Generative AI security controls

Information Security for Workload Factory

NetApp
September 16, 2026

Table of Contents

Generative AI security controls

Learn about AI controls in NetApp Workload Factory

NetApp Workload Factory includes generative AI features that accelerate diagnosis while enforcing security controls for model access, data scope, and runtime behavior.

Workload Factory enables AI diagnostics by default. You can disable generative AI features at any time through the Workload Factory **Administration** settings. Learn more about managing GenAI features.

Security-scoped AI feature overview

Workload Factory currently applies generative AI to the following features:

- Ask Me assistant (RAG with curated NetApp documentation, source-backed responses)
- Latency analysis
- EMS event analysis
- Error log analysis

AWS Bedrock overview for AI diagnostics

Workload Factory uses Amazon Bedrock for AI inference. Inference runs in your AWS account, under your configured credentials and inference profile.

Related information

- Amazon Bedrock opt-in for NetApp Workload Factory AI diagnostics
- AI tool boundaries for NetApp Workload Factory
- AI model parameters and billing for NetApp Workload Factory

Amazon Bedrock opt-in for NetApp Workload Factory AI diagnostics

NetApp Workload Factory AI diagnostics use Amazon Bedrock to analyze events, and the feature is enabled by default. Before Workload Factory can invoke the selected Bedrock model and collect diagnostic data, you must configure IAM permissions, an inference profile, and a connected Lambda link with SSH capability. If any of these components is missing or the Bedrock configuration is removed, AI diagnostics become unavailable.

AI diagnostics are enabled by default.

Before you begin

- Make sure your account includes an AWS Bedrock inference profile.
- Have a connected Lambda link with SSH capability for diagnostic data collection.

Steps

1. Grant Bedrock permissions in your IAM role:
    - `bedrock:InvokeModel`
    - `bedrock:InvokeModelWithResponseStream`
    - `bedrock:ListInferenceProfiles`
    - `bedrock:GetInferenceProfile`
2. Configure the inference profile ARN in Workload Factory settings.

If any prerequisite is missing, the system reports the specific missing component and AI features remain inactive.

Disable AI diagnostics

To disable AI diagnostics:

- Remove the Bedrock permissions from your IAM role, or
- Delete the Bedrock configuration in Workload Factory settings.

AI diagnostics become unavailable immediately, and no residual data processing occurs.

Region and cross-region inference profiles

Bedrock inference executes in the AWS region specified by your inference profile. If you use a cross-region inference profile, AWS might route the request to any region in the profile's configuration per standard AWS Bedrock behavior. Data does not leave AWS infrastructure.

AI tool boundaries for NetApp Workload Factory

NetApp Workload Factory enforces strict boundaries on how its AI features interact with your AWS and ONTAP environments, so you can trust AI-assisted diagnostics without risking unintended changes. Amazon Bedrock uses read-only AWS CLI and ONTAP CLI diagnostic tools that block mutating commands and limit output by size and execution, and any data used is only transient and not retained by NetApp Workload Factory or AWS.

Tool safety controls

Tool	Capability	Safety controls
AWS CLI (read-only)	Execute read-only AWS CLI commands (describe, list, get)	Blocks all mutating verbs (create, delete, modify, update, put, remove). Maximum 10 commands per step. Output truncated at 20 KB.
ONTAP CLI (read-only)	Execute read-only ONTAP CLI commands using SSH	Blocks all mutating verbs (delete, destroy, create, modify, move). Output truncated.

The agent cannot modify, create, or delete resources. It can only observe and diagnose.

Data retention and model training boundaries

Per AWS policy, Amazon Bedrock does not store or use your inputs and outputs to train or improve foundation models. For on-demand inference (used by Workload Factory), AWS processes your data transiently and does not retain it after returning the response.

AI model parameters and billing for NetApp Workload Factory

NetApp Workload Factory defines the model configuration and usage and billing boundaries relevant to AI governance. The configured model uses deterministic parameters and returns structured JSON to support consistent analysis results. AWS bills Amazon Bedrock inference to your account, while Workload Factory tracks AI usage and provides per-account monthly cost visibility. These sections describe the model parameters and the limits that apply to AI consumption.

Model and parameters

Parameter	Value
Model	Anthropic Claude (using a cross-region inference profile)
Temperature	0 (deterministic, no randomness)
Max output tokens	2,048
Max reasoning steps	15 per analysis
Output format	Structured JSON (summary, severity, root cause, corrective actions)

Billing and usage limits

- AWS bills Bedrock inference to your account (your inference profile, your credentials).
- Workload Factory tracks AI usage internally.
- Workload Factory provides per-account monthly cost visibility.

  Track storage costs in NetApp Workload Factory

Ask Me AI assistant

Ask Me security architecture for NetApp Workload Factory

Ask Me is a generative AI-powered assistant embedded within NetApp Workload Factory. It provides real-time, conversational support for Amazon FSx for NetApp ONTAP and Workload Factory topics. Responses are grounded in verified NetApp documentation, and Ask Me cannot access the public internet or use customer data to train the model. Multiple security layers help block malicious queries, limit responses to supported domains, and prevent user input from overriding system instructions. When Ask Me uses

tools, authorization boundaries, read-only query enforcement, an ephemeral sandbox, and audit logging help constrain and monitor execution.

