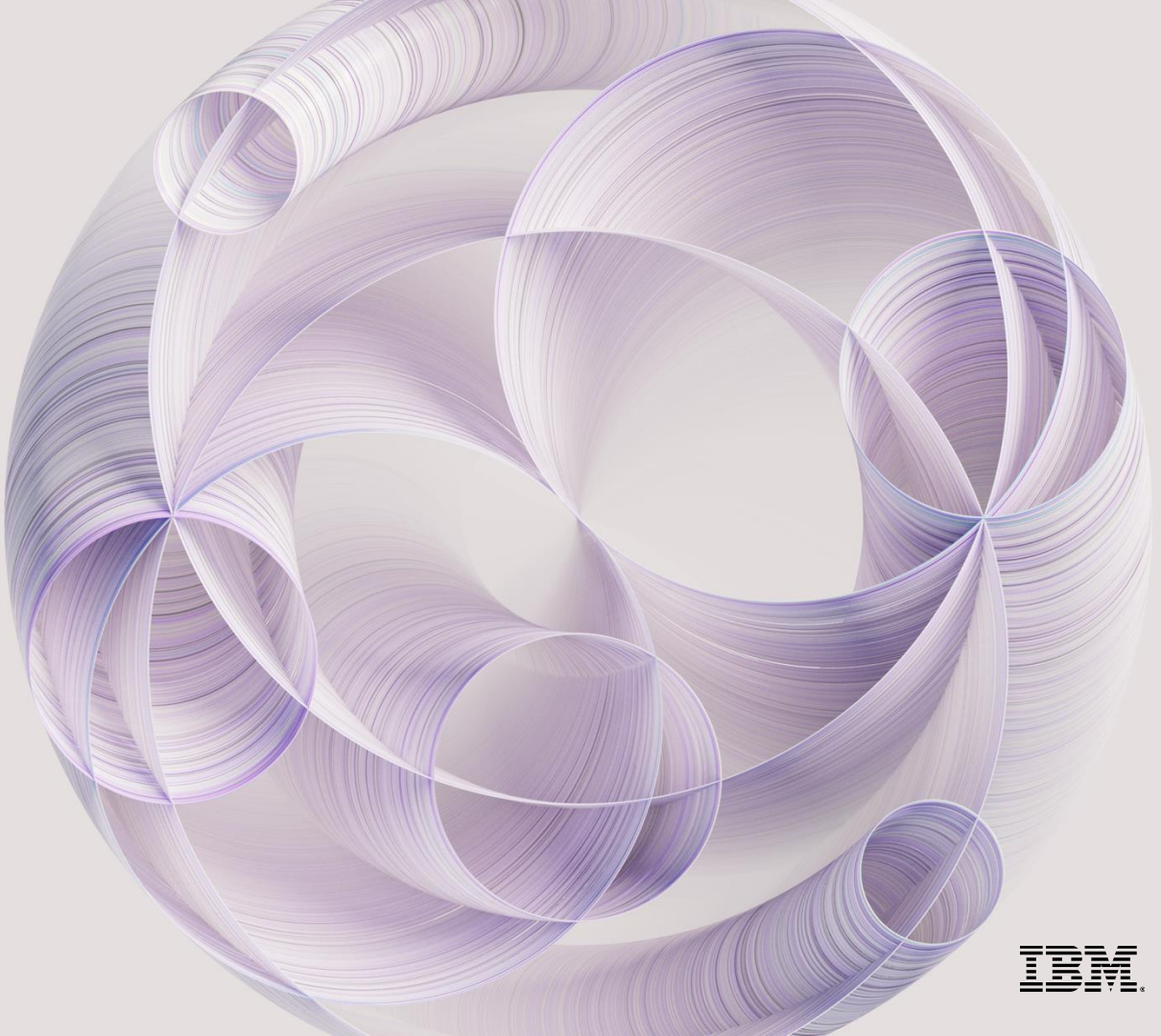
# Streamline the development of AI applications with watsonx.ai



watsonx.ai

IBM

# IBM Software

## Hybrid Cloud

Unify on-prem, public, private clouds and edge to scale virtualization and AI across environments

## Transaction Processing

Deliver unmatched transactional performance, security and reliability

## Automation

Automate technology lifecycle management with AI for productivity, resiliency and spend optimization

## Data

Access trusted and secure data to drive AI productivity

AI

Open and Trusted

# Data

Access trusted and secure data to drive AI productivity

## AI Productivity

Reinvent how work is done with AI agents/assistants

– AI Assistants

## AI/ML Ops

Work with AI models, tools and governance that's built for business—engineered to ensure trust and scalability in applications

– AI Models
– AI Tools
– AI Governance

## Data Fabric

Bring all your business data together and optimize how it moves through your systems to scale analytics and AI in your applications while protecting it.

