SUSE AI Factory Release Notes

SUSE AI Factory Release Notes

Publication Date: 2026-11-12

https://documentation.suse.com

Contents

1 Release 2.4.0 1

1.1 What's new 1

1.2 Key fixes 2

1.3 Upgrade notes 2

1.4 Known issues and limitations 3

2 Release 2.3.1 4

2.1 Key improvements and fixes 4

2.2 Upgrade notes 5

3 Release 2.3.0 6

3.1 What's new 6

Blueprint customization 6 • Custom blueprint catalogs 7 • Custom chart repositories 7 • Repository verification and readiness 8 • Workload monitoring 8 • Air-gapped deployments 9 • Application catalog 9 • Packaging 9

3.2 Key fixes 9

3.3 Upgrade notes 10

3.4 Known issues and limitations 11

4 Release 2.2.0 13

4.1 What's new 13

Application catalog 13 • Air-gapped deployments 14 • Workload recovery 14 • Usability improvements 15 • Packaging 15

4.2 Key fixes 15

4.3 Upgrade notes 16

4.4 Known issues and limitations 17

5 Release 2.1.0 18

5.1 What's new 18

Application catalog 18 • Git-backed Rancher Prime catalogs and Rancher Prime API access 19 • Extension delivery and security 19 • Workload and blueprint management 20

5.2 Key fixes 20

5.3 Upgrade notes 21

6 Release 2.0.1 22

6.1 Key improvements and fixes 22

7 Release 2.0.0 23

7.1 What's new 23

Core features 23

7.2 Known issues and behaviors 25

Installation and deployment 25 • NVIDIA blueprints 25 • NVIDIA AI Library 26 • Blueprint customization 26 • Deprecations 26

1 Release 2.4.0

This release builds on the 2.3.x line with new partner blueprints, new SUSE AI applications in the catalog, and unique default namespaces for new installations, plus fixes for secret references on the **Settings** page, large Git-hosted charts and extension access checks.

1.1 What's new

- **New partner blueprints**: The SUSE partner blueprint catalog now includes blueprints from three partners: `AlchemiStudio + vLLM` (Zentience Labs), `Tensormesh Platform on SUSE AI (vLLM)` (Tensormesh), and `Traefik Hub AI and MCP Gateway Installer` and `Traefik Hub AI and MCP Governance Demo` (Traefik Labs). Support for each blueprint is provided by its partner. Large partner charts such as AlchemiStudio now install from Git chart repositories #348 (https://github.com/SUSE/aif/pull/348).

- **New SUSE AI applications**: `vLLM Production Stack`, which supersedes the existing vLLM application, plus `Slurm Operator` and `Slurm Bridge` for running Slurm alongside Kubernetes #357 (https://github.com/SUSE/aif/pull/357).

- **NVIDIA catalog refresh**: The weekly NVIDIA AI Enterprise catalog refreshes are included. `DOCA Weave Service` (`dpf-weave`) is added to the NVIDIA AI Library #335 (https://github.com/SUSE/aif/pull/335), #350 (https://github.com/SUSE/aif/pull/350), #358 (https://github.com/SUSE/aif/pull/358).

- **Random default namespace for new installations**: The application and blueprint installation wizards now suggest a namespace with a random suffix, such as `coredns-k3x9q2a`, instead of `coredns-system`. Each new installation gets its own namespace, so installing the same application twice no longer fails because a deployment with the same name already exists in the shared namespace. You can still change the namespace, and existing installations keep theirs #338 (https://github.com/SUSE/aif/pull/338).

- **Incomplete secret references are caught before saving**: *Apply* on the **Settings** page is now blocked when a secret field has a secret selected but no key. The message names each such field and opens its section. The operator API rejects these partial references with HTTP `400`, and the CRDs reject them from any client #349 (https://github.com/SUSE/aif/pull/349).

1.2 Key fixes

- Setting the SUSE Application Collection, SUSE Registry or NVIDIA credentials, or the Rancher Prime catalog token and CA certificate secrets, to `None` now saves. Previously, the page reloaded with the old credentials filled in, or saving failed with errors such as `spec.applicationCollection.tokenSecretRef.name: Required value`. This also happened when the reference was set by Helm values or discovered by the operator #349 (https://github.com/SUSE/aif/pull/349).
- Clearing every field of a section on the **Settings** page no longer fails with `spec.rancherCatalog in body must be of type object` #349 (https://github.com/SUSE/aif/pull/349).
- Secret fields of custom repositories no longer silently drop a secret that has no key. Previously, saving appeared to succeed but stored `None` #349 (https://github.com/SUSE/aif/pull/349).
- Charts from Git chart repositories that are larger than 1 MiB unpacked now install. Previously, they failed with `ChartTooLarge` or `etcdserver: request is too large`. Chart files are now stored compressed, so a 2 MB chart takes about 0.43 MB #348 (https://github.com/SUSE/aif/pull/348).
- Global administrators are no longer refused access to the UI extension on their first navigation, for example, right after creating a cluster, with the message `Access check failed`. The extension now waits up to 30 seconds for Rancher Prime to load its management data before checking access, and a failed check is retried on the next navigation #302 (https://github.com/SUSE/aif/pull/302).

1.3 Upgrade notes

- **Updated `Settings` and `InstallAIExtension` CRDs**: The `name` and `key` of secret references must not be empty, and the `fleet`, `applicationCollection`, `suseRegistry`, `nvidia`, `appCatalog` and `rancherCatalog` sections of `Settings` accept `null`. The CRD hook job of the chart applies the CRDs on upgrade #349 (https://github.com/SUSE/aif/pull/349).
- **Existing partial secret references**: On Kubernetes 1.30 and later, a `Settings` resource that already stores a secret reference without a key, or a key without a secret, can still be updated as long as that field is not changed. Complete or clear these references the next time you edit the settings #349 (https://github.com/SUSE/aif/pull/349).

