



AiX AI Model Management

v2026.3

DDX Business Solutions Limited

Customer Reference

Table of Contents

1. AI Model Management	3
1.1 Overview	3
1.2 Providers & Runtimes	3
1.3 Local Models	3
1.4 External Models	4
1.5 Model Configuration	4
1.6 Assignment & Routing	4
1.7 Testing & Validation	4
1.8 Upgrade & Replacement	5
1.9 Troubleshooting	5

1. AI Model Management

1.1 Overview

■# AI Model Management

AiX supports local and approved external AI models. This section documents the product-level model lifecycle, configuration principles, validation, operations, and troubleshooting.

Mermaid diagram source

```
flowchart LR
    Provider[Model Provider / Runtime] --> Register[Register Model]
    Register --> Configure[Configure Capabilities]
    Configure --> Test[Test]
    Test --> Enable[Enable]
    Enable --> Route[Assign / Route]
    Route --> Monitor[Monitor]
    Monitor --> Upgrade[Upgrade / Replace]
    Upgrade --> Disable[Disable / Remove]
```

Document only currently supported functionality as supported. Clearly identify roadmap or planned capabilities.

Download Documentation

Prefer an offline or printable copy of this section?

Download AI Model Management (PDF)

The PDF contains the complete AI Model Management documentation for this release and is generated from the same documentation source as this website.

1.2 Providers & Runtimes

Document the AI model providers and runtimes supported by the current AiX release.

Type	Provider / Runtime	Deployment	Status	Notes	--- --- --- --- ---	Local	Ollama	Customer environment
Supported / TBD		Local	vLLM	Customer environment	Supported / TBD		Local	vLLM-Ascend
Ascend environment	Supported / TBD		External	Approved API provider	External	Supported / TBD		

Customer-specific endpoints, enabled models and credentials belong in the customer documentation. Never store API keys or tokens here.

1.3 Local Models

Purpose

Local models allow AiX deployments to execute approved AI workloads within customer-controlled infrastructure.

Record for Each Supported Model

- Model name and version

- Runtime
- Quantization, where applicable
- Context size
- Capabilities
- Minimum/recommended CPU, memory and accelerator requirements
- Installation or pull procedure
- AiX registration/configuration
- Health check
- Upgrade and rollback procedure
- Removal procedure

Lifecycle

Mermaid diagram source

```

flowchart TD
    Install[Install Runtime] --> Obtain[Obtain Approved Model]
    Obtain --> Register[Register with AiX]
    Register --> Validate[Validate]
    Validate --> Enable[Enable for Use]
    Enable --> Monitor[Monitor]
    Monitor --> Replace[Upgrade / Replace]
    Replace --> Retire[Retire Old Model]

```

1.4 External Models

Document supported external model/API integrations, required connectivity, supported authentication method, data-routing considerations, and validation procedures.

Do not include customer API keys, secrets or tokens. Customer-specific provider selections belong in `aix-docs-<customer>`.

1.5 Model Configuration

Document AiX model registration and configuration fields, including model identifier, endpoint/runtime, capabilities, context limits, defaults, timeout behavior, and any supported generation settings.

Where configuration differs between local and external models, document the differences explicitly.

1.6 Assignment & Routing

Document how AiX selects or assigns models to supported modules and workloads. Include default-model behavior, per-step overrides, local/external selection rules, and any location/departments routing supported by the current release.

Avoid describing planned routing features as currently available.

1.7 Testing & Validation

Before enabling a model for production use, validate:

- Connectivity and health
- Authentication, where applicable
- Expected model identity/version
- Context/token limits
- Representative prompts
- Response latency
- Structured output behavior where used
- RAG/tool/function behavior where applicable
- Failure and timeout handling

Record customer-specific acceptance evidence in the customer documentation or approved evidence repository.

1.8 Upgrade & Replacement

Treat model changes as controlled changes. Record the old and new model/runtime versions, compatibility considerations, validation performed, rollback approach, and release/customer approval where required.

Prefer validating replacement models in UAT before enabling them in Production.

1.9 Troubleshooting

Check, in order:

1. Runtime/service health.
2. Network/DNS connectivity.
3. Model availability and identifier.
4. AiX endpoint/model configuration.
5. Authentication for external providers.
6. CPU/GPU/NPU memory and capacity.
7. Timeout/context-limit errors.
8. AiX and runtime logs.

Never paste secrets into troubleshooting records.