– Databases
– Data Intelligence
– Data Integration
– Data Security

## Data Storage

Store data across edge, core and clouds

– Software-defined Storage

# Data

Access trusted and secure data to drive AI productivity

## AI Productivity

Reinvent how work is done with AI agents/assistants

### AI Assistants

watsonx Code Assistant™

watsonx Orchestrate™

Planning Analytics

## AI/ML Ops

Work with AI models, tools and governance that's built for business—engineered to ensure trust and scalability in applications

### AI Models

Granite™

Meta Llama

Mistral

### AI Tools

watsonx.ai™

### AI Governance

watsonx. governance™

## Data Fabric

Bring all your business data together and optimize how it moves through your systems to scale analytics and AI in your applications while protecting it

### Databases

watsonx.data™

### Data Intelligence

Data Product Hub

Knowledge Catalog

Manta Data Lineage

### Data Integration

DataStage®

Databand®

Streamsets

### Data Security

Guardium® Data Security Center

## Data Storage

Store data across edge, core and clouds

### Software-defined Storage

Storage Ceph®

# watsonx

A portfolio of AI products that accelerates the impact of generative AI in core workflows to drive productivity.

## watsonx.ai

Enterprise-grade AI studio that helps AI builders innovate with all the APIs, tools, models, and runtimes to build AI solutions

Featuring **IBM Granite,** and popular third-party models including **Mixtral, Llama** series

## watsonx.data

The **hybrid, open data lakehouse** to power AI and analytics with all your data, anywhere

## watsonx.governance

End-to-end toolkit for AI governance to manage **risk and compliance across the entire AI lifecycle**.

## watsonx Orchestrate

An enterprise-ready solution that helps create, deploy, and manage AI assistants and agents to automate processes and workflows.

## watsonx Code Assistant

Accelerate development, **application modernization, and assist with IT Operations**

# Why IBM watsonx for scaling enterprise AI to drive productivity

## Open

→ Offers choice to train the right foundation models, including open-source models, and the choice of data, tools, and frameworks to achieve desired business outcomes.

→ Run AI wherever the business needs to, across any cloud, at scale.

## Trusted

→ Built with open and transparent technology to give enterprises confidence in their AI and meet regulatory compliance demands.

→ Responsible AI and protected data backed by enterprise governance and security controls.

