

Red Hat AI/ML or AWS AI/ML

Where can you deploy AI: Red Hat vs AWS

	Baremetal on-prem	SaaS	AWS	OpenShift Dedicated/ GCP	Edge	ARM	IBM Power, Z
RHO AI	✓	IBM Cloud with ✓ WatsonX	✓ ROS A	✓	✓	H2 2025	✗
AWS	✗	✓	✓	✗	✓ Outposts	✗	✗

AWS AI offerings and Red Hat

AWS AI offering	Red Hat approximate
Q - A generative AI assistant- insights on their data and accelerate their tasks; agentic capabilities	Can build with Red Hat AI but not OOTB
Bedrock - Managed service offers foundation models from AI21 Labs, Anthropic, Cohere, Luma,, Meta, Mistral AI, poolside, Stability AI, and Amazon through a single API; customize with fine-tuning and (RAG), and build agents that execute tasks using enterprise systems	RHEL AI but with Granite or bring your own
SageMaker - Managed service - tools to enable (ML): build, train and deploy ML models using notebooks, debuggers, profilers, pipelines, MLOps; governance. Access to pre-trained models.	OPC AI
App Studio - AWS App Studio natural language to build business applications requiring deep software development skills.	None



AWS ML offering	Red Hat approximate
Deep Learning AMIs - Frameworks, dependencies, and tools for deep learning on EC2. Amazon Linux and Ubuntu, AMIs preconfigured with TensorFlow, PyTorch, NVIDIA CUDA drivers and libraries, Intel MKL, Elastic Fabric Adapter (EFA), and AWS OFI NCCL plugin,.	RHELAI & RHOAI
Deep Learning Containers - Prepackaged Docker images.; optimized model training for TensorFlow, PyTorch, and Apache MXNet.; Build ML workflows for training, validation, and deployment integration with Amazon SageMaker, Amazon EKS, and Amazon ECS.	RHEL Image mode, RHEL AI
Hugging Face on AWS - Access, evaluate, customize, and deploy FMs through Amazon SageMaker on NVIDIA GPUs, AWS Trainium and AWS Inferentia. AWS and Hugging Face collaborate on open-source Optimum Neuron library in DLCs built for AWS AI chips.	None
TensorFlow on AWS - Fine-tune applications with visualization tools, including histograms and graphs, to quickly train deep neural networks.	RHOAI
Pytorch on AWS - PyTorch on AWS is an open-source deep learning (DL) framework that accelerates the process from ML research to model deployment.	RHOAI
ApacheMxNet - Apache MXNet on AWS is a fast and scalable training and inference framework with an easy-to-use, concise API for machine learning. MXNet includes the Gluon interface.	None
Jupyter - Managed Jupyter notebooks	RHOAI Workbench
Transcribe - Convert speech to text	None OOTB - but can build using Red Hat AI
Polly - Convert text to speech	None OOTB - but can build using Red Hat AI and Docling
Textract - Extract text handwriting and data from scanned documents	None OOTB - but can build using Red Hat AI and Docling
Rekognition - Extract info and insights from videos	None OOTB - but can build using Red Hat AI
Translate - Language translation	None OOTB - but can build using Red Hat AI
Lex - Conversational interfaces for natural language models	None OOTB - but can build using Red Hat AI

