Data Sheet

SUSE AI

Sovereign, Private AI for Every Cluster and Every Enterprise.

AI Suite

SUSE

21.04.2026

At A Glance

Built on the foundation of SUSE Rancher Prime, SUSE AI is purpose-built to deploy, run, and manage Generative AI (GenAI) workloads with a focus on security, sovereignty, and choice.

Move from PoC to production (ie., the production chasm) with the SUSE AI Factory that provides a unified product layer designed to industrialize Private Enterprise AI .

Product Overview

In 2026, the biggest challenge for enterprises is no longer how to build an AI model, but how to run it without losing control of their data.

SUSE AI turns the 'Wild West' of Generative AI into boringly reliable infrastructure. We provide the hardened plumbing that makes AI predictable and compliant, allowing teams to focus on innovation rather than troubleshooting.

- Sovereignty & Choice: SUSE AI does not require a specific vendor or cloud. It runs on-premises, in the cloud, or even in air-gapped environments. Choose the Large Language Models (LLMs) and components that fit your needs
- Zero-Trust Security: Using a "Never Trust; Always Verify" framework. SUSE AI includes built-in tools to detect PII (Personally Identifiable Information) leaks and can automatically shut down applications to ensure compliance with regulations like the EU AI Act or GDPR.
- The "AI Factory" Model: It delivers a turnkey digital assembly line. It connects the workflows of AI engineers with Platform Engineers, ensuring that the transition from development to global scale is seamless.

Key Benefits

SUSE AI inherits all the features from SUSE Rancher Prime, a recognized leader in the Gartner Magic Quadrant. With the introduction of the SUSE AI Factory, users get a comprehensive, end-to-end "assembly line" for the entire AI lifecycle, designed to move projects from local development to scalable enterprise production.

A platform you can trust.

SUSE AI provides you with a streamlined multi cluster and hybrid/multi-cloud management with enhanced enterprise-grade security. Enterprises also get long-term support, and

> "SUSE AI is a complete, CNCF-supported solution that works from the get go—whether on an enterprise-grade infrastructure or on an ARM device at the edge."
>
> **Johan van Amersfoort**
> Chief Evangelist, AI Lead
> ITQ

simplified developer operations through a unified interface. In essence, companies will get a production-ready platform with critical security updates, access to enterprise support, and features like robust authentication, policy management, and observability.

Insights that matter.

AI shouldn't be a 'black box' that keeps your Platform Engineers up at night. SUSE AI provides integrated observability and makes managing LLMs boringly transparent. Observability provides the transparency, control, and insights necessary to manage AI workloads effectively. This will maximize value, optimize costs, and ensure responsible, reliable AI operations at scale without the drama of unexpected downtime.

Sovereignty you can trust.

Between hardening and increasing regulations, including the EU AI Act, companies must ensure that their IP data and their customers' data stays under tight control. Losing control of data not only risks violating regulations, but also losing customer trust and competitive advantage.

With SUSE AI your data remains private. You run the entire AI workload (models, data, infrastructure) in your environment (on-premises, private cloud, or air-gapped) to prevent sensitive data from leaving your control.

In addition with zero trust security, SUSE AI protects the network and automatically learns what is "normal" behavior for each container and instantly blocks any unauthorized network connection or activity (process, file access) to prevent spread of attacks or data theft.

A factory for building AI workloads.

In order to bridge the gap between local development projects and scalable enterprise production, a comprehensive, end-to-end "assembly line" for the entire AI lifecycle is needed. The SUSE AI Factory is that assembly line.

SUSE AI Factory is a unified product layer designed to industrialize Private Enterprise AI integrating blueprints, tools and components, observability, security, and broader governance capabilities into a single implementation. It is designed to:

- Bridge the gap between developer experimentation and production deployment

- Deliver pre-validated, industry-aligned blueprints for Generative AI and Machine Learning
- Provide a highly specific, opinionated, and prescriptive Private Enterprise AI solution

With the SUSE AI Factory, the entire AI application lifecycle is simplified, from development to production. In addition, it also provides robust management, monitoring, and security for the entire AI workload.

Key Features

SUSE AI inherits all its proven features of SUSE Rancher Prime, a leader in container management solutions, including:

- **Unified Cluster Management:** A single pane of glass to deploy, operate, and upgrade Kubernetes clusters (RKE, K3s, AKS, EKS, GKE, etc.) wherever your AI workloads reside – on-premises, in public clouds, or at the edge.
- **GitOps-Driven Automation (Fleet):** Automate the deployment and configuration of AI applications across thousands of clusters, ensuring consistency, auditability, and rapid iteration.
- **Simplified Operations:** Streamline day-2 operations for your AI infrastructure. While the AI outputs are revolutionary, infrastructure management should be boringly consistent. Reduce manual effort and operational overhead with automated lifecycle management.

In addition, SUSE AI has specific features that make it valuable for AI workloads.

Security and trust.

Integrated zero trust security means SUSE AI operates on a “Never Trust; Always Verify” framework, delivering robust, zero-trust security across the entire application lifecycle, from development to daily operations.