- **Use matching operator and UI extension versions**: The fixes for `None` and partial secret references change both the operator API and the **Settings** page, and require both the 2.4.0 operator and the 2.4.0 UI extension #349 (https://github.com/SUSE/aif/pull/349).
- **New default namespace pattern**: New installations get a `<APP>-<RANDOM>` namespace instead of `<APP>-system`. If your scripts or policies expect `<APP>-system`, set the namespace explicitly in the installation wizard. The *Manage* action still finds instances installed in the earlier `<APP>-system` namespace #338 (https://github.com/SUSE/aif/pull/338).

1.4 Known issues and limitations

- Charts from Git chart repositories are limited to a 4 MiB archive, 32 MiB unpacked and 1 MiB of compressed files in the Fleet `Bundle`.
- Non-administrator users get access to the UI extension through a `cluster-owner` role binding for their user on the `local` cluster. Users who are cluster owners only through a group, for example, an LDAP group or a GitHub organization, are refused access.
- The vLLM application is superseded by `vLLM Production Stack` and will be deprecated in a future release.

Full Changelog: aif-operator-2.3.1...aif-operator-2.4.0 (https://github.com/SUSE/aif/compare/aif-operator-2.3.1...aif-operator-2.4.0)

2 Release 2.3.1

This is the 2.3.1 release of SUSE AI Factory, bringing comprehensive AI workload management capabilities to Rancher Prime.

2.1 Key improvements and fixes

- Added a default SUSE partner blueprint catalog, controlled by the `partnerBlueprints` Helm value and enabled by default. Catalogs declared through Helm are now read-only on the **Settings** page and in the operator API, and are marked `Managed by Helm` #322 (https://github.com/SUSE/aif/pull/322).
- Added a `Community` blueprint source. Older operators reject blueprints that use `source: Community` #324 (https://github.com/SUSE/aif/pull/324).
- The Rancher Prime `Extensions` page now shows the logo and title of the SUSE AI Factory extension instead of a `Third-Party` card. The operator installs the extension the same way Rancher Prime does, and the logo is served from inside the cluster, so it also loads in air-gapped installations #325 (https://github.com/SUSE/aif/pull/325).
- Hardened the installation and cleanup of the UI extension. The operator refuses an extension that has the same name as its server release, and the finalizer only deletes a release that its own resource installed #325 (https://github.com/SUSE/aif/pull/325).
- Fixed workloads stuck in `Degraded` on clusters whose Fleet name differs from the Rancher Prime cluster ID. The installation wizards now show cluster display names instead of `c-m-…` IDs #330 (https://github.com/SUSE/aif/pull/330).
- Fixed the verification panel on the **Settings** page hiding the NGC chart access results when no NVIDIA mirror is set #321 (https://github.com/SUSE/aif/pull/321).
- The weekly NVIDIA AI Enterprise catalog refresh no longer removes applications that lose NVIDIA AI Enterprise support. It only removes their `Supported` label. Each refresh also produces a report of what changed #309 (https://github.com/SUSE/aif/pull/309), #323 (https://github.com/SUSE/aif/pull/323).

2.2 Upgrade notes

- **Partner blueprint catalog in air-gapped installations**: The partner catalog is synchronized from GitHub. In air-gapped installations, disable it with `partnerBlueprints.enabled=false` or point it at an internal mirror #322 (https://github.com/SUSE/aif/pull/322).
- **Duplicate partner catalogs**: If you added a catalog on the **Settings** page that points to the SUSE partner-blueprints repository, remove it after upgrading. Otherwise, two Fleet `GitRepo` resources synchronize the same blueprints and conflict #322 (https://github.com/SUSE/aif/pull/322).
- **Server-side apply for Helm-declared catalogs**: Keeping catalogs added on the **Settings** page next to catalogs declared in `blueprintCatalogs` requires server-side apply. Releases first installed with Helm 3 keep using client-side apply until you run one `helm upgrade --server-side=true` #322 (https://github.com/SUSE/aif/pull/322).
- `Community` **blueprint source**: SUSE AI Factory Operator 2.3.0 and earlier reject blueprints with the `Community` source. Do not use it until 2.3.1 is the minimum operator version in your environment #324 (https://github.com/SUSE/aif/pull/324).

Full Changelog: aif-operator-2.3.0...aif-operator-2.3.1 (https://github.com/SUSE/aif/compare/aif-operator-2.3.0...aif-operator-2.3.1)

3 Release 2.3.0

This release builds on the 2.2.x line with customizable blueprint deployments, custom chart repositories and Git-sourced blueprint catalogs, clearer repository and workload health reporting, and a new workload details view, plus several stability and installation improvements.

3.1 What's new

3.1.1 Blueprint customization

- **Customize blueprint deployments**: A new `Configuration` step in the blueprint installation and management wizard lets you override Helm values and enable or disable individual blueprint components before deployment. Disabled components are left out of the deployment and removed from the cluster. Blueprint workloads can be reconfigured later with the *Manage* action #265 (https://github.com/SUSE/aif/pull/265).

- **Configuration drift tracking**: Workloads report whether their configuration differs from the blueprint defaults, and the `Workloads` page shows a `Customized` badge next to the blueprint source. Only the values you changed are saved, not the full set of component values #287 (https://github.com/SUSE/aif/pull/287).

