Red Hat AI

What's New, What's Next

Q3 2026 edition

Red Hat AI Product Management



Version nur

Red Hat

Red Hat AI - What's New and What's Next Series

Stay informed about the latest developments in Red Hat AI

- **For Q/A and discussions**
  - **Customers and partners**: YouTube Chat
  - **Red Hatters**: use the slack channel #forum-ai-field-feedback
- **Recording and slides:**
  - redhat.com/whats-new-red-hat-ai
- **Red Hat AI Roadmap sessions are hosted Quarterly**

Any forward looking statements are **subject to change**

Generative AI customer adoption challenges



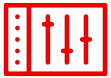
Cost

Generative AI frontier model services are cost prohibitive at scale for most enterprise customer use cases.



Complexity

Turning enterprise data into reliable, governed, production-ready AI is too complex.



Control

Increasing concerns with data privacy, security, and latency are compelling organizations to adopt hybrid strategies.

Red Hat AI

Inference

Fast, flexible and efficient inferencing

Data

Connecting data to models and agents

Agents

Scale AI agents with governance and control

Platform

Governance and security from metal to agents

Run any model or agent on any infrastructure!

Red Hat AI

The Red Hat AI family of products

Red Hat AI Enterprise
For enterprise AI, deploy and scale efficiently, anywhere

Red Hat AI Inference
For optimized inference of large language models

Red Hat OpenShift AI
For the distributed deployments in OpenShift

Red Hat Enterprise Linux AI
For inference in individual Linux server environments

Trusted, Consistent and Comprehensive foundation

NVIDIA AMD intel Hardware Acceleration Google aws IBM

Physical Virtual Private Cloud Sovereign Cloud Public Cloud Edge

Red Hat

What's New In RHAI 3.5 - Highlights

Safety and Observability at the Core of Enterprise AI

Flexible and Efficient Inference

- MaaS external models, OIDC support and showback (TP)
- Controlled deployments of llm-d with automatic rollback
- Flow Control – priority-aware serving on shared GPU pools (GA)
- Llm-d on third-party Kubernetes - EKS, AKS, CKS
- vLLM Omni (DP)

Connecting Agent & Models to Data

- AutoRAG (TP): multilingual
- AutoML (TP) with built-in serving runtimes
- Post-training with GRPO/RLVR (GA)
- Distributed fine-tuning with Ray



Agentic AI

- Responses and RAG APIs (GA)
- MCP Management including MCP G/W
- Agent Management
- OpenShell sandboxing (DP) for secure execution
- GenAI Studio enhancements

AI Platform

- Eval Hub (GA) - Server, SDK, CLI, MCP Server
- NeMo Guardrails as the default guardrail service
- AI Safety And Governance Orchestration
- User/tenant Observability(TP)
- GPU-as-a-Service dashboard
- Multi-tenancy with OpenShift Virt and Hosted Control Planes (GA)

Run any model or agent on any infrastructure!

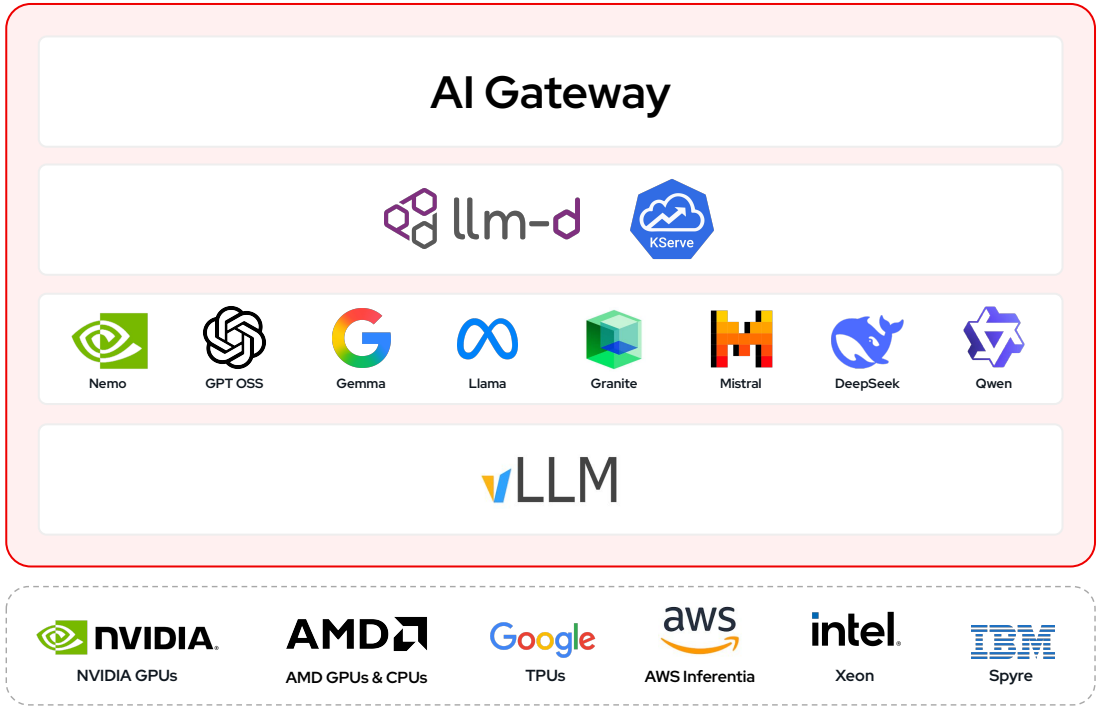
Red Hat

Fast, flexible and efficient inference

Red Hat

Enterprise GenAI Inference Platform

Operationalize Model as a Service across the hybrid cloud



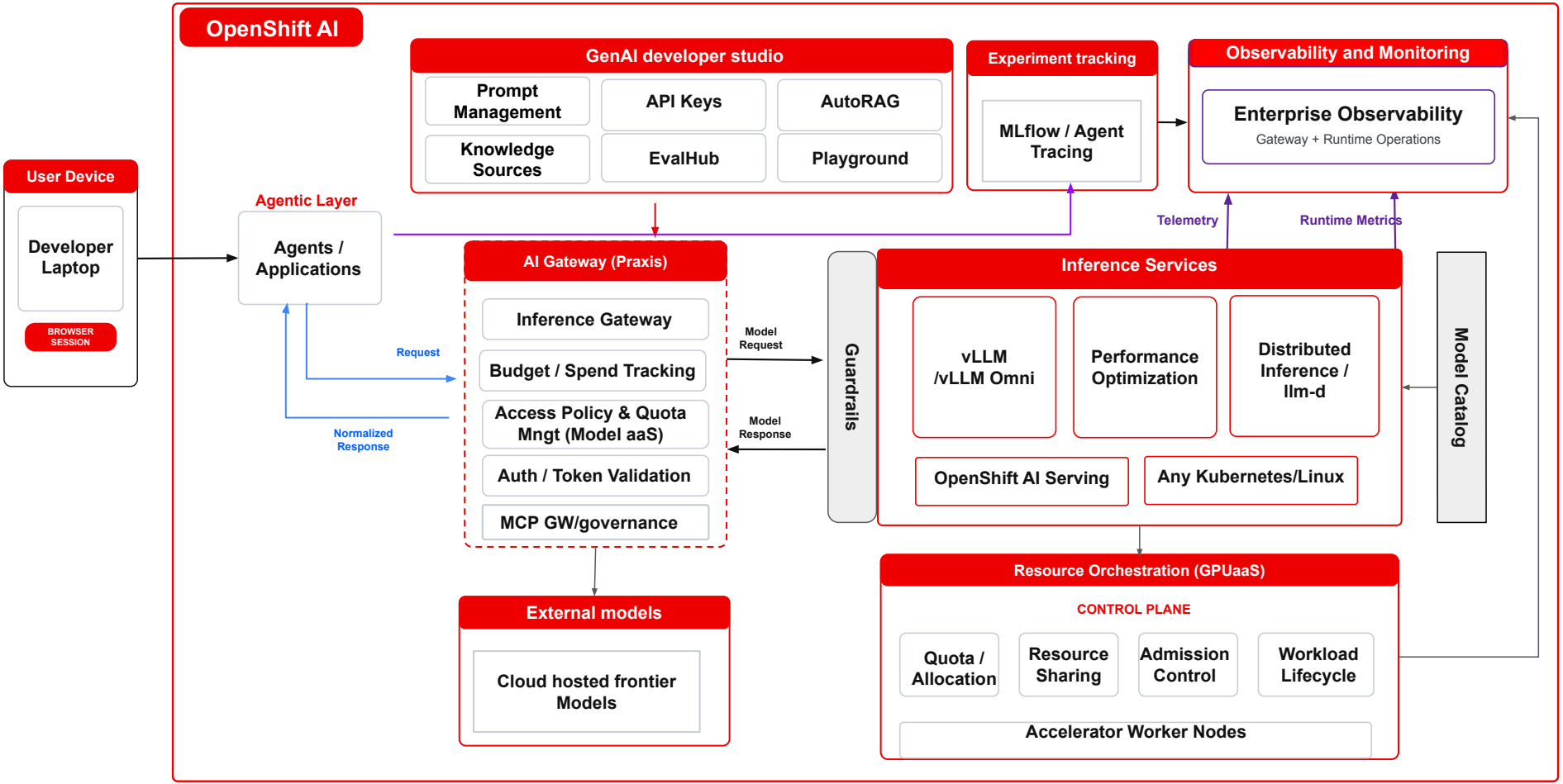
API Management, Token Tracking, Authentication for any model

Distributed Inference

Validated, Optimized Models & Techniques

Inference Server

Enterprise Agentic Inference Platform on Openshift AI



vLLM: The Core Engine for Inferencing

What is it?

- SOTA model coverage (esp. open-weight)
- Hardware portability across accelerators
- Performance is the product for tokenomics
- Production-grade disaggregated serving
- Multi-modality: Omni any-to-any modality (text, vision, audio and video)

Why it matters

- Core engine of your inference factory to set P&L tokenomics
- Quantization, Speculative Decoding, KV offload and disaggregation save orders of magnitude $/million tokens on the right workload

What's Next (RHAI 3.6 Fast 1, vLLM 0.26.0 and 3.6 GA)

- Nvidia Vera Rubin (3.6), vLLM granular cache accounting (3.6 Fast 1)
- Omni TP with real time voice intelligence (3.6 Fast 1)
- vLLM Watermarking for AI-generated context (3.6)

What's New (Q3 2026, RHAI 3.5 vLLM 0.24.0)

- **Day 0 models + Optimizations**. GLM5.2/5.3, DeepSeek V4, Kimi K3, Inkling, Qwen3.8. Optimized with llm-d for enterprise workload including agentic coding blog
- **Beyond autoregressive**. DiffusionGemma for text. Diffusion/DiT image and video through Omni.
- **Faster Tokens.** NVFP4 KV cache/fused MoE, MXFP4, TurboQuant, Speculative decoding with MTP, DFlash, DSpark.
- **Resilience & Scalability:**
  - Hardened Prefill-Decode Disaggregation with request draining for scaling; DP-aware load balancing
  - Hardened DeepEP and elastic-EP for large open weight models.