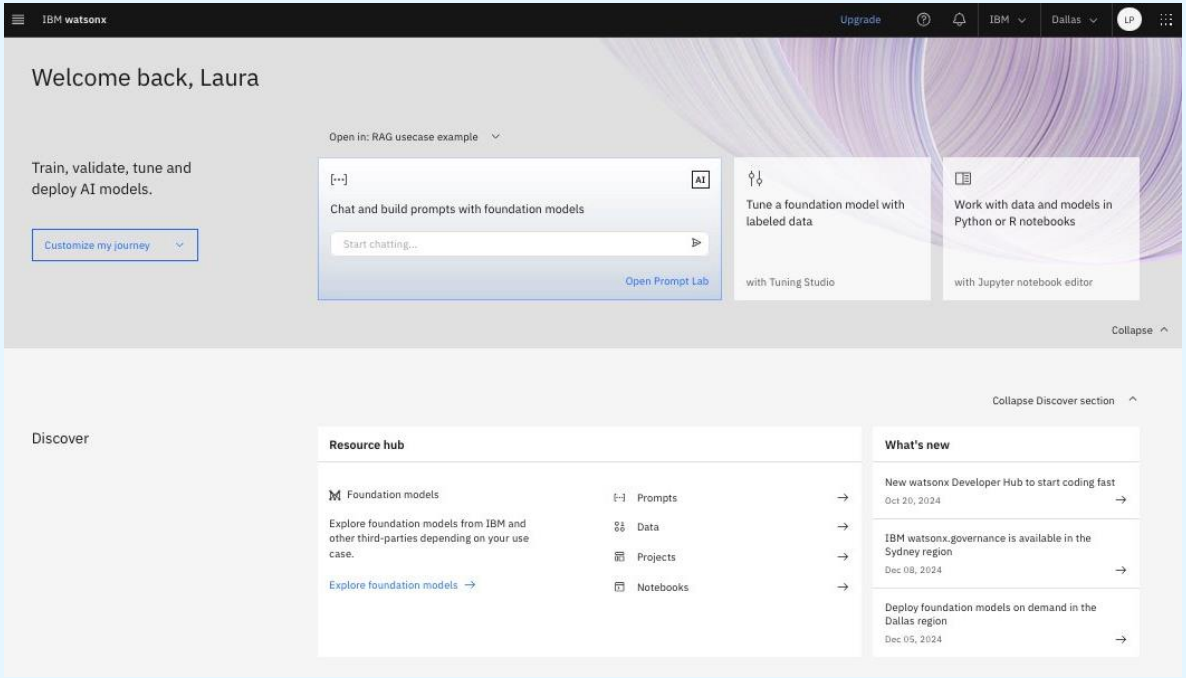
## Integrated

→ Integrates technology seamlessly into existing infrastructures, systems, and processes with choice of cloud to transform the enterprise and drive productivity from within.

→ Embedded AI for targeted use cases that drives enterprise scale productivity.

# Streamline the development of AI applications with watsonx.ai

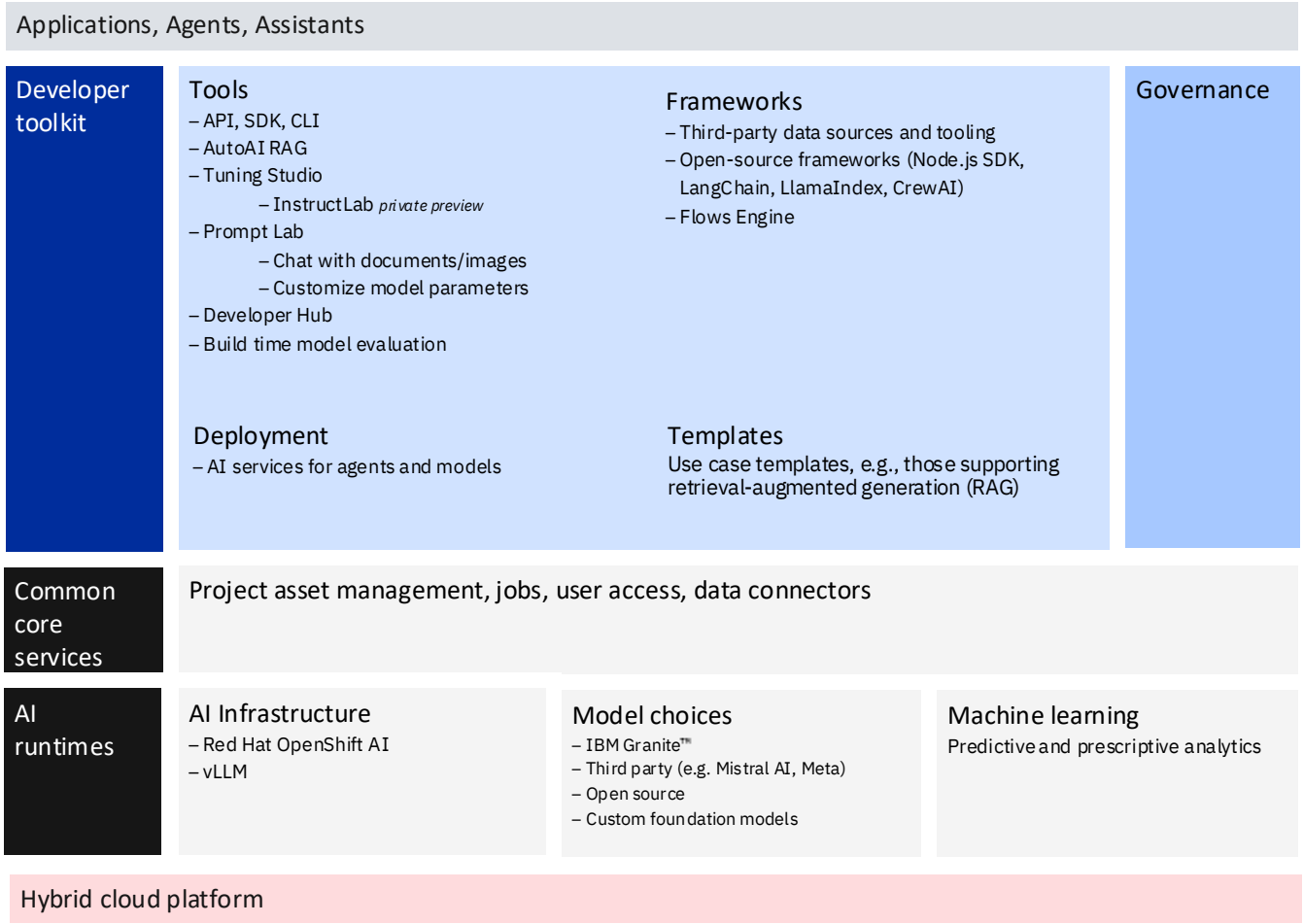
- Model selection
- Model customization
- AI development and deployment
- AI governance