You can opt out of Ask Me at any time through the Workload Factory **Administration** settings, which disables the assistant for your user account. Learn more about managing GenAI features.

RAG architecture in Ask Me

Ask Me uses retrieval-augmented generation (RAG).

- Curated knowledge base: responses are grounded in a managed corpus of verified, published NetApp documentation.
- Retrieval scoring and filtering: only sufficiently relevant content is included in the model context.
- No internet access: the model cannot fetch, browse, or scrape external content.
- Source attribution: responses include references to the source documents used.

Multi-layer prompt security

- Layer 1: Security classifier (LLM-as-a-judge) blocks malicious queries (prompt injection, privilege escalation, data exfiltration, social engineering, policy violations).

  The system logs blocked queries, and the generative model never sees them.

- Layer 2: System prompt hardening prevents user input from overriding system instructions.
- Layer 3: Domain scoping limits responses to FSx ONTAP and Workload Factory topics.
- Layer 4: Tool-use governance validates parameters and runs queries in a hardened sandbox; the model cannot execute arbitrary operations.

Model data isolation and privacy

- No model training on customer data
- Request-level isolation (no cross-tenant data leakage)
- Session-scoped conversation context with bounded history
- No persistent model memory
- Managed inference infrastructure

Tool-use and agentic security

- Authorization-scoped tools (user permission boundaries enforced)
- Strict tool execution timeouts
- Ephemeral, session-scoped sandbox for retrieved data
- Read-only query enforcement with allowlisting
- Tool call auditability with sensitive parameter scrubbing

Ask Me access control for NetApp Workload Factory

In NetApp Workload Factory, Ask Me access security centers on authenticated sessions,

user-level accountability, and protected secret handling during runtime operations. Authentication controls validate each session and associate interactions with a specific user. Sensitive credentials are retrieved from a managed secrets vault at runtime and redacted from logs. The service also applies infrastructure controls, including non-root container execution and a restricted CORS allowlist.

Authentication

- Signed JWT authentication with cryptographic validation (including audience, issuer, and algorithm checks such as RS256)
- Middleware-enforced authentication at the service boundary
- User identity scoping so interactions are attributable to a specific user

Credential protection and secret handling

- Secure vault storage: sensitive credentials used by the service are stored in a managed secrets vault and retrieved at runtime. No secrets are stored in application code, configuration files, or container images.
- Log scrubbing: sensitive values (credentials, tokens, passwords, access keys, certificates) are redacted from logs before they are written.

Infrastructure security

- Non-root container execution
- CORS restricted to an explicit allowlist of trusted NetApp domains

Ask Me execution modes for NetApp Workload Factory

NetApp Workload Factory provides multiple execution modes for Ask Me, each with different security and privacy characteristics. When Ask Me uses tool integrations, for example, querying the customer's own Workload Factory resources, tool execution is controlled through layered safeguards. Tool-enabled execution operates within the authenticated user's permissions and applies limits to the duration and scope of each operation. Retrieved data remains in an ephemeral session sandbox that blocks external access and is destroyed when the session ends. Independent prompt and code controls enforce read-only queries, while audit records capture tool activity with sensitive values scrubbed.

Tool execution safeguards

- Authorization-scoped tools operate within the authenticated user's permissions.
- Tool execution timeout limits long-running operations.
- Ephemeral data sandbox stores tool-retrieved data in session scope, blocks external access, file I/O, network calls, and extension loading at the engine level, and is destroyed when the session ends.
- Read-only query enforcement validates generated queries through a strict allowlist.
- Prompt-enforced and code-enforced security are applied independently.
- Tool call auditability logs tool name, parameters, timestamps, and result status with sensitive value scrubbing.

Output controls

- Token limit enforcement
- Deterministic output for classification/security-critical operations (temperature 0)
- Streaming with early termination and cleanup

Ask Me privacy and availability for NetApp Workload Factory

In NetApp Workload Factory, Ask Me privacy controls limit data exposure, keep session processing isolated, and maintain privacy boundaries for user interactions. Your inputs are processed by a managed AI service that does not use them to train or improve foundation models. Retrieved operational data remains in ephemeral, in-memory storage for the duration of the session and is not persisted or shared. A multi-model fallback chain across multiple cloud regions supports availability when the primary model is throttled or unavailable, while the same security controls apply across the fallback models.

Data handling and privacy

- User data in prompts: processed by a managed AI service that does not use your inputs to train or improve foundation models.
- Knowledge base content: curated, published NetApp documentation only; your data isn’t included.
- Operational data: retrieved only when a user queries their own resources; held in ephemeral, in-memory storage for the duration of the session and not persisted or shared.
- Audit logging: query metadata (anonymized user ID, timestamps, durations) logged, with sensitive fields scrubbed.
- Feedback data: stored separately and linked to a query ID, not to personally identifiable information beyond the anonymized user ID.

Availability and isolation controls

Ask Me employs a multi-model fallback chain across multiple cloud regions.

If the primary model is throttled or unavailable, the system can fail over to alternative models.

All models are subject to the same security controls, system prompt hardening, and isolation guarantees.

Copyright information

Trademark information