Red Hat

Flexible Interface For Chained & Omni Multimodality

Optimized Enterprise Voice AI:

Omni Tech Preview (3.6 Fast 1 Release)

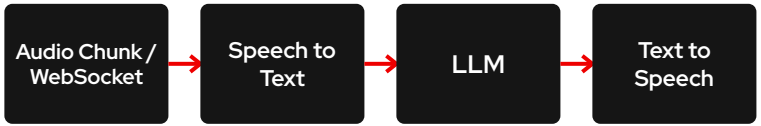
Latency Real time (~200ms E2E latency) conversations

Intelligence Detect language, tool call and reason across modalities

Naturalness Understand interruptions and preserve emotions/tones

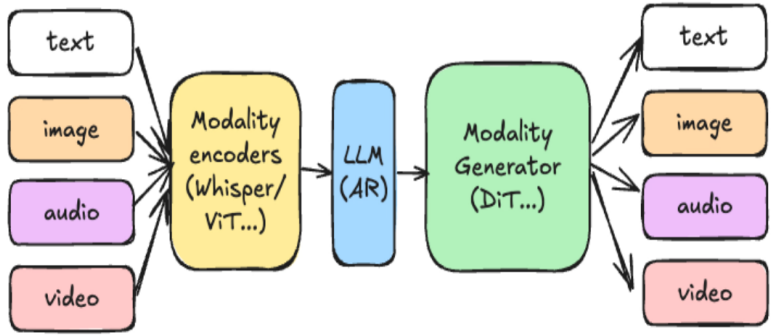
Streaming Native & Batch: WebSocket and long audio chunking

Chained Approach: Voice AI Pipeline Example



Every hop costs latency. Tone and hesitation are lost at ASR.

Omni Approach: One Container, Any Modality



Faster delivery of vLLM

What is it?

- Delivery of additional downstream supported versions through fast RHAII vLLM releases

Why it matters?

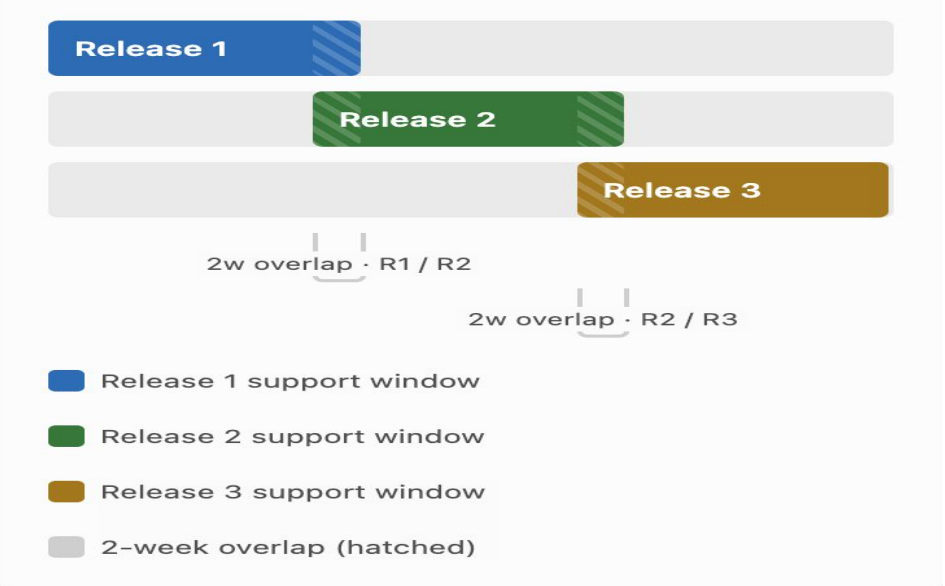
- Allows updates to vLLM without waiting for the next GA release
- Supported by Red Hat to allow customers the confidence to deploy in production

How it works

- Delivered by z stream releases (i.e. 3.5.z)
- Each vLLM will be supported for an additional two weeks after the release of the next downstream release of vLLM

What's Next (RHAI 3.6, 1H 2027)

- Model Runtime Library to deliver vLLM outside of z stream releases.



Red Hat

Models: Validation and Optimizations

What is it?

- 3rd party model validation, packaging and delivery into model catalog
- Optimized models with quantization
- Speculative Decoding - Inference-time optimization

Why it matters?

- Time to value for model deployment into production while managing risk
- Reduce GPU needs/costs and increase performance

What's Next (RHAI 3.6, 1H 2027)

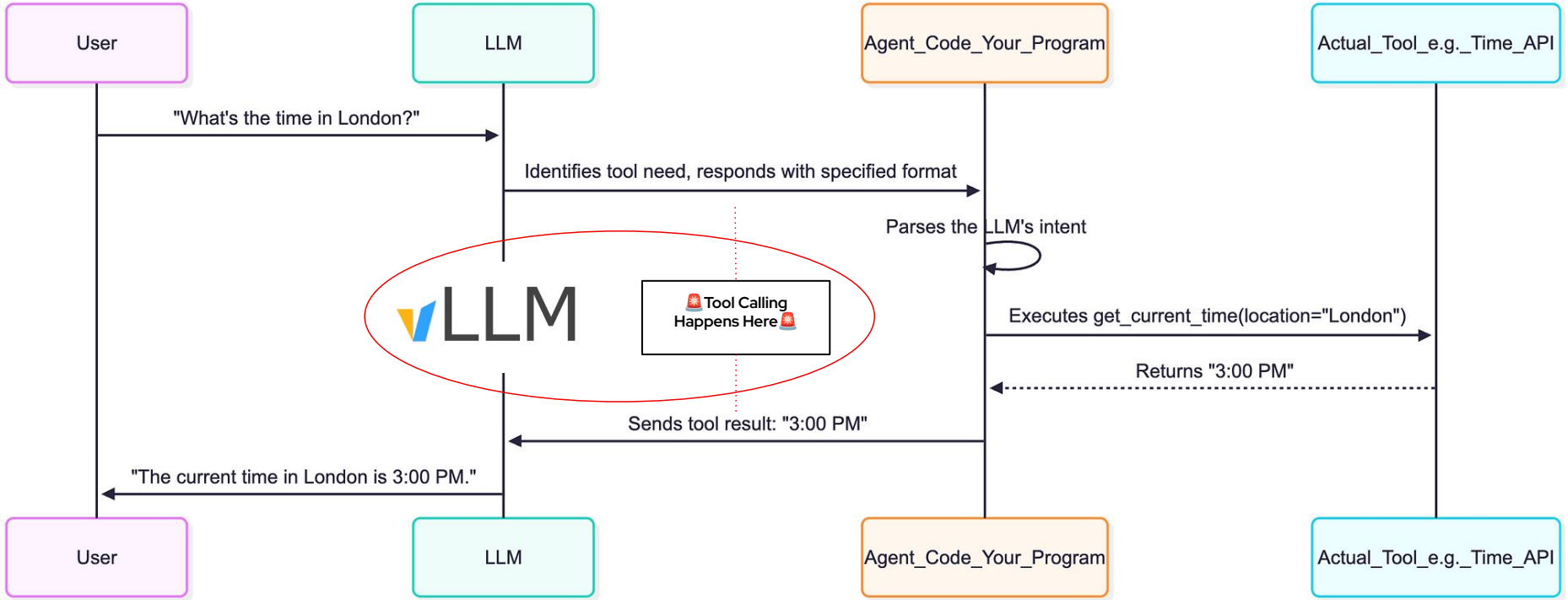
- 3.6 GA model validation batch (in-planning):
  - Muse Glimmer
  - Inkling
  - Qwen 3.8
- Evaluations integration for validated models
- Static security scan integration for validated models

What's New (Q3 2026, RHAI 3.5)

- Security scans now provided with validated models:
  - Garak red-teaming results now provided with validated models to provide customers with model risk scores
- New batch of validated models in 3.5 GA RHOAI catalog including:
  - Gemma 4
  - Nemotron Ultra
  - Qwen 3.6
- New speculators integrated into model catalog for native speculative decoding serving from RHOAI
- LLM Compressor v0.12.0
  - Transformers v5 upgrade: unified load_context() MoE loading, simplified dataset API, drops transformers<5.0.0.
  - Multi-GPU model-free PTQ + Nemotron-3-Ultra-550B FP8 examples.
- Speculators v0.6.0
  - New algorithms: P-EAGLE, MTP finetuning, DFlash sliding-window attention.
  - GuideLLM benchmarking suite + multi-modal and multi-node training.

Red Hat

What is Tool Calling?



Why does tool calling matter?



MCP Explosion

Model Context Protocol clients & servers need accurate, structured tool calling as a core primitive



AI Agents

Agentic frameworks (LangGraph, Llama Stack, CrewAI, Kagenti) drive execution loops through tool_calls



Enterprise Apps

Mandates to integrate existing business systems with AI are driving real-world tool calling at scale

The OpenAI API moved tool responsibility from clients to the inference server – a double-edged sword

Completions API

```
Manual tool prompting
Client-side response parsing
```

→

Chat Completions API

```
tools  in request
tool_calls  in response
```

→

Responses API

```
Server owns even more
Configuration matters more
```

Model Validation: Tool Calling Updates

What is it?

- Tool calling in large language models (LLMs) is a feature that lets an AI model request data or trigger actions from external systems—like databases, APIs, or code interpreters—instead of relying only on its static training data

Why it matters?

- Accesses real-time data: Fetches live facts, weather, and stock updates.
- Executes software actions: Automates tasks like booking calendars or sending emails.
- Eliminates calculation errors: Outsources complex math to precise code interpreters.
- Prevents AI hallucinations: Grounds answers in actual database and API results.
- Ensures system safety: Separates AI reasoning from direct backend execution.

What's New (Q3 2026, RHAI 3.5)

- 5 models fully validated for tool calling in RHOAI Model Catalog - verified chat-templates, tool-call-parsers, vllm configs
- 3.5 GA Models:
  - RedHatAI//gpt-oss-120b
  - RedHatAI/NVIDIA-Nemotron-3-Super-120B-A12B
  - RedHatAI/Mistral-Small-4-119B-2603
  - RedHatAI/Qwen3.6 27B
  - RedHatAI/Gemma 4-31B

What's Next (RHAI 3.6, 1H 2027)

- 3.6 GA tool calling updated validated models batch (in-planning):
  - Muse Glimmer
  - Inkling
  - Qwen 3.8
- Agentic Evaluation integration into Catalog

Red Hat

Distributed Inference with llm-d

What is it?