- **Partner icons and `Partner` source**: Blueprints accept an optional `icon`, either an `https://` URL or a base64-encoded raster `data:` URI of up to 16 KB, which is shown next to the source badge, and a new `Partner` source value. Updating a blueprint with an invalid specification now returns HTTP `422` instead of `500` #292 (https://github.com/SUSE/aif/pull/292).

3.1.2 Custom blueprint catalogs

- **Git-sourced blueprint catalogs**: Administrators can add external Git repositories that contain blueprint definitions in the `Custom Catalogs` section of **Settings** or through the `blueprintCatalogs` Helm value. The operator synchronizes each catalog into the cluster through its own Fleet `GitRepo` resource, with its own credential and CA certificate secrets #281 (https://github.com/SUSE/aif/pull/281).

- **New `BlueprintCatalog` CRD**: A cluster-scoped resource that describes a curated, branded collection of blueprints #281 (https://github.com/SUSE/aif/pull/281).

- **Browse blueprints by catalog**: The `Blueprints` page provides a catalog filter. Blueprints bundled with the chart are listed under `SUSE Blueprints` #281 (https://github.com/SUSE/aif/pull/281).

3.1.3 Custom chart repositories

- **Manage chart repositories from Settings**: Administrators can add, edit and delete Helm (HTTP or HTTPS), OCI and Git chart repositories, including credentials, in the `Repositories` section of **Settings**. The operator creates a `ClusterRepo` resource with the same name for each repository, and its applications are listed in the application catalog under their own library #244 (https://github.com/SUSE/aif/pull/244).

- **Single-chart OCI repositories**: Repositories whose URL already ends with the chart name, for example `oci://ghcr.io/nvidia/openshell/helm-chart`, install correctly with both Helm and Fleet. Use this to add the NVIDIA OpenShell charts #244 (https://github.com/SUSE/aif/pull/244), #293 (https://github.com/SUSE/aif/pull/293).

- **Verify before saving**: Custom repositories can be tested for credentials, reachability and access to a sample chart from the same form #244 (https://github.com/SUSE/aif/pull/244).

3.1.4 Repository verification and readiness

- **Three-part connection test**: In **Settings**, *Test* reports the registry connection, access to a sample chart, and the readiness of the Rancher Prime chart repository as separate results, so a successful login alone is no longer reported as success. The result names the chart and version that were tested, and denied access, missing content, invalid indexes, TLS errors and timeouts each have their own explanation #242 (https://github.com/SUSE/aif/pull/242).

- **Repository failures shown across the UI**: The `Overview` page lists failing managed repositories with links to **Settings** and Rancher Prime. Applications from an unavailable repository remain visible but cannot be installed, and the installation wizard explains missing repository indexes instead of reporting `configmaps "" not found` #242 (https://github.com/SUSE/aif/pull/242).

- **Explicit refresh**: *Test* is read-only. A new *Refresh* action asks Rancher Prime to download the index of a saved repository again #242 (https://github.com/SUSE/aif/pull/242).

3.1.5 Workload monitoring

- **Workload details**: Selecting a workload name on the `Workloads` page opens a details panel with the source, deployment strategy, target namespace and clusters, blueprint components, per-cluster status messages, value overrides, and the underlying Fleet resources or Helm release #221 (https://github.com/SUSE/aif/pull/221), #303 (https://github.com/SUSE/aif/pull/303).

- **Accurate status for Helm workloads**: Application workloads deployed with Helm now derive their status from pod and controller readiness. A crash-looping or unschedulable pod reports `Degraded` with the reason, and a pod that is still starting reports `Pending` for as long as it needs. Previously, a workload was reported as `Running` as soon as its Helm release existed #260 (https://github.com/SUSE/aif/pull/260).

3.1.6 Air-gapped deployments

- **NVIDIA repository aliases preserved in mirror mode**: Switching to a private NVIDIA mirror keeps every repository alias derived from the catalog, such as `nvidia-runai` or `nim-nvidia`, so existing workloads can still be upgraded and redeployed #237 (https://github.com/SUSE/aif/pull/237).
- **Bundled catalog logos**: Logos for 81 SUSE AI charts are bundled with the UI extension, so application tiles show the correct logo in connected and air-gapped installations, including behind private mirrors #245 (https://github.com/SUSE/aif/pull/245).

3.1.7 Application catalog

- **NVIDIA catalog refresh**: The weekly NVIDIA AI Enterprise catalog refreshes are included. NGC applications that lost their NVIDIA AI Enterprise support label, including several NeMo microservices, two NIM embedding models and `dgxc-admission-controller`, remain in the catalog without the `Supported` label. The NVIDIA GPU Operator no longer has the `Supported` label #301 (https://github.com/SUSE/aif/pull/301).

3.1.8 Packaging

- **Rancher Prime readiness check**: A pre-install and pre-upgrade hook job waits until the required Fleet and Rancher Prime catalog CRDs and the `cattle-ui-plugin-system` namespace exist before the operator is installed, and retries on temporary API errors #277 (https://github.com/SUSE/aif/pull/277), #289 (https://github.com/SUSE/aif/pull/289).

3.2 Key fixes

- The UI extension no longer remains in the `Failed` state after the operator re-creates its `aif-ui-config` ConfigMap. The operator now takes Helm ownership of an existing ConfigMap instead of failing every installation and upgrade with an `invalid ownership metadata` error #218 (https://github.com/SUSE/aif/pull/218).
- Custom UI extension settings are retained after uninstalling. The `aif-ui-config` ConfigMap was previously deleted with the release, so custom operator connection and catalog settings were lost on reinstallation #218 (https://github.com/SUSE/aif/pull/218).