# IBM **watsonx.ai**

The enterprise-grade AI developer studio

Integrated AI platform with all the APIs, tools, models, and runtimes to simplify and scale the development and deployment of AI



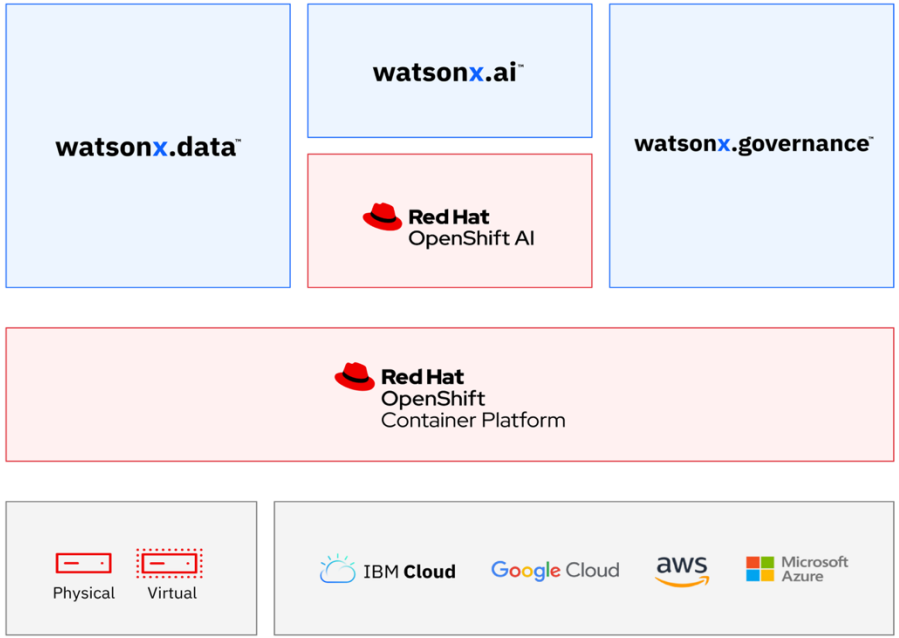

© 2025 IBM Corporation

# watsonx.ai and Red Hat

**Watsonx.ai** integrated with **Red Hat OpenShift AI** & **Red Hat OpenShift Container Platform** delivers an accelerated time to value through an optimized generative AI and ML workflow.

Key benefits:

- A complete platform for managing AI workloads, training models and scaling AI deployments
- Access to high quality, governed, and trusted AI through curated IBM-developed, third-party, and open-source foundation models
- Deploy to on-premises environments, private or public clouds
- Integrations to the wider watsonx and Cloud Pak for Data platforms.