- Kubernetes-native orchestration that uses request-level signals to route, scale, and schedule inference workloads, built on vLLM

Why it matters?

- More predictable performance and better GPU utilization, with control over how inference runs.

What's Next (RHAI 3.6, 1H 2027)

- **Multi Cluster routing**: Reduce response time and GPU work by routing requests across clusters based on inference state and cached context
- **Fast model server startup**: Reduce cold starts during scale-up, restart, and model switching
- **llm-d orchestration for vLLM Omni**: Coordinate Multi Modality Input and Output (DP)

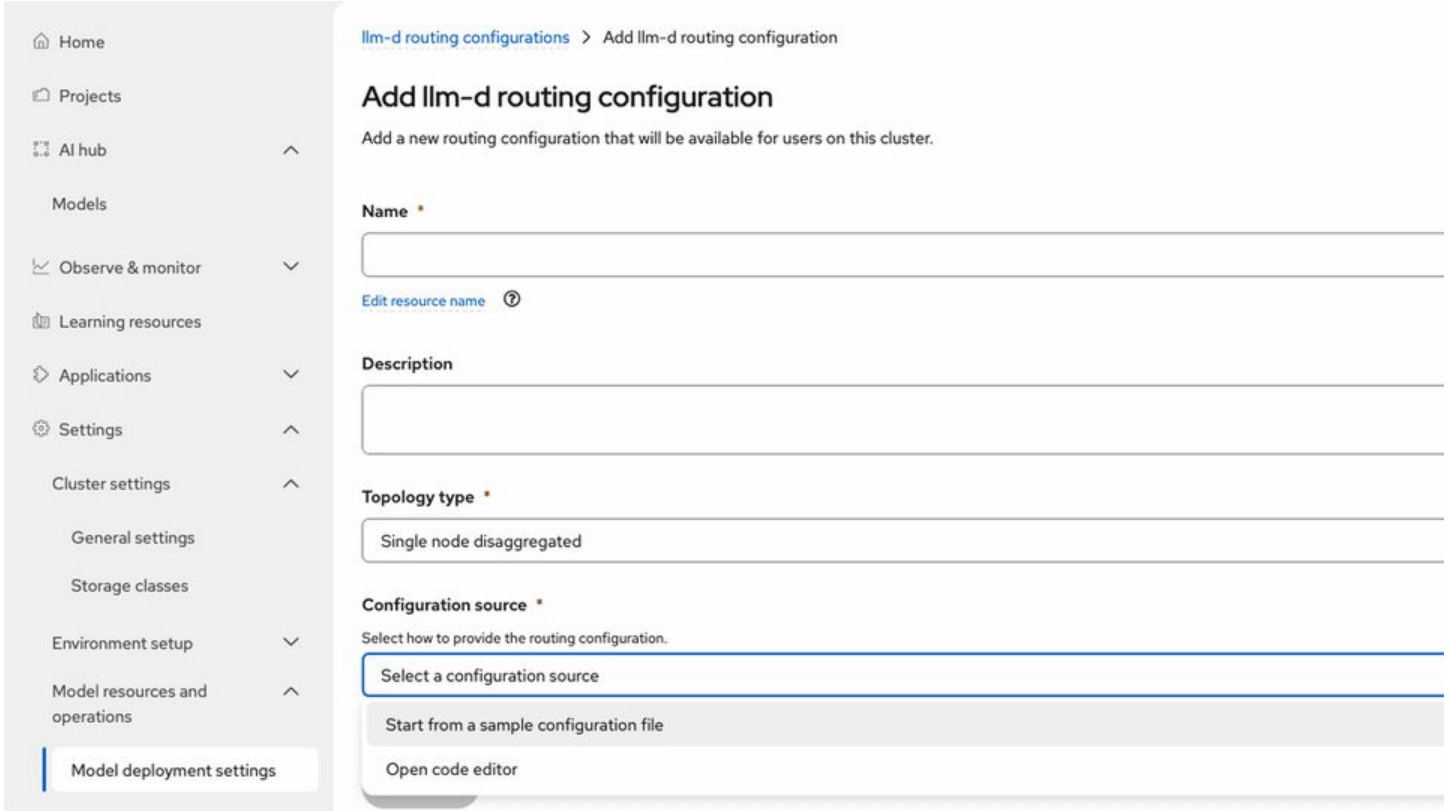
What's New (RHAI 3.5)

- **Flow control**: Prioritize critical traffic under load (GA)
- **Tiered KV-cache offload**: Reuse cached prompts across GPU, CPU (GA) and POSIX filesystem storage (DP)
- **Controlled deployments**: Test model updates gradually and roll back without disrupting traffic (GA)
- **Inference observability**: Troubleshoot serving health through Perses dashboards, Service Level Metrics and Distributed Tracing (GA)
- **Distributed Inference**: Azure AKS (GA), CoreWeave CKS (GA); Amazon EKS (TP)

Red Hat

Guided model deployment with llm-d

Platform Operators define reusable topology and routing configurations. Users select them through the form and can inspect the generated YAML.



Red Hat

Flow control: Priority Aware Serving for shared GPU Inference - llm-d

What is it?

- Priority-aware admission for mixed workloads on a shared model pool
- Central queues organized by priority and tenant

How it works

- Configure priority bands and tenant identity
- Set the saturation threshold and fairness policy
- Detects when the model pool is saturated
- At saturation, llm-d queues requests and advances higher-priority traffic first

Why it matters?

- Protects latency-sensitive traffic under load
- Prevents one tenant from consuming capacity
- Reduces idle GPU capacity reserved for traffic peaks
- Runs more workloads on the existing GPU fleet

Go deeper in our developer blog
red.ht/flow-control-developer-blog

Red Hat

Model Gateway (Model as a Service)

What is it?

- Multi-tenant governance layer for LLM serving
- Resource Governance: Token Rate Limiting & Tenant Authentication

Why it matters?

- **Enabled enterprise-grade governance:** token rate limiting, RBAC, usage tracking, and multi tenant isolation

What's New (Q3 2026, RHAI 3.5)

- Multi-provider external model support (GA)
- OpenAI-compatible body-based routing (GA)
- External OIDC support (GA)
- Multi tenancy (TP)
- User showback/ usage dashboard (TP)
- External metering and token-level cost attribution (DP)

What's Next (RHAI 3.6, 1H 2027)

- Multi-tenancy improvements
- Guardrails integration
- Embedding model support (TP)
- Consumer Portal (TP)

Red Hat

AI Gateway - Praxis Data Plane

Replacing the Envoy-based Model as a Service Gateway & More

What is it?

- **Today:** Model as a Service routes inference through an Envoy-based gateway - IPP for provider translation, Istio/Kuadrant for Auth, Rate Limiting, HttpRoutes, etc.
- **Next:** The Praxis-based AI Gateway will become the data plane for that same MaaS control plane - Subscriptions, API Keys, AITenant, & Gen AI Studio Stay; Praxis will eventually replace Envoy as the inference hop
- **One Gateway for All:** OpenAI-compatible URL for self hosted (llm-d/vLLM) and external models (Bedrock, Vertex, Anthropic, OpenAI) - policy, translation, credentials, & telemetry on one purpose-built proxy, not a chain of generic Envoy filters

Why it matters?

Envoy was the right bootstrap for MaaS early on, but AI Gateway features - body-based routing, provider API translation, credential injection, GenAI Observability, Guardrails, and MCP - need a data plane built for inference, not bolted onto a general-purpose proxy. Praxis will replace fragmented Envoy + IPP + Shim stacks with one governed hop: same RBAC & Quotas (Kuadrant/RHCL), fewer moving parts, and a path to GA without maintaining two gateway implementations.

Red Hat

AI Gateway

Control Plane that manages and governs Agent traffic

What is it?

- Centralized control plane for AI and agent traffic to models, tools, memory, guardrails and observability
- Centralized authn and RBAC (OAuth 2.0), ensures agents use scoped tokens instead of master keys
- The AI Gateway is the next generation control plane & data plane that manages and governs agent traffic

Why it matters?

Replaces fragmented model access, duplicated agent infra, and inconsistent guardrails with centralized policy enforcement, intelligent routing, shared memory, and end-to-end observability.

What's Next (RHAI 3.6, Q4 2026)

- Token Tracking, Usage, Rate Limiting, Advanced Metrics
- Content Filters, Guardrails, Runtime Safety Filters
- AI Abstraction Layer
- Content Based Routing / Model-Aware Routing

What's New (Q3 2026, RHAI 3.5)

- MCP Gateway has made it to Tech Preview w/ GA targeted in RHAI 3.6
- Enhanced Traceability Metrics for a better developer experience
- Egress Gateway Policy Support
- gRPC Policy Support

Red Hat

Agentic AI

Red Hat

Agentic APIs

What is it?

A secure, OpenAI-compatible API layer (OGX - formerly Llama Stack) standardizing agentic workflows,including orchestration and RAG, across any environment.

Why it matters?

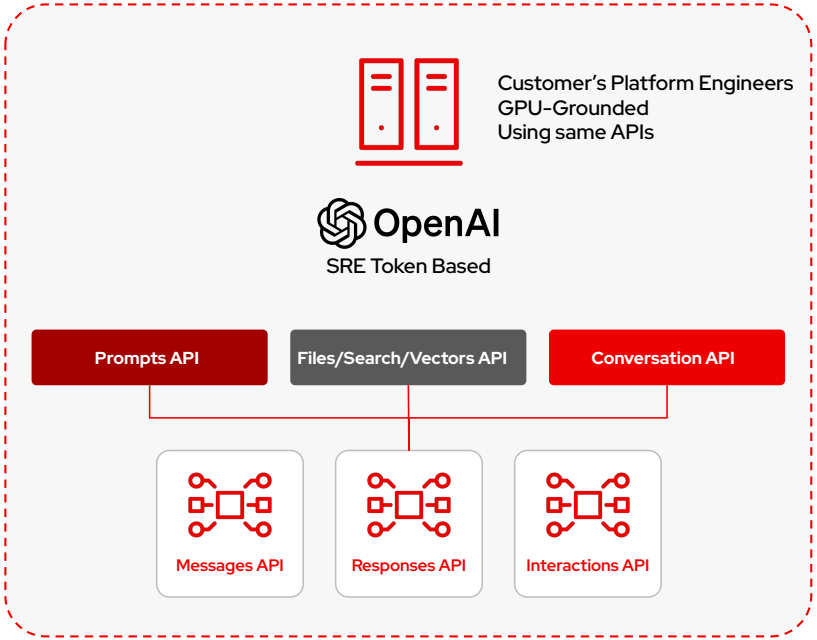
Simplifies building portable, production-ready agents by providing a consistent interface across any model, provider, or infrastructure.