- Failing Helm workloads are no longer reported as `Running`. For example, a NIM pod that did not fit in GPU memory was reported as `Running` while it was crash-looping #260 (https://github.com/SUSE/aif/pull/260).
- Kubeflow, LiteLLM and Qdrant no longer fail on the first installation wizard screen when chart access is denied or the Rancher Prime repository is disconnected despite a successful registry login #242 (https://github.com/SUSE/aif/pull/242).
- Applications such as Ollama show their own logo instead of a generic icon when their metadata contains a logo URL #245 (https://github.com/SUSE/aif/pull/245).
- Switching to an NVIDIA mirror no longer breaks later upgrades of applications installed from NVIDIA team repositories #237 (https://github.com/SUSE/aif/pull/237).

3.3 Upgrade notes

- **Mirror the `kubectl` image for air-gapped installations**: The new Rancher Prime readiness hook job uses the `registry.suse.com/suse/kubectl:1.35` image, which you can configure with the `global.kubectl.image` values. The job waits until Rancher Prime is ready, limited only by the installer timeout, such as `helm --timeout` or the Fleet `helm.timeoutSeconds` value #277 (https://github.com/SUSE/aif/pull/277), #289 (https://github.com/SUSE/aif/pull/289).
- **New CRD and settings**: The `blueprintcatalogs.ai-factory.suse.com` CRD and the `blueprintCatalogs` and `customRepos` settings are added, and the `AIWorkload` and `Blueprint` CRDs are updated. The CRD hook job of the chart applies them on upgrade #281 (https://github.com/SUSE/aif/pull/281), #244 (https://github.com/SUSE/aif/pull/244), #265 (https://github.com/SUSE/aif/pull/265).
- **Use matching operator and UI extension versions**: Chart access checks, blueprint customization and partner icons require both the 2.3.0 operator and the 2.3.0 UI extension #242 (https://github.com/SUSE/aif/pull/242), #265 (https://github.com/SUSE/aif/pull/265), #292 (https://github.com/SUSE/aif/pull/292).
- **Helm workload status may change**: Application workloads that were previously reported as `Running` may now report `Pending` or `Degraded` if their pods are not ready #260 (https://github.com/SUSE/aif/pull/260).

- `aif-ui-config` **is retained on uninstall**: Uninstalling the UI extension now leaves the `aif-ui-config` ConfigMap in place. Delete it manually for a clean reinstallation #218 (https://github.com/SUSE/aif/pull/218).
- **Standalone installations of the UI extension chart**: Installing or upgrading the `aif-ui` chart with `helm` without the operator, over an existing `aif-ui-config` ConfigMap, requires the `--take-ownership` option, available in Helm 3.15 and later #218 (https://github.com/SUSE/aif/pull/218).
- **Reserved custom repository names**: Custom repositories cannot use the names of the built-in SUSE AI Factory repositories, the default Rancher Prime repositories, the NVIDIA team repositories, or a name already used by another `ClusterRepo` resource #244 (https://github.com/SUSE/aif/pull/244).
- `Partner` **blueprint source**: SUSE AI Factory Operator 2.2.0 rejects blueprints with the `Partner` source. Do not use it until 2.3.0 is the minimum operator version in your environment #292 (https://github.com/SUSE/aif/pull/292).

3.4 Known issues and limitations

- Air-gap support covers private chart repositories, private Git repositories, repository credentials and CA trust. Preparing the cluster remains a separate task: node and containerd registry mirrors, DNS, host proxies and operating system repositories are not configured by SUSE AI Factory.
- Container image references inside blueprint values are not rewritten. Your containerd registry mirrors must resolve them transparently.
- Git over SSH is not supported for blueprint input and GitOps output. Use HTTPS with credentials and, where required, a private CA certificate.
- Blueprint partner icons specified as `https://` URLs do not load in air-gapped installations. Use base64-encoded `data:` URIs instead.
- A custom HTTPS repository that uses a certificate issued by a private CA is not trusted through the CA bundle alone. As a workaround, enable `Skip TLS verification` for that repository.
- Anonymous connection tests against OCI registries that require repository-scoped tokens, such as GitHub Container Registry, report an authentication failure even though the charts can be pulled anonymously.

- Chart access is only checked for `https://` Helm and `oci://` repositories. A plain `http://` Helm repository can pass the connection test but reports a configuration error for chart access.
- Running chart access checks repeatedly against repositories with very large indexes can exceed the operator memory limit and restart the operator.
- The OpenShell gateway chart requires the Agent Sandbox CRDs and controller (`agents.x-k8s.io`) on the target cluster.
- A Helm workload whose pod is stuck for a reason that is only reported as a pod event, such as `FailedMount` for a missing secret or ConfigMap, remains `Pending` instead of `Degraded`.
- Upgrade, rollback and retry apply to blueprint workloads only. A workload that has never reached a healthy, fully rolled out state has no rollback target.

Full Changelog: aif-operator-2.2.0...aif-operator-2.3.0 (https://github.com/SUSE/aif/compare/aif-operator-2.2.0...aif-operator-2.3.0)

4 Release 2.2.0

This release builds on the 2.1.x line with enhanced support for air-gapped deployments, recovery operations for blueprint workloads, an NVIDIA AI Enterprise application catalog that is generated and refreshed automatically, and a range of reliability fixes in the operator's Helm lifecycle.