© 2025 IBM Corporation

# Creating Generative AI Solutions

# Model Selection

# watsonx.ai Provided Models

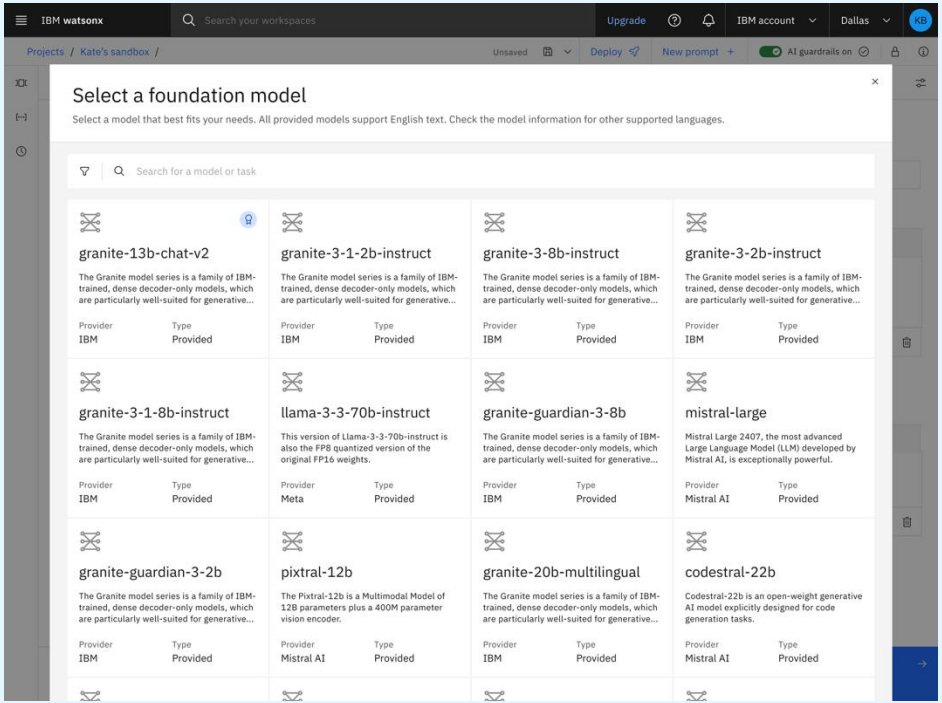
Most models available in all regions

A highly curated library with IBM-built, open-source and third-party foundation models that are pre-deployed and ready to use in watsonx.ai for a wide range of business requirements and budgetary considerations. These models are deployed on multi-tenant GPU's.

Key benefits:

- Great for experimentation and prototyping to determine which model is best for your use case
- Pay only for the tokens that are consumed, no time commitment
- Ideal for clients that need periodic or infrequent use of foundation models
- Indemnification offered for IBM and select models

Learn more in documentation




© 2025 IBM Corporation

# Deploy on Demand Models

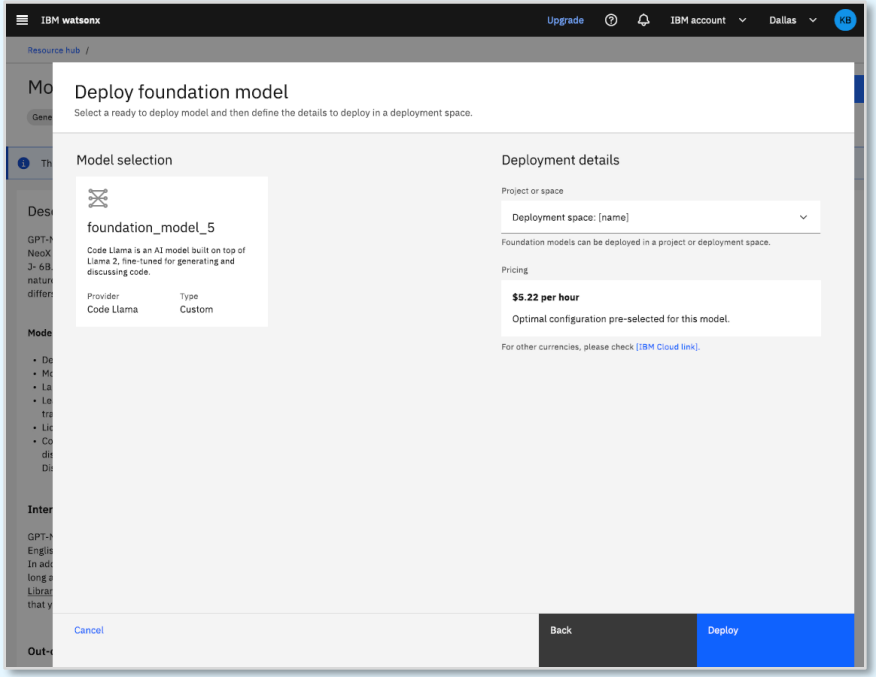
Dallas region, Standard Plan only

A curated collection of high quality and popular foundation models that can be easily deployed to a dedicated GPU with just a few clicks.