What's New (Q3 2026, RHAI 3.5)

- **Responses API (GA via OGX):** agentic loop with tool use, multi-turn, streaming
- **RAG main and supplementary APIs GA:** Files API + VectorStores for end-to-end RAG + Conversations API for server-side turn history
- Multi-tenancy with OAuth2, JWKS, ABAC

What's Next (RHAI 3.6, 1H 2027)

- **A2A protocol support** for multi-agent workflows for AI Gateway
- **Messages API** in the AI Gateway
- **Coding agent qualification:** Codex, Claude Code against AI Gateway



Red Hat

MCP Management

What is it?

A platform approach for discovering, deploying, securing, and governing MCP servers

Why it matters?

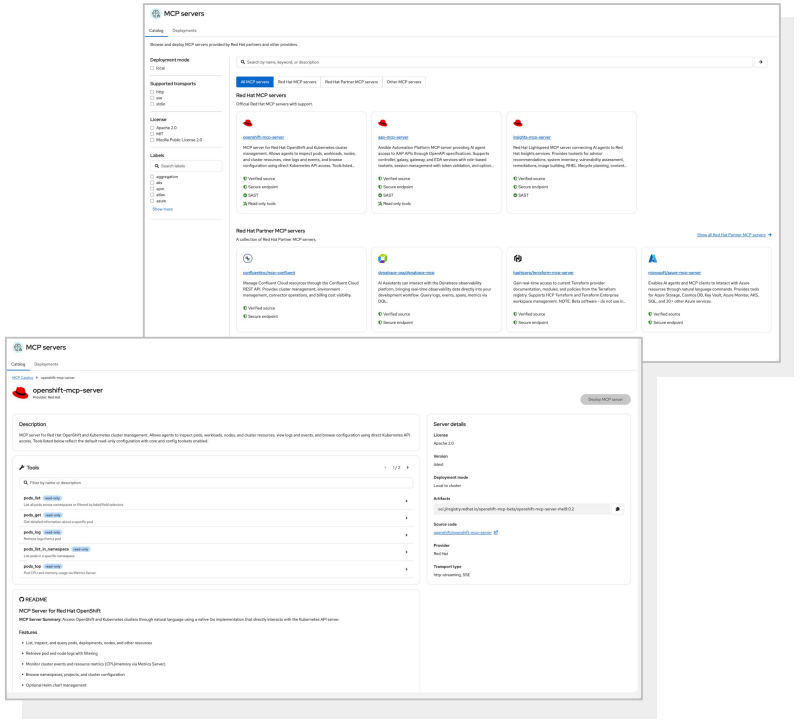
Simplifies the development of production-ready agentic apps

What's New in RHAI 3.5

- MCP Gateway (TP)
- MCP Catalog (DP)
- MCP Lifecycle (TP)
- Consumption in Gen AI Studio

What's Next for RHAI 3.6 and 1H'27?

- MCP Gateway (GA)
- MCP Catalog (GA)
- MCP Lifecycle (GA)
- MCP Registry (TP → GA)
- Registered MCP consumption in Gen AI Studio
- 1H'27: trust tiers, approvals, signed provenance (TP)



Red Hat

Agentic Skills Management

What is it?

A platform approach for discovering, governing, and versioning the skills agents use

Why it matters?

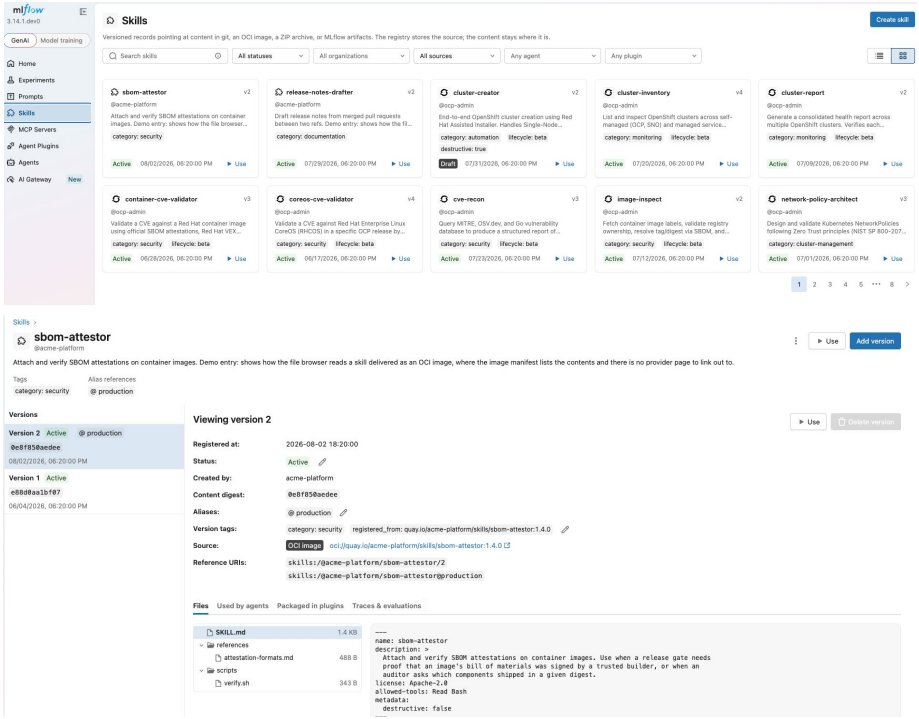
Turns Red Hat, partner, and in-house expertise into governed, reusable agent capabilities

What's New in RHAI 3.5

- Skill Registry design merged upstream (MLflow RFC-0008)
- Agent Plugins v1 adopted as the skill package format
- Agent Plugins registry RFC opened upstream (RFC-0010)
- Red Hat agentic skills collections published upstream

What's Next for RHAI 3.6 and 1H'27?

- Skills Catalog (TP)
- Skills Registry (TP)
- Curated Red Hat skills via the trusted AI asset pipeline
- 1H'27: Skills Catalog and Skills Registry (GA)



26

Red Hat

Agent Management

What is it?

A platform approach for discovering, deploying, sandboxing, and governing agents

Why it matters?

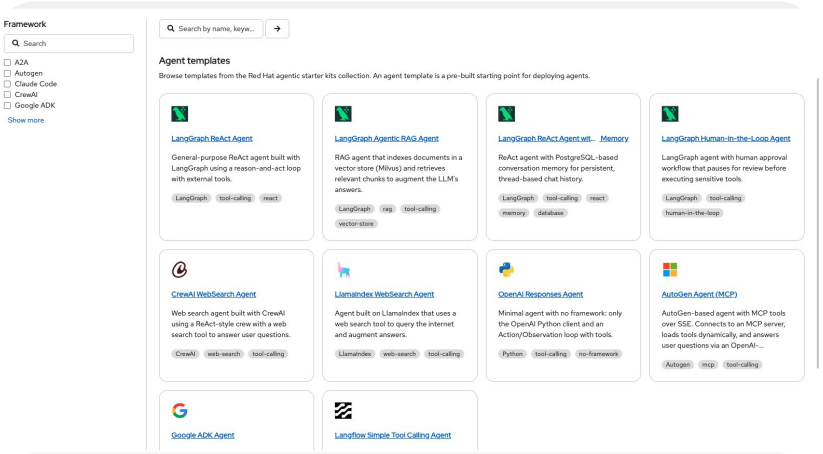
Moves agents from experiment to production with inventory, identity, and containment

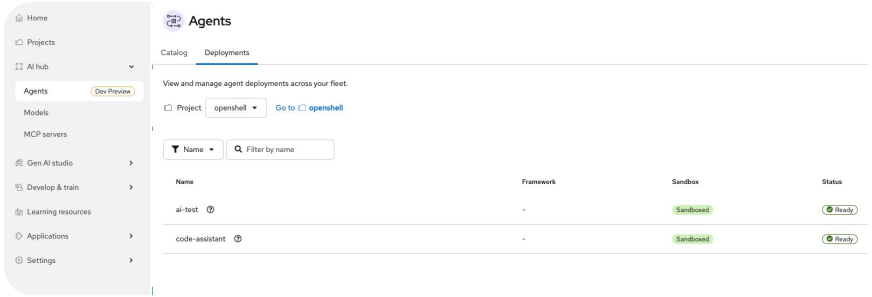
What's New in RHAI 3.5

- Agent Catalog - Red Hat starter and harness kits (DP)
- Agent deployments view (DP)
- Agent metadata schema based on the A2A Agent Card

What's Next for RHAI 3.6 and 1H'27?

- **Agent Catalog**: deploy button via OpenShell Go SDK + harness images
- **Agent Eval Starter Kit :** scorer templates, methodology
- **Build and deploy agent configuration**s from Gen AI Studio
- **1H'27:** Agent Registry (TP)





Red Hat

Gen AI Studio - AI Assets and Agent Builder Playground

What is it?

Interactive playground for AI engineers to discover, test, and prototype models, agents, and associated components

Why it matters?

Accelerates time to market by centralizing model, prompt engineering, RAG, and agent prototyping

What's New in RHAI 3.5

- Tracing in playground sessions (TP)
- Prompt Templates (TP)
- Model and configuration comparison
- Multi-modal support for image and audio (TP)
- Saved agent configurations (DP)

What's Next for RHAI 3.6?

- Gen AI Studio playground (GA)
- Deploy saved configurations as agents
- Consumes the MCP Registry
- Document uploads directly to context
- Model Prompt Pairings

What's Next for 1H'27?

- AutoRAG integration
- Skills registry consumed by playground
- Register saved agents in Agent Registry
- Standalone chat experience with feedback loop
- Agent Configuration Comparison

Red Hat

Secure Agent Onboarding + Agent Platform

What is it?

Secure agent execution with identity management, sandboxed isolation, and compliance-grade audit logging via OpenShell on OpenShift. Joint development with NVIDIA.

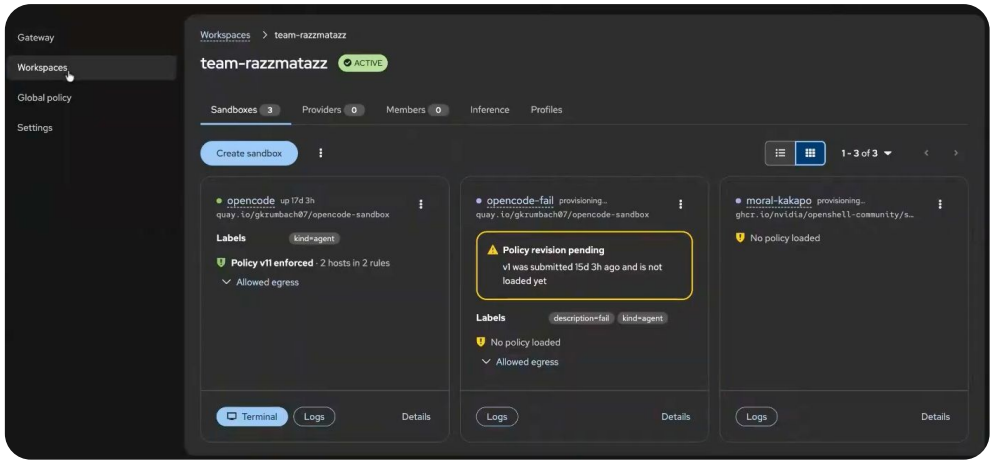
Why it matter?