4.1 What's new

4.1.1 Application catalog

- **Generated NVIDIA AI Enterprise catalog**: The NVIDIA AI Library is generated from the NVIDIA GPU Cloud (NGC) catalog and gated on NVIDIA AI Enterprise support, replacing the previously hand-maintained label refresh #223 (https://github.com/SUSE/aif/pull/223).
- **Automated catalog refresh**: A scheduled job regenerates the catalog weekly, so the `Supported` labels on NVIDIA applications stay current between releases #232 (https://github.com/SUSE/aif/pull/232).
- **NGC repository classification**: NGC repository paths are classified as organization, public, gated or excluded. Gated repositories are provisioned with the authentication secret, and unrecognized paths fall back to anonymous access #223 (https://github.com/SUSE/aif/pull/223).
- **NGC image pull secret injection**: NVIDIA charts receive the `ngc-secret` image pull secret under both supported chart value conventions, which resolves `403` errors caused by an empty default suppressing the injection #223 (https://github.com/SUSE/aif/pull/223).
- **Catalog scoped to managed repositories**: The `Apps` page lists applications only from `ClusterRepo` resources that SUSE AI Factory created. Previously, an unrelated `ClusterRepo` could add applications to the SUSE AI or NVIDIA libraries #182 (https://github.com/SUSE/aif/pull/182).

4.1.2 Air-gapped deployments

- **Private chart and Git sources**: Blueprints reference a chart repository by its identity, while the registry endpoints in **Settings** determine where that repository is hosted. The same blueprints therefore deploy unchanged against a local mirror. Private HTTPS Git is supported for both blueprint input and GitOps output #224 (https://github.com/SUSE/aif/pull/224).
- **Private CA certificate propagation**: You can select a CA certificate secret for SUSE Application Collection, SUSE Registry and NVIDIA in **Settings**. The operator propagates it to Rancher Prime and Fleet, and handles CA certificate and credential rotation #200 (https://github.com/SUSE/aif/pull/200).
- **NVIDIA blueprint repository in mirrored mode**: Both the NVIDIA application and NVIDIA blueprint repository identities are preserved when mirrored and pointed at your private OCI URL, so NVIDIA blueprints continue to resolve #201 (https://github.com/SUSE/aif/pull/201).
- **Image registry override**: A single `global.imageRegistry` value redirects every image shipped by the operator and UI extension charts #209 (https://github.com/SUSE/aif/pull/209).

4.1.3 Workload recovery

- **Upgrade, roll back and retry**: Blueprint workloads can be upgraded to a new blueprint version, rolled back, or retried after a failed deployment directly from the UI. Application workloads do not support these operations #214 (https://github.com/SUSE/aif/pull/214).
- **Verified rollback target**: A blueprint version becomes a rollback target only after it has been fully rolled out across all target clusters, so a workload cannot be rolled back to a version that never reached a healthy state #214 (https://github.com/SUSE/aif/pull/214).
- **Per-component and per-pod status**: Workloads report their status for each component and pod, including while a deployment is failing, which makes it easier to identify the component that is blocking the rollout #214 (https://github.com/SUSE/aif/pull/214).

4.1.4 Usability improvements

- **Open workloads in Rancher Prime**: Each workload row provides an *Open in Rancher* action that opens the corresponding Helm release or namespace, with a cluster selector for multi-cluster workloads #206 (https://github.com/SUSE/aif/pull/206).
- **Library filter is preserved**: Cancelling an application installation returns you to the library you were browsing, and filtered catalog views can be bookmarked and shared #215 (https://github.com/SUSE/aif/pull/215).
- **Catalog metadata for the UI extension**: The extension is listed as SUSE AI Factory with the correct metadata in the Rancher Prime `Apps` and `Marketplace` views #230 (https://github.com/SUSE/aif/pull/230).
- **Clearer confirmation dialogs**: The upgrade and rollback confirmation dialogs explain the effect of the operation #229 (https://github.com/SUSE/aif/pull/229).

4.1.5 Packaging

- **CRD lifecycle**: CRDs are installed natively on first installation, and a hook job applies schema upgrades. The new `crds.rbac.create` and `crds.serviceAccountName` values support restricted environments where the installer cannot create cluster-scoped RBAC #202 (https://github.com/SUSE/aif/pull/202).
- **Per-component Helm release names**: Blueprint components accept an optional `releaseName` value that pins the Helm release name of a component. It defaults to the chart name #202 (https://github.com/SUSE/aif/pull/202).
- **UI extension installed by the operator**: The operator installs the UI extension itself, so `aif-ui` is no longer installed as a separate Helm chart #233 (https://github.com/SUSE/aif/pull/233).

4.2 Key fixes

- Restarting the operator no longer leaves the UI extension's Helm release in a `failed` state without recovering #199 (https://github.com/SUSE/aif/pull/199).
- The UI extension no longer remains on an incorrect version after consecutive upgrades and downgrades #177 (https://github.com/SUSE/aif/pull/177).

- Deleting an application workload now uninstalls the underlying Helm chart instead of leaving the release in the `Uninstalling` state with orphaned resources #193 (https://github.com/SUSE/aif/pull/193).
- The operator no longer downloads the UI extension chart on every reconcile, which previously caused continuous registry traffic and could exhaust public registry rate limits #185 (https://github.com/SUSE/aif/pull/185).
- Attaching image pull secrets no longer removes other service account fields, such as labels, annotations, owner references and finalizers #192 (https://github.com/SUSE/aif/pull/192).
- Several user interface strings and tooltips that previously displayed raw translation keys now render correctly #211 (https://github.com/SUSE/aif/pull/211).
- Long Helm release names no longer prevent the UI extension from installing. The generated service name could exceed the Kubernetes 63-character limit and be rejected #210 (https://github.com/SUSE/aif/pull/210).