SUSE Security provides for comprehensive container security, including:

- Network Segmentation: Micro-segmentation of container traffic to limit lateral movement in case of a breach.
- Vulnerability Management: Continuous scanning of images and running containers for known vulnerabilities.
- Runtime Security: Real-time threat detection and behavioral learning to identify and block anomalous activity.
- Policy Enforcement: Automated enforcement of security policies across your Kubernetes clusters.

Integrated security is particularly key for AI projects. Many popular AI projects and frameworks are open-source and constantly evolving. SUSE AI’s integrated security is invaluable for:

- Tracking & Verification: Providing the ability to continuously track the security posture of components from these upstream projects.
- Secure Control: Ensuring that even widely adopted open-source tools are running in a controlled, compliant, and secure environment within your enterprise.
- Mitigating Supply Chain Risks: Verifying the integrity of AI components from development to deployment

Integrated observability extensions.

SUSE's integrated observability has a powerful 4T Data Model, which provides a complete picture of your IT infrastructure and applications. Let's briefly recap the 4Ts:

- **Telemetry:** Provides the raw visibility needed to know exactly what is happening across your stack in real-time.
- **Tracing:** Maps the journey of a single request to show you where bottlenecks or errors are hiding in complex workflows.
- **Topology:** Delivers a dynamic "map" of your environment so you can see how your AI components are connected and interacting.
- **Time:** Adds the historical context required to analyze when performance shifted, allowing for accurate trend prediction and capacity planning.

The specific observability dashboards take the mystery out of the "black boxes" of agentic workflows. They provide a way to quickly identify performance bottlenecks, diagnose issues, optimize resource usage, and ensure the trustworthiness and effectiveness of AI applications in production.

In essence, SUSE Observability for AI provides:

- **LLM Cost and Token Usage Monitoring:** Gain clear insights into the financial impact of your LLM deployments. Track token consumption (input/output) to optimize model usage and predict costs, preventing budget overruns.
- **LLM Performance and Drift Detection:** Monitor real-time performance metrics (latency, throughput, accuracy) and detect model drift (degradation in prediction quality over time). Proactively identify and address issues before they impact business outcomes.
- **GPU Performance Monitoring:** GPUs are expensive and critical for AI workloads. SUSE AI provides deep insights into GPU utilization, temperature, and health, ensuring optimal resource allocation and preventing bottlenecks.
- **Vector Database Performance:** For Retrieval Augmented Generation (RAG) and other AI applications, vector databases are crucial. SUSE AI monitors their performance, ensuring efficient data retrieval and overall application responsiveness.
- **Prompt/Response Logging & Tracing:** Capture and analyze the prompts sent to LLMs and their corresponding responses. This is vital for debugging, understanding model behavior, identifying prompt injection attempts, and refining prompts for better results. It also enables auditing for compliance.

Integrated SUSE AI Factory

The SUSE AI Factory offers a number of key features that consolidate disjointed tools into a cohesive experience. These include:

- **Pre-Validated Blueprints:** Users receive a collection of tightly integrated architectural blueprints for common use cases, which reduces setup time from months to days.
- **A "One-Click" Install Experience:** This "easy button" for enterprise AI at scale provides a single point for installing frameworks, tools, and observability dashboards.
- **Iterative Dev-to-Prod Pipeline:** This bridges the "Persona Gap" by allowing developers to rapid-prototype in local sandbox environments and move seamlessly into a production environment.

- **Choice and Flexibility:** Users have absolute digital sovereignty, with the freedom to choose where their intelligence lives—whether in core data centers, private clouds, or fully air-gapped tactical edge sites.
- **Industrial-Scale Automation:** Support for both intuitive, UI-driven "ClickOps" for prototyping and declarative "GitOps" workflows for consistent, large-scale production management.

System Requirements

The SUSE AI stack consists of multiple applications. We recommend running each application on nodes that meet or exceed the corresponding hardware requirements.

Minimum hardware requirements for High Availability

- RAM: 64 GB or more per node is recommended for larger clusters or to run applications with high resource demands.
- CPU: At least 8 cores, ideally 16 or more cores, depending on the expected load.
- Disk Space: For larger-scale clusters or persistent storage applications, 100 GB or more of disk space per node might be required. Using high-performance SSDs is recommended, especially for workloads with high I/O requirements, such as databases or AI/ML model training.
- Networking: Ensure a low-latency, high-throughput network for efficient communication between nodes, especially if deploying in multi-region or multi-cloud environments.

Disclaimer: These requirements are for the Management Layer only, and that workload nodes should be sized based on specific LLM parameters and GPU requirements.

For detailed product specifications and system requirements, visit: https://documentation.suse.com/suse-ai/1.0/html/AI-requirements/index.html#ai-reqs-rancher-hw

SUSE

2026

SUSE Software Solutions
Germany GmbH

Frankenstraße 146
90461 Nürnberg
Germany

www.suse.com

For more information, contact SUSE at:

+1 800 796 3700 (U.S./Canada)

+49 (0)911-740 53-0 (Worldwide)

Innovate Everywhere