Agents run with shared secrets (no identity), untrusted code runs uncontained (no isolation), and there is no trace of what agents do (no visibility). OpenShell closes all three gaps.



What's New in RHAI 3.5

- **OpenShell (DP):** secure agent sandboxing with policy enforcement, SPIFFE/SPIRE, Policy, OSCF Logs.



What's Next for RHAI 3.6 and 1H'27?

- **OpenShell with** Token exchange and delegated identity landing: agents authenticate as themselves to external services
- **Native OTEL tracing** via OTEL relay.
- **Sandbox lifecycle** (stop/start, eviction protection, graceful termination
- **OpenShell UI**

Connecting data to models and agents

Red Hat

Optimize for cost, accuracy and reduce time to production

Through enterprise context, customization, and optimization

Trusted enterprise context

- Discovery, connection, ingestion*
- Data processing
- Synthetic data generation
- Feature store

Operationalized AI expertise

- AutoML
- AutoRAG
- AutoEval, etc (Future)

Customization and optimization

- Prompt design (via GenAI Studio)
- RAG
- Post-training
- Inference-time scaling

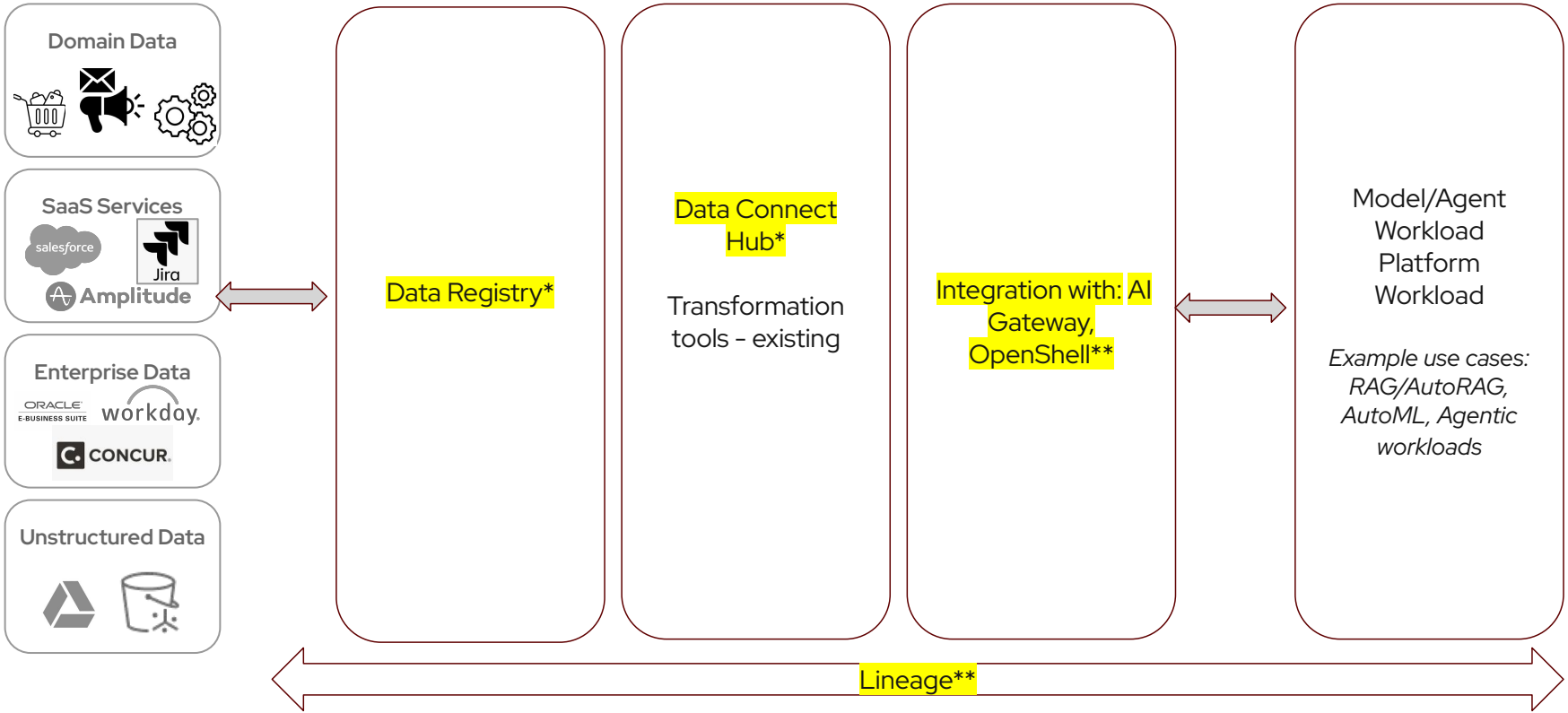
Item* - roadmap item

Red Hat

RH AI - Plans for Data Strategy

Data Producers

Data Consumers



Highlighted items denote net new engineering work.
Item* - roadmap for 3.6
Item ** - roadmap planned for 3.7

Data Processing - Transformation tools

What is it?

- Kubeflow Spark Operator integration with workbenches: Allows users to interactively submit and manage batch data processing jobs from within workbenches (DP)
- New docling image with new versions of the library (SDK and API), as well as models baked in for disconnected environments

What's Next for RHAI 3.6?

- A way to discover and use data artifacts on the RH AI platform
- Advanced credential management for users

Red Hat

AutoRAG

Automate and optimize your RAG pipeline in minutes instead of days

What is it?

- **Automated Pipelines**: Automatically test multiple RAG algorithms and techniques for your use case
- **Hyper-parameter Optimization**: Evaluates variables such as chunk size, embedding models, vector stores and top-K retrieval settings.
- **Evaluation:** Uses advanced metrics (like Answer Correctness and Faithfulness)

Why it matters?

- Reduces complexity with RAG
- Accelerate Time-to-Market
- Ease of use

What's New (Q3 2026, RHAI 3.5) - Advanced TP

- Multi-lingual support (German, Spanish, Japanese)
- Responses API Export
- Chat with AutoRAG patterns in AI Playground
- Inline Evaluation creator
- Expanded Vector database support
- Updated visual pipeline

What's Next (RHAI 3.6, 1H 2027) - GA

- AutoRAG (GA)
- Graph AutoRAG (TP)
- Multimodal (Audio) AutoRAG (TP)

Red Hat

Auto ML

AutoML for Scalable Predictive Intelligence

What is it?

- Automate **end to end development** of machine learning workflows built using open source ML libraries
- **Feature Engineering**
- **Hyperparameter Optimization/Model Selection -** Selects best model for your use case by training several models with a range of hyperparameter configurations
- **Validation & Testing:** uses Metrics such as confusion matrix, ROC curve, F1 score

Why it matters?

- Reduces complexity with RAG
- Accelerate Time-to-Market
- Ease of use

What's Next (RHAI 3.6, 1H 2027)

- Auto ML (GA)
- Integration with MLFlow
- Expose AutoML models as MCP tools for agentic workflows

What's New (Q3 2026, RHAI 3.5)

- Auto Gluon as a First-Class KServe Runtime/Out of the box Supported Runtime for predictive models in Red Hat Openshift AI
- Advanced Experiment Settings – *Preset Support (Faster vs Better Quality)* & Evaluation Metric Override

Red Hat

Post Training

Training Libraries and Distributed Training Backends

What is it?

- Enable enterprise customers to finetune task specific small language models
- Training Hub Library: offers post training algorithms such as LoRA, SFT, OSFT, GRPO, LoRA-GRPO
- Distributed Training Backends: Kube Flow Trainer, Ray Train

Why it matters?

- Sovereign AI; own your model weights instead of renting intelligence
- Small purpose built models could be more useful for agentic workflows for tasks done at repeated volume

What's Next (RHAI 3.6, 1H 2027)

- Training as a Service (DP)
- Amortized (Conversational experience for building SLMs)

What's New (Q3 2026, RHAI 3.5)

GRPO / RLVR Fine-Tuning (GA)

- Train models to reason using programmatic reward functions; no human-labeled data or separate reward model
  - Just write a function that checks if an answer is right (valid tool calls, correct SQL, well-formed JSON)
- Same workflow as SFT/LoRA - Use Training Hub + Kubeflow SDK from a Jupyter notebook
- **Universal Training Image** ships vLLM, ART backend, Training Hub & Kubeflow SDK pre-installed

Distributed Fine-Tuning with Ray (GA)

- Run SFT, OSFT, LoRA, and GRPO on Ray, the industry-standard distributed compute framework
- Submit jobs via CodeFlare SDK from a notebook; clusters scale elastically; GPUs provisioned on demand, released when done
- Ray CUDA runtime image ships Training Hub pre-installed; works in air-gapped environments

Red Hat

A tool call can fail two ways

One is a format problem, solved at serve time. The other is a judgment problem, and only training fixes it.

At serve time — vLLM

Is the call well-formed?

```
{
  "name": "refund_order",
  "arguments": {
    "order_id": "A-4417",
    "amount": 89.50
  }
}
```

- ✓ Raw output parsed into a typed call
- ✓ Shape matches the declared schema
- ✓ Your agent code gets a contract, not a string

The serving layer guarantees the format.

In the weights — Post Training

Is it the call you wanted?

```
{
  "name": "lookup_order",
  "arguments": {
    "order_id": "4417",
    "reason_code": "CX"
  }
}
```

- ✗ Wrong tool — looks up, never refunds
- ✗ Order ID missing its prefix
- ✗ reason_code isn't in your schema

Parses perfectly. Still wrong. Only training fixes it.

How GRPO / RLVR fixes it — no labeled data, no reward model

Try	→	Check	→	Compare	→	Reinforce
The model takes several attempts at the same prompt		A program grades each one: right tool, valid args, call succeeds		GRPO ranks the attempts against each other in the group		Weights shift toward the attempts that passed

↻ Repeat until the model reaches for the right tool on your schemas, every time

Training Hub ships built-in reward functions for tool-call correctness — or bring your own checker

Inference Time Scaling

More accuracy at runtime: Scaling intelligence through smarter inference instead of retraining.

What is it?

- Additional "thinking" at inference time: parallel candidate generation + reward-model verification to select the optimal answer.

Why it matters?

- Smaller models achieve frontier-level accuracy at inference time — no expensive fine-tuning required
- Ideal for high-stakes domains (finance, healthcare) where correctness matters more than latency

What's New for Q3, RHAI 3.5

- Smarter Self-Consistency: Adaptive and Beta methods let easy queries stop early while harder queries use the full budget, lowering average inference cost without sacrificing accuracy.