4.3 Upgrade notes

- **CRDs are retained on uninstall**: `helm uninstall` no longer removes CRDs or the `Settings` resource. Bundled blueprints and the UI extension are still removed. Schema upgrades are applied by a pre-install and pre-upgrade hook job #202 (https://github.com/SUSE/aif/pull/202).
- **Mirror the `kubectl` image for air-gapped installations**, because the CRD hook job requires it. Alternatively, set `crds.manageWithJob=false` and apply the CRDs manually with `kubectl apply --server-side -f crds/` before each upgrade #202 (https://github.com/SUSE/aif/pull/202).
- **The UI extension is no longer installed separately**: If your automation installs the `aif-ui` chart as a separate step, remove that step and the `aiExtension.enabled: false` override #233 (https://github.com/SUSE/aif/pull/233).
- **Deleting workloads outside the UI requires a Rancher Prime token**: Deleting an `AIWorkload` resource with `kubectl` or through GitOps only uninstalls the chart when a token is configured in **Settings** > **Rancher API Access**. Without a token, the deletion is held rather than orphaning the Helm release #193 (https://github.com/SUSE/aif/pull/193).
- **Manually created chart repositories are ignored**: Only `ClusterRepo` resources created by SUSE AI Factory contribute to the application catalog #182 (https://github.com/SUSE/aif/pull/182).

- **HTTPS Git authentication**: The authentication type setting is deprecated, and existing configurations continue to work. For new authenticated HTTPS configurations, select a password or personal access token. The user name is optional unless the Git server requires an account name. SSH is not supported #224 (https://github.com/SUSE/aif/pull/224).
- **Unused settings were removed**: The SUSE Application Collection API endpoint, catalog discovery and image rewrite settings, and the **Advanced** section of **Settings** were removed #224 (https://github.com/SUSE/aif/pull/224).
- **New workload validation**: Review existing `AIWorkload` resources before upgrading. The deployment strategy is now immutable, blueprint workloads require at least one target cluster, and GitOps workloads cannot combine the local cluster with downstream clusters #214 (https://github.com/SUSE/aif/pull/214).
- **Image overrides for air-gapped installations**: The `global.imageRegistry` and `global.imagePullSecrets` values of the `aif-operator` chart are now forwarded to the managed `aif-ui` chart. Existing nested overrides take precedence #209 (https://github.com/SUSE/aif/pull/209).

4.4 Known issues and limitations

- Air-gap support covers private chart repositories, private Git repositories, repository credentials and CA trust. Preparing the cluster remains a separate task: node and containerd registry mirrors, DNS, host proxies and operating system repositories are not configured by SUSE AI Factory.
- Container image references inside blueprint values are not rewritten. Your containerd registry mirrors must resolve them transparently.
- Git over SSH is not supported. Use HTTPS with credentials and, where required, a private CA certificate.
- Upgrade, rollback and retry apply to blueprint workloads only. A workload that has never reached a healthy, fully rolled out state has no rollback target.

Full Changelog: aif-operator-2.1.0...aif-operator-2.2.0 (https://github.com/SUSE/aif/compare/aif-operator-2.1.0...aif-operator-2.2.0)

5 Release 2.1.0

This release builds on the 2.0.x line with an operator-served application catalog that is now the default, git-backed Rancher Prime catalog support, expanded NVIDIA integration, hardened extension chart delivery, and a range of workload/UI reliability fixes.

5.1 What's new

5.1.1 Application catalog

- **Operator-served catalog, now the default**: The operator serves the application catalog through a static catalog configuration and a CRD lifecycle job, and the UI now loads applications from this bundled catalog by default (`useStaticCatalog=true`). Runtime discovery from live chart repositories remains available by setting the flag to `false` #145 (https://github.com/SUSE/aif/pull/145), #172 (https://github.com/SUSE/aif/pull/172).
- **NVIDIA AI Enterprise library**: The NVIDIA AI Library is populated with NVIDIA AI Enterprise-supported Helm charts, and the default catalog ships the org-level NVIDIA repositories out of the box #155 (https://github.com/SUSE/aif/pull/155), #172 (https://github.com/SUSE/aif/pull/172).
- **Application sorting and support labels**: The `Apps` page sorts supported applications first (default) or by name, shows NVIDIA support status through labels, and defaults the library filter to all libraries. The sort control is shared across the `Apps`, `Blueprints`, and `Workloads` screens #155 (https://github.com/SUSE/aif/pull/155), #172 (https://github.com/SUSE/aif/pull/172).
- **NGC team repositories**: The operator supports NVIDIA NGC team repositories with gated authentication. The default catalog currently ships only the org-level NVIDIA repositories, so team-repository provisioning is dormant until the catalog expands #163 (https://github.com/SUSE/aif/pull/163).

5.1.2 Git-backed Rancher Prime catalogs and Rancher Prime API access

- **Git-backed `ClusterRepo` support**: Blueprint components and application workloads can now be deployed from git-backed Rancher Prime `ClusterRepo` resources (`spec.gitRepo`/`spec.gitBranch`), including the built-in `rancher-charts` and `rancher-ui-plugins` repositories, which previously failed. The operator fetches the chart from the Rancher Prime catalog API on the management cluster and ships it as an embedded Fleet `Bundle`, so air-gapped downstream clusters deploy without each cluster needing to reach Rancher Prime or hold an API token. Identical reconciles are skipped through a chart fingerprint #156 (https://github.com/SUSE/aif/pull/156).

