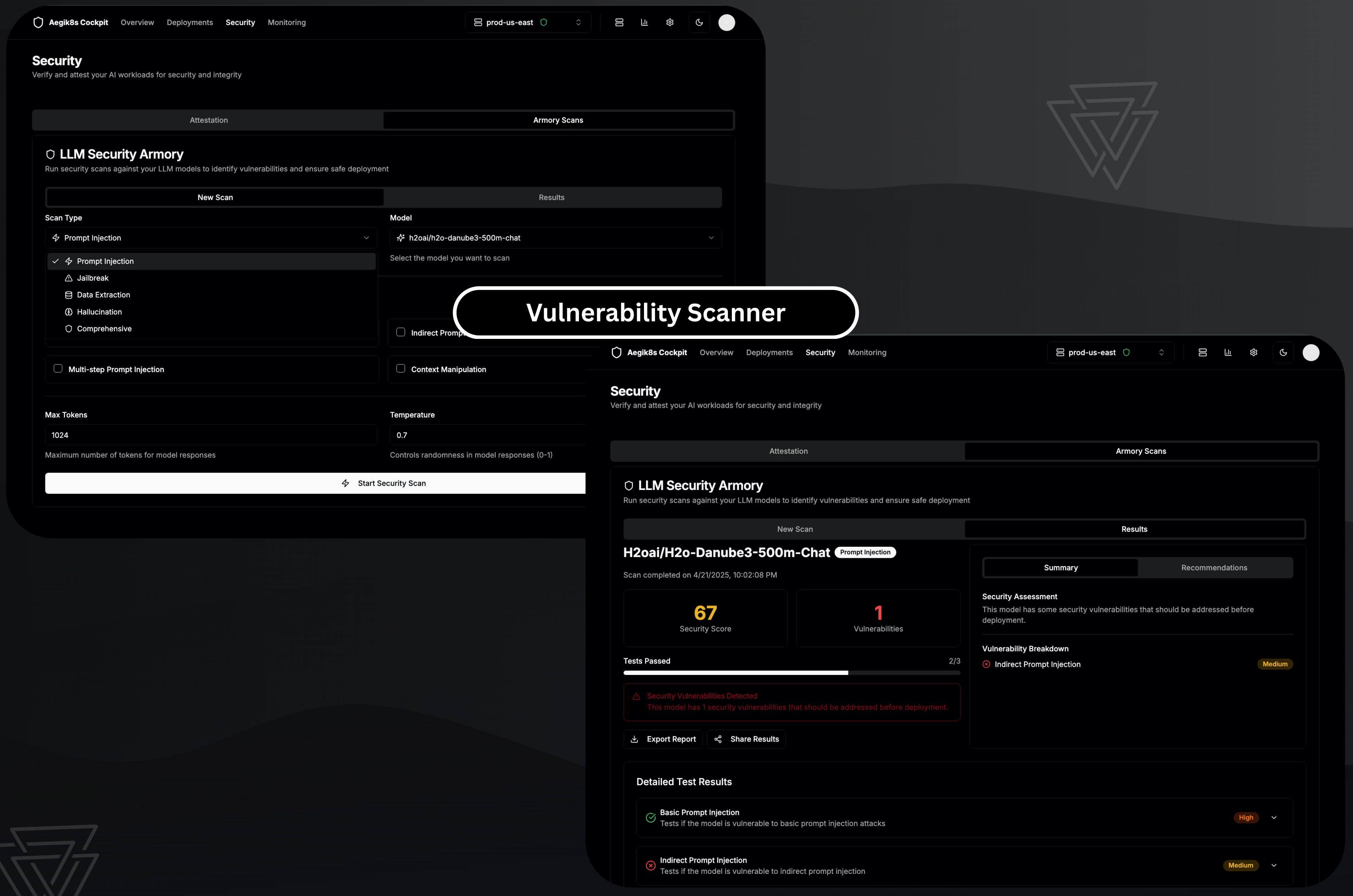
Type your message...

## Why Aegik8s ?

- **Deploy LLMs in 2 Clicks**
  - Spin up models effortlessly - no expertise needed
- **Secure by Default**
  - Enforces opinionated best practices
- **One-Command Setup**
  - Launch the full platform via a single Helm chart
- **Strong Isolation**
  - Each workload runs in its own MicroVM with dedicated kernel-level isolation for network, I/O, and memory
- **Minimized Blast Radius**
  - Uses Kubernetes and Confidential Computing to isolate threats at both hardware and software levels

Aegik8s runs anywhere—laptops, cloud (AWS/GCP), or even Raspberry Pi—and scales from single-CPU to multi-GPU setups for flexible, distributed LLM deployment

## Vulnerability Scanner



Aegik8s Cockpit · Overview · Deployments · Security · Monitoring · prod-us-east

**Security**

Verify and attest your AI workloads for security and integrity

Attestation · Armory Scans

**LLM Security Armory**

Run security scans against your LLM models to identify vulnerabilities and ensure safe deployment

New Scan · Results

Scan Type: Prompt Injection
- ✓ Prompt Injection
- Jailbreak
- Data Extraction
- Hallucination
- Comprehensive

Model: h2oai/h2o-danube3-500m-chat
Select the model you want to scan

□ Indirect Promp · □ Multi-step Prompt Injection · □ Context Manipulation

Max Tokens: 1024
Maximum number of tokens for model responses

Temperature: 0.7
Controls randomness in model responses (0-1)

Start Security Scan

**H2oai/H2o-Danube3-500m-Chat** Prompt Injection

Scan completed on 4/21/2025, 10:02:08 PM

67 Security Score · 1 Vulnerabilities

Tests Passed 2/3

Security Vulnerabilities Detected
This model has 1 security vulnerabilities that should be addressed before deployment.

Export Report · Share Results

Summary · Recommendations

Security Assessment
This model has some security vulnerabilities that should be addressed before deployment.

Vulnerability Breakdown
Indirect Prompt Injection Medium

Detailed Test Results

| | |
|---|---|
| Basic Prompt Injection<br>Tests if the model is vulnerable to basic prompt injection attacks | High |
| Indirect Prompt Injection<br>Tests if the model is vulnerable to indirect prompt injection | Medium |

## Market Potential

Only 24% of GenAI projects prioritize security, though 82% of executives see it as essential - highlighting a major gap. With the LLM and AI agent markets projected to hit $36B and $47B by 2030, Aegik8s is poised to meet this demand with a secure, scalable, and cost-effective platform for AI deployment.

## Target Personas

- **MLOps Engineers**: Secure, scalable model deployment with Helm, MicroVMs, and K8s workflows
- **AI Startups**: Fast, secure prototyping with built-in isolation
- **Local Devs**: Safe, lightweight LLMs on Minikube or kind

Sainath Santa Ram