What's Next for RHAI 3.6 and 1H 2027

- Integration with Red Hat AI Gateway to make ITS capabilities easily accessible across the platform.

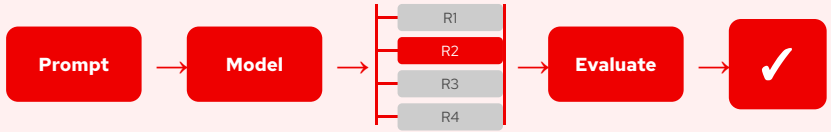
Today: single-shot inference

One attempt per question. Same compute whether trivial or hard.



ITS: multiple attempts, best answer

Generate N responses in parallel. Evaluate each one. Select the best.



Red Hat

AI Platform

Red Hat

AI Evaluations with EvalHub

What is it?

- **Comprehensive Evaluation:** Analyze AI systems, including models, agents, skills, tools, guardrails, and LLM traces
- **Assessment Capabilities:** Evaluate AI systems for suitability, scalability, capabilities, and potential risks.
- **Safety & Performance Benchmarking:** Measure models and guardrails against AI safety, risk, and performance.
- **Curated Evaluation Collections:** Utilize test suites tailored to specific industries and use cases.

Why it matters?

- **Solve Fragmentation:** Eliminate the need for manually stitching together disparate frameworks, benchmarks, and formats.
- **Define Metrics:** Provide clear guidance for teams struggling to identify the appropriate metrics for their specific needs.
- **Streamline Governance:** Embed compliance and governance directly into AI evaluation workflows.

What's New (Q3 2026, RHAI 3.5)

- **EvalHub Server,** SDK, CLI, **Agent Skills** and **MCP Server** move to General Availability (GA)
- **Two multi-tenant modes:** cluster level (shared instance) & tenant level (tenant dedicated instance)
- **Kueue Integration** for Multi-Tenant Evaluation
- **Agent-Discoverable** Evaluations
- OTEL Meter Provider & Service Monitor auto-provisioning
- **EvalCard Generation** per Evaluation Run
- EvalHub Local Mode (`pip install "eval-hub-sdk[server,cli]"`)
- **EDD_workflow for agents** (e.g. EDD during vibe coding sessions)
- Use **datasets** and model artifacts **from PVC and git repos**
- **New Eval Collections:** Safety & Farines, Open Telco, Long Context, Toxicity and Ethical Principles, Coding, and others.

Red Hat

Advancing AI Evaluation Strategies with EvalHub

What's New (Q3 2026, RHAI 3.5) (Dev Preview)

- **Community collections:** Korean language
- **Community adapters:** DeepEval, MTEB, RULER, SWEBench, WildGuard

What's Next (rolling)

- Community adapter for **MiDojo from asago**
- Extend **community adapters for AI security**
- Extend **community Evaluation Collections**
- Ability to export evaluation results in **EvalEval schema**

What's Next (Q4 2026, RHAI 3.6)

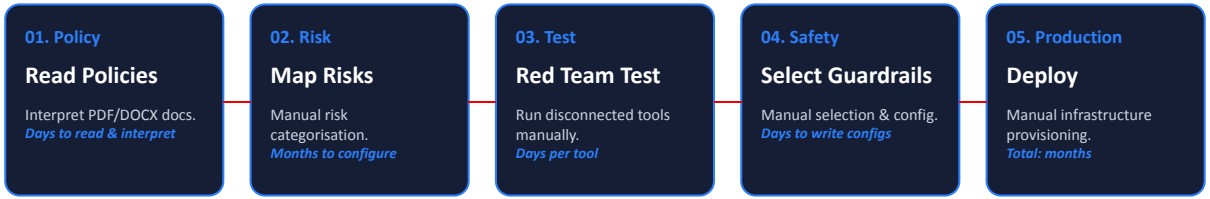
- **New GA UI with Evaluation Gallery experience**
- **UI experience for creation/customization of Evaluation Collections**
- **UI experience for importing Evaluation Collections** or catalogs from git repository
- **GA EvalHub adapters:** Inspect AI, Garak, LightEval, RAGAS, GuideLLM, IBM CLEAR
- **Sandboxing of adversarial AI evaluations**
- Native **support for ATIF** Agent Trajectory format (Harbor format)
- Support for **Harbor Hub benchmarks** via Inspect AI (inspect-harbor)
- **Deprecate LMEvalJob** in favor of native LM-Eval-Harness EvalHub adapter

Red Hat

asago

asago.ai

AI Safety And Governance Orchestration



What is it?

- **New AI Safety community: asago** (AI Safety and Governance Orchestration) automates AI safety and governance from policy to production.
- Translating a custom **AI governance policy document into a safely deployed AI system** takes months of manual work, with no end-to-end traceability.

Key components in (Dev Preview):

- **MiDojo:** Red-teams AI agents by injecting malicious payloads into fetched data to test for harmful side effects. (e.g. AI-driven **Adversary-in-the-middle** attack)
- **Policy-Mapper:** Converts AI policy documents into structured risk lists.
- **Scenario Generator:** Translates identified risks into technology-agnostic test scenarios.
- **Artifact Generator:** Converts scenarios into framework-specific red teaming tools.

Press release: Red Hat Launches asago Community to Automate AI Safety and Governance from Policy to Production

Join the asago community
https://red.ht/asago-slack

Red Hat

AI Safety: Securing AI for Enterprise Production

What are Guardrails?

- Guardrails are **independent checkpoints** in your application pipeline that **ensure secure, consistent safe behavior** of your AI application.
- **They operate in critical phases:** input validation, output filtering and prompt construction.

Why it matters?

Addressing safety concerns is the key to unlocking AI adoption.

Guardrails remove these primary blockers, enabling you to:

- **Security & Defense:** Protects applications from prompt injections, jailbreaks, and unauthorized data exfiltration.
- **Simplify Compliance & Safety:** Implement standardized, easy-to-manage security layers that ensure your AI applications remain reliable and secure.
- **Accelerate Time-to-Market:** Confidently deploy AI solutions by mitigating risks that would otherwise delay production.

What's New (Q3 2026, RHAI 3.5)

- NeMo Guardrail as the **default guardrail service**
- NeMo Guardrail **at AI Gateway (TP)**
- **Deprecation of FMS Guardrails**
- **MiDojo** for Adversarial-in-the-Middle red teaming (DP)

What's Next (Q4 2026)

- **Support for responses and messages API**
- **Support Guardrails for MCP Servers**
- **Source-app based guardrail selection & enforcement**
- Stand-alone guardrail runtime with **template rails** (built-in rails catalog)
- **Removal of FMS Guardrails**
- Refactoring of TrustyAI Service **from Java to Python**
- Renaming TrustyAI Operator to **AI Safety Operator**

Red Hat

AI Observability

What is it?

A native, unified monitoring & diagnostic framework that provides visibility into the health, performance, and behavior of entire AI stack.

Why it matters?

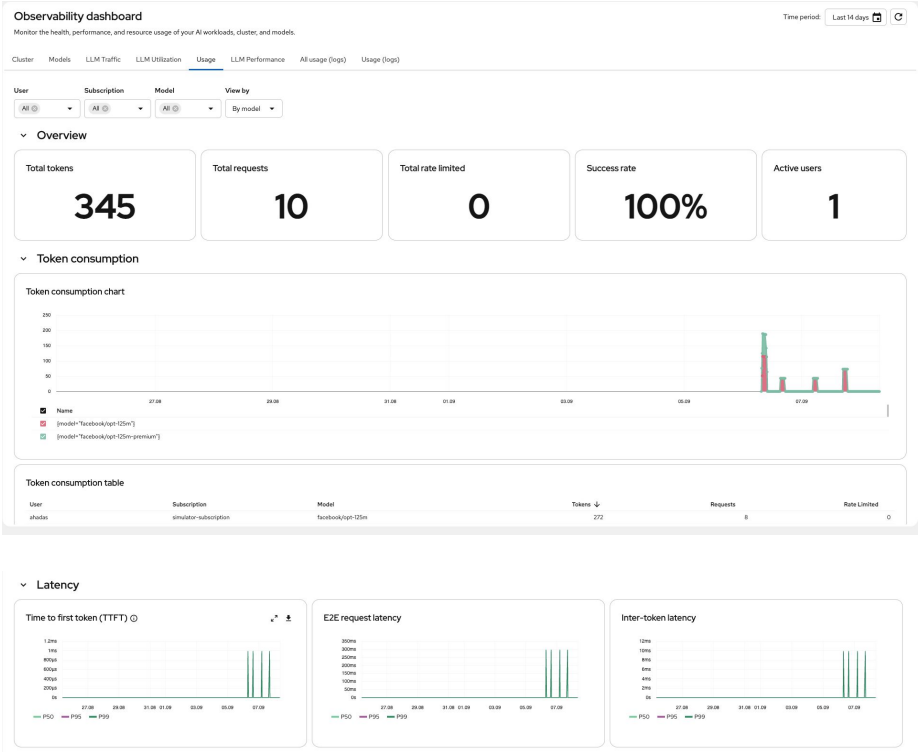
Provide native experiences to ensure HW efficiency, model accuracy, and cost-predictability in production.

What's new for Q3 2026, RHAI 3.5?

- Non-admin Usage Observability dashboard (TP)
- Loki based MaaS usage showback (TP)

What's next for 3.6

- GenAI Semantic Conventions compliant traces/metrics/logs for inference stack
- AI powered Observability and troubleshooting with backend correlation (TP)
- Dashboard as Code(DaC) platform support
- Day 0 Observability stack



Red Hat

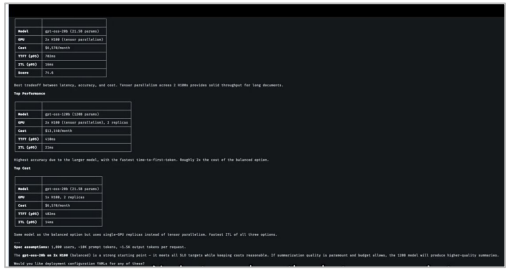
Workflow navigator

What is it?

Cluster-aware agentic orchestration platform to launch AI workflows within RHOAI

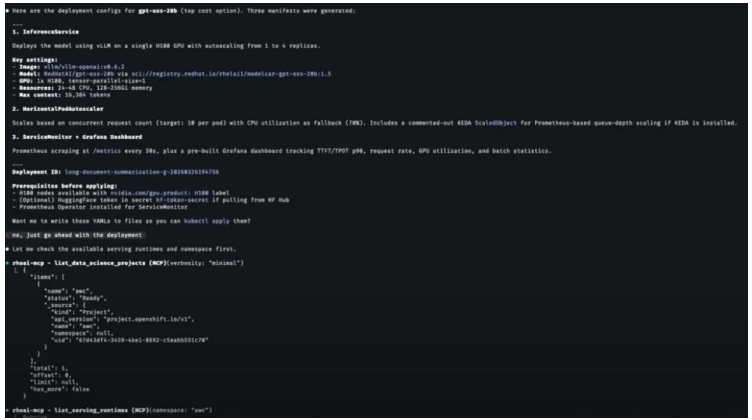
Why it matters?