- **{*ranchera}** API access settings*: A new **Settings > Rancher API Access** section manages the Rancher Prime catalog credential end to end. Selecting *Authorize* mints a Rancher Prime API token as the signed-in user, *Test* validates a token before saving, and the section shows the granted expiry and warns before it lapses. When the token expires, select *Authorize* again to renew it. The Rancher Prime CA certificate is discovered automatically when none is configured, so TLS verification stays enabled by default. The configuration is hot-reloaded without an operator restart #156 (https://github.com/SUSE/aif/pull/156).

- **Oversized git charts fail fast**: Charts whose unpacked Fleet `Bundle` would exceed the Kubernetes object-size limit fail terminally with `ChartTooLarge` and guidance to host them in an OCI or HTTP `ClusterRepo` instead, rather than retrying indefinitely #156 (https://github.com/SUSE/aif/pull/156).

5.1.3 Extension delivery and security

- **Authenticated and TLS registries**: `InstallAIExtension` can pull the UI extension chart from authenticated and TLS-enabled registries. The `aif-operator` Helm chart exposes the corresponding Helm authentication and TLS settings for the bundled extension #162 (https://github.com/SUSE/aif/pull/162).

- **Registry hardening**: A registry-host allowlist guards against confused-deputy requests (with a start-up warning when the allowlist is empty), insecure registry TLS is gated behind two explicit opt-ins, plain-HTTP OCI chart pulls are no longer performed, and `https+tls` chart URLs that are not a direct `.tgz` are rejected. Terminal reconcile failures now correctly set `Ready=False` #162 (https://github.com/SUSE/aif/pull/162).

5.1.4 Workload and blueprint management

- **Pre-flight credential checks**: Blueprint credentials are validated before installation, so missing credentials are reported up front instead of surfacing as a failed deployment #148 (https://github.com/SUSE/aif/pull/148).
- **Clearer failure reporting**: Workload failure reasons are shown on the `AI Workloads` list, and `ClusterRepo` resolution failures are surfaced on the `AIWorkload` status #148 (https://github.com/SUSE/aif/pull/148).
- **Protected bundled blueprints**: Deletion of bundled blueprints is rejected by the API, and the delete action is hidden for them in the UI #122 (https://github.com/SUSE/aif/pull/122).
- **Updated blueprint**: The `simple-chatbot-with-rag` blueprint is updated to version 1.0.2. The default model changes from `gemma:2b` to `qwen2.5:3b` (larger context window); `qwen2.5:3b` is provided for research use only under the Qwen Research License #154 (https://github.com/SUSE/aif/pull/154).

5.2 Key fixes

- Rancher Prime and operator index caches are refreshed when the extension version changes, self-healing a stale index cache after an extension upgrade #152 (https://github.com/SUSE/aif/pull/152).
- `InstallAIExtension` status is patched to avoid `resourceVersion` conflicts #140 (https://github.com/SUSE/aif/pull/140).
- The family label is preserved when saving an edited blueprint as a new version #139 (https://github.com/SUSE/aif/pull/139).
- Pull-secret namespaces are no longer created on the management cluster for downstream-only workloads #130 (https://github.com/SUSE/aif/pull/130).
- The UI clarifies that the blueprint installation namespace is only a default #141 (https://github.com/SUSE/aif/pull/141).
- Deployment type cards are aligned to equal height #129 (https://github.com/SUSE/aif/pull/129).

5.3 Upgrade notes

- **CRD lifecycle on upgrade**: The operator chart applies and upgrades its custom resource definitions (CRDs) through a pre-upgrade job (`crds.manageWithJob=true`), which only creates or updates CRDs and never deletes them. In restricted environments where the job cannot run, set `crds.manageWithJob=false` and apply the CRDs from the chart's `crds/` directory manually before upgrading #145 (https://github.com/SUSE/aif/pull/145).
- **Git-backed catalogs require a Rancher Prime token**: To deploy from git-backed `ClusterRepo` resources, authorize a Rancher Prime API token under **Settings** > **Rancher API Access** after upgrading. Charts too large to ship as a Fleet `Bundle` fail with `ChartTooLarge`; host those in an OCI or HTTP `ClusterRepo` instead #156 (https://github.com/SUSE/aif/pull/156).
- **Chatbot default model change**: After upgrading `simple-chatbot-with-rag` to 1.0.2, new chats default to `qwen2.5:3b` (`gemma:2b` is removed); existing chats remain readable #154 (https://github.com/SUSE/aif/pull/154).

Full Changelog: aif-operator-2.0.1...aif-operator-2.1.0 (https://github.com/SUSE/aif/compare/aif-operator-2.0.1...aif-operator-2.1.0)

6 Release 2.0.1

This is the 2.0.1 release of SUSE AI Factory, bringing comprehensive AI workload management capabilities to Rancher Prime.

6.1 Key improvements and fixes

- Fixed pull secrets not being delivered to certain downstream clusters, which left workloads stuck in ImagePullBackOff. The pull-secret bundle now targets the downstream cluster by the `management.cattle.io/cluster-name` label instead of clusterName #107 (https://github.com/SUSE/aif/pull/107).
- Published `simple-chatbot-with-rag` and `simple-chatbot-with-rag-vllm` 1.0.1 with updated image revisions (#132 (https://github.com/SUSE/aif/pull/132)):
  - `open-webui`: 0.6.41-14.20 to 0.6.41-15.2
  - `vllm-openai`: 0.19.0 to 0.19.0-5.29
- Reduced resource requests for the Phi-3-mini-4k-instruct model.
- Marked the 1.0.0 simple-chatbot-with-rag, simple-chatbot-with-rag-vllm, and suse-inference-endpoint blueprints as deprecated as their 1.0.0 have blocking issues.