Maps user intent to actionable workflows backed by evidence saving time on guesswork



What's new for RHAI 3.5, Q3 2026?

- Intelligent model recommendation (DP)
- Optimized vLLM Model deployment (DP)
- Role based access control(DP)



What's next for 3.6 2026 , 1H 2027?

- Tech Preview and General Availability
- Optimized llm-d model deployment
- Troubleshooting common Observability issues
- Security, Trust, Safety and Governance

Red Hat

Notebooks Images, Server, and Extensions

What is it?

- Notebook **images** are pre-built environments (PyTorch, TensorFlow, CUDA, etc.) with the OS, runtime, and libraries baked in
- The **server** (workbench) is the running compute instance, CPU/GPU/memory and persistent storage, that hosts your chosen image
- **Extensions** let you add custom images, packages, or JupyterLab plugins on top, plus hook into pipelines/model serving

Why it matters?

- Standardized, reproducible environments eliminate setup headaches and speed onboarding
- Extensibility keeps the platform adaptable to new tools and the full MLOps lifecycle, without waiting on platform upgrades

What's New (Q3 2026, RHAI 3.5)

- **Kubeflow Automated Pipelines Engine (KALE integration for Developer and Technical Preview)**: data scientists can annotate Jupyter cells and submit pipelines to AI Pipelines without rewriting notebook logic in KFP SDK code
- **Notebooks 2.0 (DP):** a next-gen workbench model built around WorkspaceKinds, configurable templates for IDE type, resources, images, secrets, and storage
- **RStudio Deprecation:** now self-build recipes for customers who want to run their own unsupported RStudio image
- **Hermetically Built Notebooks Images for RHAI Index Adoption:** notebook images are built from pinned, reproducible inputs and faster security fixes and image refreshes

What's Next (RHAI 3.6, 1H 2027)

- **Support for ARM Architecture:** delivers multi-arch parity: native aarch64 builds of supported OOTB workbenches
- **KALE GA:** supported, production-ready notebook-to-pipeline experience
- **Notebooks 2.0 Tech Preview:** trial the new model while a migration path from Notebooks 1.x is defined
- **OOTB Coding Assistance with code-server and Jupyter IDEs:** self-hosted AI coding assistance into RHOAI workbenches so users don't need GitHub Copilot, Cursor, or manual setup

Red Hat

Run distributed workloads: different kinds

training, notebooks, and pipleines

What is it?

- **Kubeflow Trainer: abstracts the infrastructure complexity of multi-node distributed fine-tuning** – scaling, orchestration, and gang lifecycle are handled for you
- **Ray: the distributed compute engine** for parallel workloads – hyperparameter tuning, data-parallel preprocessing, and scale-out training/serving
- **Kueue: cluster-level job queueing and quota management** – admits training jobs only when GPU quota is available, with gang admission so distributed jobs start whole or not at all
- **Together** they form the distributed workloads stack: Training Hub defines what to train, **Kubeflow Trainer/Ray define how it** runs distributed, Kueue decides when and where it's admitted

Why it matters?

- **Multi-node training** becomes a config change, not an infrastructure project
- **Quotas, priorities, and preemption enable fair**, governed multi-tenant GPU sharing
- **Data scientists stay in notebooks**; platform teams **enforce policy underneath** – separation of concerns that scales

What's New (Q3 2026, RHAI 3.5)

- MLflow integration for AI Pipelines and training environments
- Deprecated Kubeflow Training operator v1
- Known issue: RHOAIENG-87834 - Kueue-managed TrainJobs are not scheduled when using Red Hat Build of Kueue 1.4 or later

What's Next (RHAI 3.6, 1H 2027, uncommitted)

- **Deeper Kueue integration** across the platform
- **Better visibility for users and admins**: quota hierarchy visualisation, queue and workload visibility for data scientists, and per-tenant GPU usage export
- Smarter placement: **topology-aware scheduling with Kueue** to co-locate distributed jobs on well-connected hardware

Red Hat

Managing workloads with GPU-as-a-Service

What is it?

Integrated GPU visibility, scheduling, and governance built on Kueue, MIG/DRA, and GPU metrics.

Why it matters?

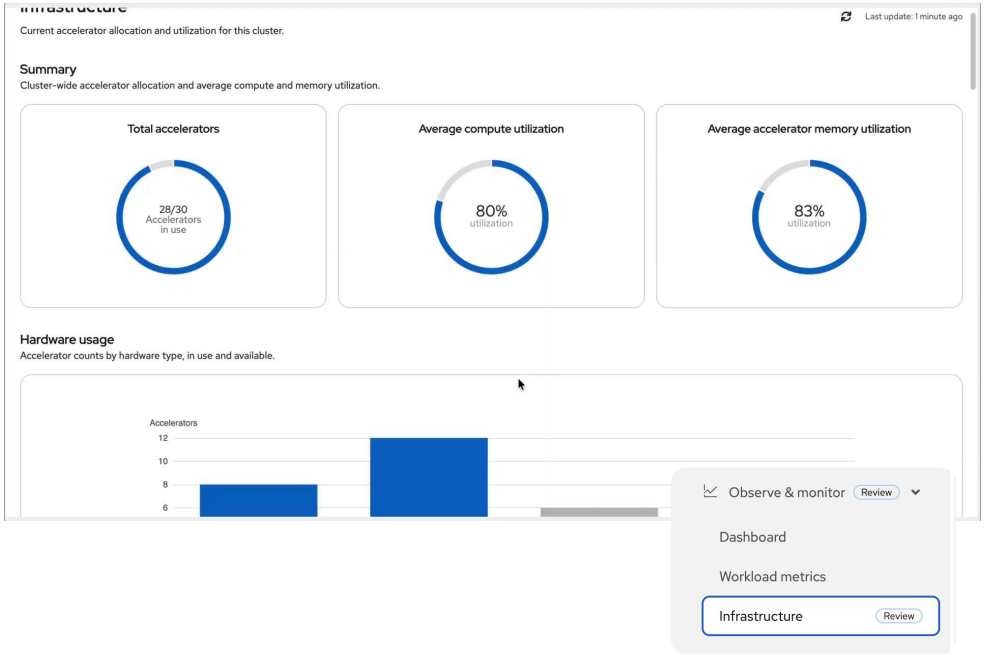
See your entire GPU fleet, enforce fair-share quotas across teams, and give users self-service access to the right compute.

What's New for RHAI 3.5?

- GPU Topology and Utilization Dashboard
- Scheduling visibility in Workbenches

What's Next for RHAI 3.6 and 1H'27?

- Kueue integration for Kserve workload preemption
- Kueue UI: org-to-team quota tree
- Fine-grained GPU allocation via Dynamic Resource Allocation: DRA support for fine-grained allocation and for the workload APIs (TP)
- Per-tenant GPU usage export to enable showback
- Kueue Topology Aware Scheduling



Red Hat

Native Multi-Tenancy for Neo and Sovereign Cloud Providers

What is it?

- Red Hat AI now officially supports running on Red Hat OpenShift Hosted Control Planes deployed on Red Hat OpenShift Virtualization
- Each tenant gets a dedicated OpenShift control plane, while their AI workloads run in dedicated VMs on shared, GPU-enabled infrastructure

Why it matters?

- Neoclouds, GPU cloud providers, and regulated/sovereign AI customers need hardware-to-software tenant isolation — not just namespace-level separation
- Infrastructure providers still manage and upgrade the entire fleet from a single control point, even though every tenant gets their own dedicated, self-service control plane
- Turns one shared bare-metal GPU pool into a multi-tenant AI cloud without standing up a physical cluster per tenant

What's New (Q3 2026, RHAI 3.5)

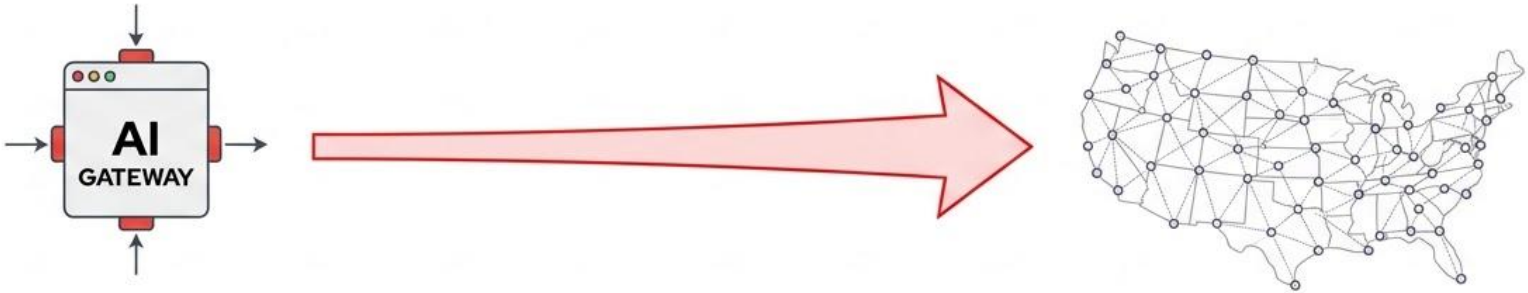
- GA: officially supported configuration for OpenShift AI on Hosted Control Planes running on OpenShift Virtualization
- Worker nodes are automatically provisioned as VMs via HyperShift + KubeVirt — no manual per-tenant installs

What's Next (RHAI 3.6, 1H 2027, uncommitted)

- Fractional GPU sharing (NVIDIA vGPU / MIG) on HCP-on-Virtualization tenant clusters,
- Scale validation to 5+ concurrent tenant clusters sharing one management cluster
- Per-tenant GPU usage/metering reference architecture

Red Hat

AI Gateway -> AI Grid



AI Gateway

From Centralized Traffic Management
to Distributed Workload Placement

AI Grid

MLflow in RHOAI

What is it?

A unified, open-source engine for tracking, evaluating, and managing the e2e lifecycle of both gen AI agents and predictive ML models.

Why it matters?

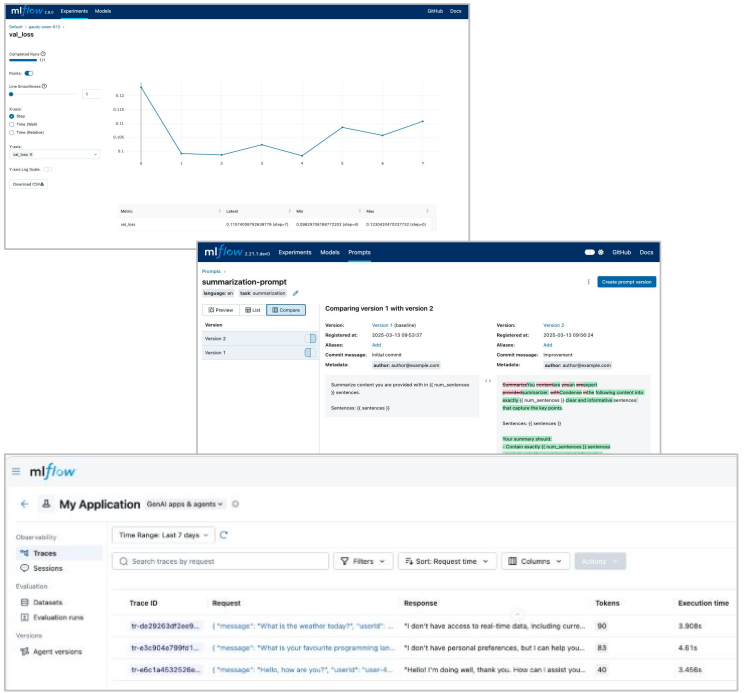
Provides a centralized system of record needed to move AI from experiment to production.

What's new for RHAI 3.5, Q3 2026?

- MLflow integration for AI Pipelines and training environments
- Enhancements to run agent observability at scale.
- Shared prompt templates
- Multi-model provider support
- Multi-tenancy with OAuth2, JWKS, ABAC

What's next for 1H 2027?

- Integration with Inference and AI gateway
- Agent Registry
- Skills registry



Red Hat

Red Hat AI Inference Roadmap Summary

RHAII 3.5

August 2026

AI Accelerator enablement

- AMD MI350P PCIe
- NVIDIA DGX Spark with the GB10 SoC (vLLM)
- NVIDIA RTX 6000D (3.5.1 and fast.1)

New features

- **Monthly Supported FAST release of RHAII vLLM**
- **Beyond autoregressive.** DiffusionGemma for text. Diffusion/DiT image and video through Omni.
- **Faster Tokens.** NVFP4 KV cache/fused MoE, MXFP4, TurboQuant, Speculative decoding with MTP, DFlash, DSpark.
- **Hardened Prefill-Decode Disaggregation with requesdraining for scaling;** DP-aware load balancing, Hardened DeepEP and elastic-EP for large open weight models.

- **Flow control:** Prioritize critical traffic under load (GA)
- **Tiered KV-cache offload:** Reuse cached prompts across GPU, CPU (GA) and POSIX filesystem storage (DP)
- **Controlled deployments:** Test model updates gradually and roll back without disrupting traffic (GA)
- **Inference observability:** Troubleshoot serving health through Perses dashboards, Service Level Metrics and Distributed Tracing (GA)
- **Distributed Inference:** Azure AKS (GA), CoreWeave CKS (GA); Amazon EKS (TP)

RHAII 3.6

November 2026

AI Accelerator enablement

- NVIDIA Vera Rubin

New features

- **vLLM Rust front-end for high-throughput low-latency inference**
- **vLLM granular cache** accounting for cache usage tracking
- **Omni TP on RHOAI ServingRuntime** with real time voice intelligence
- vLLM Watermarking for AI-generated context (3.6 GA)

- **Multi Cluster routing:** Reduce response time and GPU work by routing requests across clusters based on inference state and cached context
- **Fast model server startup:** Reduce cold starts during scale-up, restart, and model switching
- **llm-d orchestration for vLL Omni:** Coordinate Multi Modality Input and Output (DP)

DP: Developer Preview - **TP**: Tech Preview - **GA**: General availability

Red Hat

Red Hat OpenShift AI Roadmap Summary

RHOAI 3.5

August 2026

Inference
- MaaS - external models & OIDC support (GA); multitenancy & user showback dashboards (TP)
- llm-d: flow control (GA), tiered KV cache offload (GA), observability (GA)
- Vllm - optimized serving for newer/validated models
- Model validation for tool calling
- Faster delivery of vllm to accelerate support for new models

Model development & Alignment
- AutoRAG (TP)- multilingual support, AI playground integration
- AutoML (TP)- OOTB serving runtime support, enhanced experimentation
- GRPO / RLVR fine-tuning (GA)
- Distributed fine-tuning with Ray enhancements (GA)

Agentic
- OpenShell (DP)
- Agent Catalog (DP)
- Responses & RAG APIs (GA)
- Gen AI Studio enhancements (TP)

AI platform
- Eval Hub (GA)
- AI Safety - MiDojo (DP), Policy-Mapper (DP)
- Observability - user/non-admin dashboards (TP)
- MLflow integration with AI Pipelines (GA)
- Workflow Navigator (TP)

RHOAI 3.6

November 2026

Inference
- AI GW / MaaS - multitenancy (GA), guardrails integration, consumer portal, intelligent routing (DP)
- Multimodel / omni models (TP)
- Llm-d - fast model server startup
- Expanded validated models, including tool calling

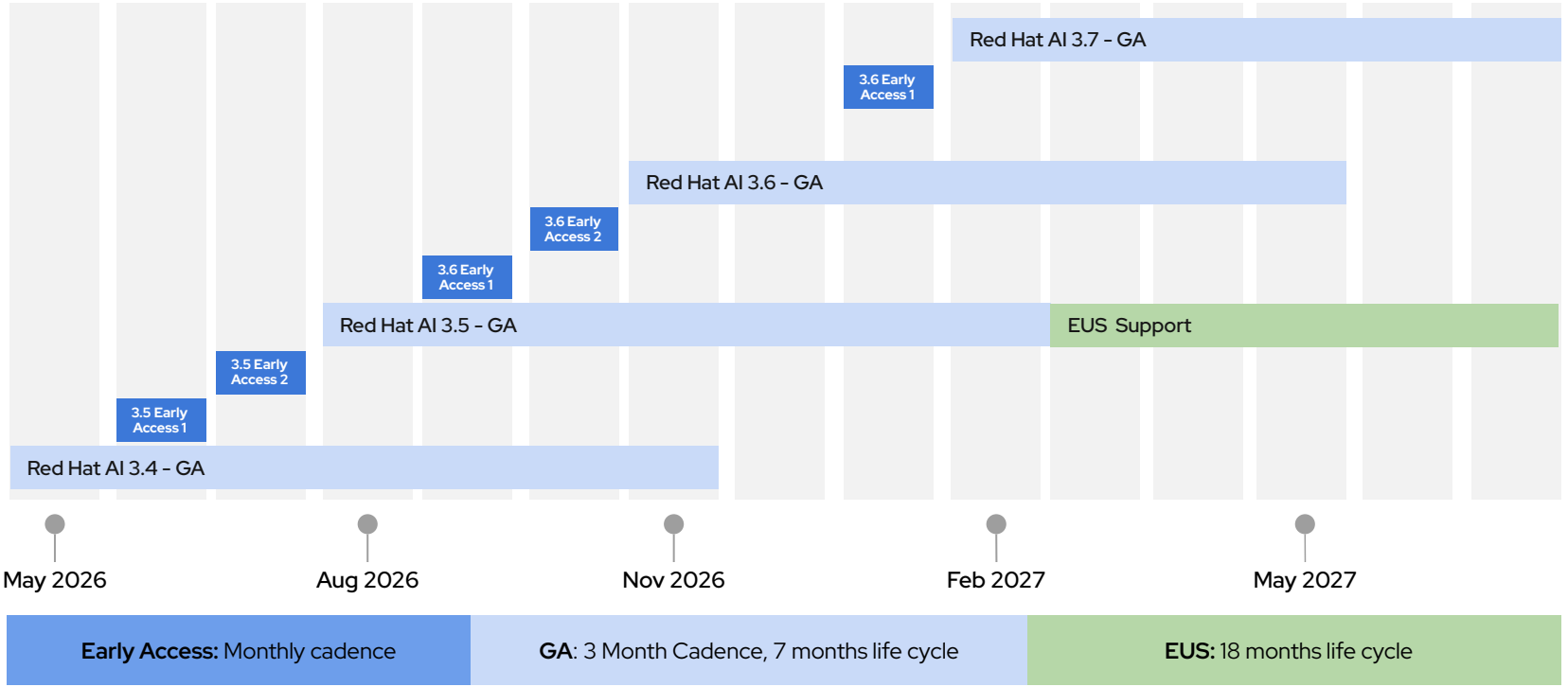
Model development & Alignment
- AutoRAG (GA)
- AutoML (GA)
- Graph AutoRAG (TP)
- Discover & use data artifacts

Agentic
- OpenShell (TP)
- MCP Catalog, Gateway, Registry & lifecycle (GA)
- Skills catalog & registry (TP)
- GenAI Studio Playground (GA)

AI platform
- Eval Hub - new & enhanced UI experiences
- AI Safety - support for responses & messages APIs & guardrails for MCP servers
- AI powered observability and troubleshooting
- Notebooks 2.0 (TP)
- Enhance e2e GPUaaS experience

Red Hat

Red Hat AI life cycle



Red Hat

Learn more, and try out, Red Hat AI

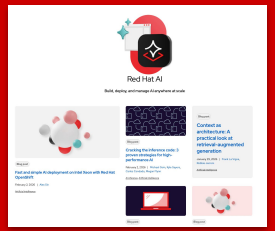
Accelerate your AI journey with the Red Hat **AI quickstarts**

red.ht/blog-ai-quickstarts

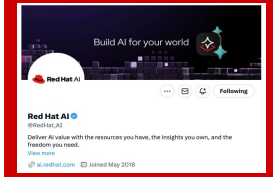
AI quickstarts home page: red.ht/ai-quickstarts

Subscribe to the **Red Hat AI blog**

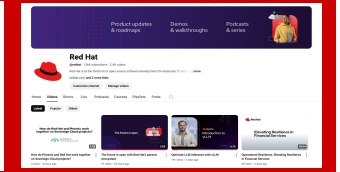
red.ht/rhai-blog



Follow us on X and YouTube



x.com/RedHat_AI



youtube.com/@RedHat

Interactive demos

- red.ht/InferenceBenchmark
- red.ht/intro-inference
- red.ht/model-optimization
- red.ht/model-customization
- red.ht/intro-llm-d
- red.ht/MaaS

60-day trial of **Red Hat AI Enterprise**
red.ht/AI-enterprise-trial

60-day trial of **Red Hat AI Inference**
red.ht/AI-Inference-trial

60-day trial of **Red Hat OpenShift AI**
red.ht/RHOAI-trial

Red Hat

Red Hat AI

Thank you

Red Hat is the world's leading provider of enterprise open source software solutions. Award-winning support, training, and consulting services make Red Hat a trusted adviser to the Fortune 500.


linkedin.com/company/red-hat


youtube.com/user/RedHatVideos


facebook.com/redhatinc


twitter.com/RedHat

Red Hat