Full Changelog: aif-operator-2.0.0...aif-operator-2.0.1 (https://github.com/SUSE/aif/compare/aif-operator-2.0.0...aif-operator-2.0.1)

7 Release 2.0.0

This is the **first official release** of SUSE AI Factory, bringing comprehensive AI workload management capabilities to Rancher Prime.

SUSE AI Factory is a Rancher Prime UI Extension for managing AI workloads and applications across Kubernetes clusters. This extension provides a unified interface for installing, managing and monitoring AI applications in Rancher Prime-managed clusters.

Note

This extension requires an active SUSE AI Factory (https://www.suse.com/products/ai/) subscription to access the application catalog.

7.1 What's new

7.1.1 Core features

7.1.1.1 AI workload management

- **Multi-cluster AI workload deployment**: Deploy and manage AI applications across multiple Rancher Prime-managed Kubernetes clusters from a single interface
- **Blueprint-based application templates**: Preconfigured application blueprints for common AI use cases:
  - Simple Chatbot with RAG (Retrieval-Augmented Generation)
  - Simple Chatbot with RAG using vLLM
  - NVIDIA RAG Minimal deployment
  - SUSE Inference Endpoint

- **Declarative workload definition**: Kubernetes-native Custom Resource Definitions (CRDs) for AI workloads, blueprints, and settings
- **GitOps integration**: Fleet-based GitOps workflow support for managing AI workload configurations

7.1.1.2 User interface

- **Unified dashboard**: Overview page with at-a-glance statistics and recent activity
- **Application catalog**: Browse and install AI applications from SUSE AI Factory and NVIDIA AI Library
- **Multi-step installation wizard**: Guided workflows for deploying applications and blueprints
- **Application instance management**: Monitor and manage deployed AI workloads and their metrics
- **Blueprint management**: Create, edit, and manage custom blueprints
- **Settings management**: Helm and Git repositories configuration

7.1.1.3 Operator and backend

- **SUSE AI operator**: Kubernetes operator managing the lifecycle of AI extensions and workloads
- **Custom resource definitions**:
  - `AIWorkload`: Defines AI application instances
  - `Blueprint`: Defines reusable application templates
  - `InstallAIExtension`: Manages UI extension installation
  - `Settings`: Centralized configuration management
- **Multiple deployment strategies**: Helm for simple deployments to local clusters, Fleet integration for GitOps for downstream and multi-cluster deployments
- **Multi-vendor support**: Seamless integration with both SUSE and NVIDIA AI components
- **Credential management**: Automatic injection of registry credentials and pull secrets

7.1.1.4 Integration and connectivity

- **SUSE application collection integration**: Access to curated AI application catalog
- **NVIDIA NGC integration**: Direct access to NVIDIA AI applications and models
- **Fleet integration**: GitOps-based deployment workflows
- **Multi-cluster support**: Deploy to local and downstream Rancher Prime clusters

7.2 Known issues and behaviors

7.2.1 Installation and deployment

- Fleet must be installed before deploying the AIF operator, the installation of the AIF operator may fail.
- Standalone UI extension installation through Rancher Prime Extensions interface is not supported. Use the AIF operator default values to install it.
- Images are distributed via GitHub Container Registry (ghcr.io). In the following releases, images will be distributed via the SUSE Registry.

7.2.2 NVIDIA blueprints

- **RAG v2.5.1+**: rag-server image version 2.5.1 does not exist in the registry. See #618 (https://github.com/NVIDIA-AI-Blueprints/rag/issues/618).
- **RAG v2.5.0+**: nv-ingest deployment renders duplicate `OTEL_EXPORTER_OTLP_ENDPOINT` environment variable, causing Fleet/ArgoCD deployments to fail. See #687 (https://github.com/NVIDIA-AI-Blueprints/rag/issues/687).
- **RAG v2.6.0**: NIM embedding and reranking charts do not pass `model.profiles` to NIMCache, causing all profiles (~11) to download instead of the selected one. See #681 (https://github.com/NVIDIA-AI-Blueprints/rag/issues/681).
- **RAG v2.5.1+**: Service names use hardcoded `nvidia-blueprint-rag` prefix and ignore `nameOverride`/`fullnameOverride` values. See #662 (https://github.com/NVIDIA-AI-Blueprints/rag/issues/662).

- **RAG v2.6.0**: **Critical** - NIMs cannot be deployed in air-gapped environments; chart missing `spec.storage.pvc.name` field to use existing PVC. See #693 (https://github.com/NVIDIA-AI-Blueprints/rag/issues/693).
- **AIQ v2.1.0**: Chart hardcodes `ns-aiq` namespace, ignoring release namespace; incompatible with Fleet/ArgoCD. See #290 (https://github.com/NVIDIA-AI-Blueprints/aiq/issues/290).
- **AIQ v2.1.0+**: Local inference configuration requires manual ConfigMap creation; not documented or exposed via Helm values. See #302 (https://github.com/NVIDIA-AI-Blueprints/aiq/issues/302).

7.2.3 NVIDIA AI Library

- Not all NVIDIA Helm charts in the catalog are NVIDIA AI Enterprise supported; verify support status at the NGC Helm catalog (https://catalog.ngc.nvidia.com/search?resourceType=helm_chart&label=NVIDIA+AI+Enterprise+Supported).

7.2.4 Blueprint customization

- Bundled blueprints cannot be modified directly during the deployment. Users must create a copy to customize configuration or component versions.

7.2.5 Deprecations

SUSE AI Deployer and SUSE AI Lifecycle Manager are now superseded by SUSE AI Factory. Update as soon as possible.