Key benefits:

- Faster interactions with a dedicated, single-tenant deployment, hosted 24/7 until de-provisioned by client
- Predictable hourly hosting price compared to variable token-based pricing
- Supports full context length of the model
- No rate limits on the inference requests per second (vs. 8 req/s limit on pre-deployed models)

Learn more in documentation




© 2025 IBM Corporation

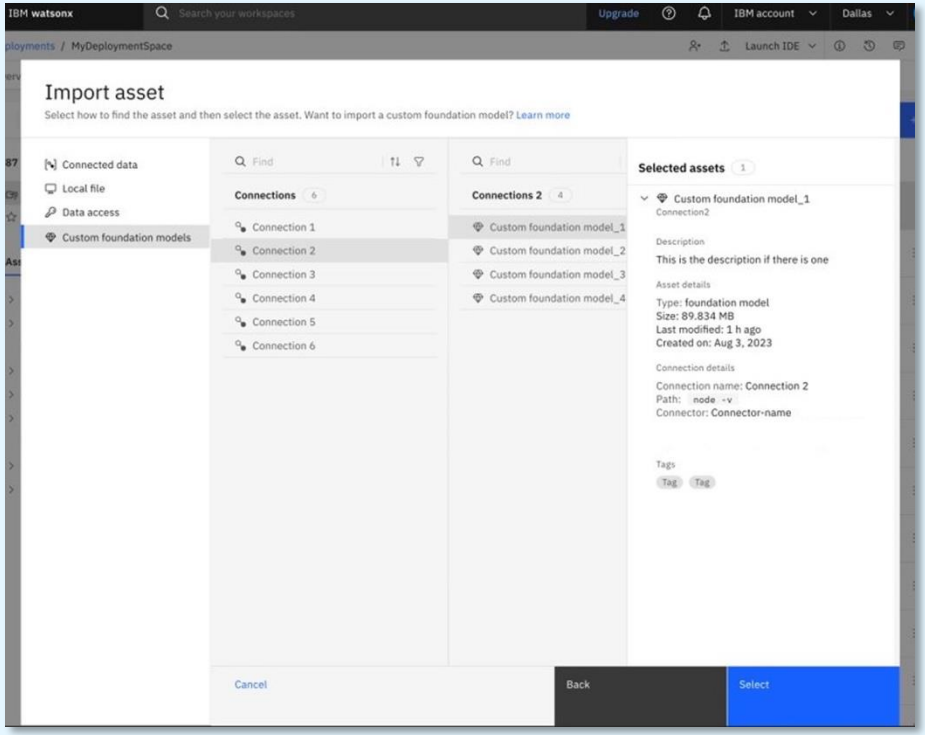
# Custom Foundation Models

Import custom foundation models for maximum flexibility in how generative AI solutions are developed.

Key benefits:

- Leverage externally fine-tuned LLMs that support a specific language or are customized for an industry or business domain
- Import a model that is not already provided by IBM.
- Compatible with 1000's of models from repositories like Hugging Face, which provides access to a huge selection of open-source foundation models

Learn more in the blog, tutorial or documentation




© 2025 IBM Corporation

# Access **1000+ models** in watsonx.ai from a mix of platform-provided models and custom foundation models available for import

| | Provided Models (Ready to use models) | Deploy on Demand | Import Custom Foundation Models | Software |
|---|---|---|---|---|
| **Overview** | Multi-tenant SaaS offering where capacity is shared | Single-tenant SaaS offering with curated set of models from IBM | Single-tenant SaaS offering with customer-provided models | On-premises deployment on customer infrastructure |
| **Cost Structure** | Token-based pricing, billed on both input and output tokens; available for all SaaS Plans | Hourly pricing - includes hosting and inference; billed while model is deployed; Standard Plan only | Hourly pricing - includes hosting and inference; billed while model is deployed; Standard Plan only | No additional model pricing; client is billed on software entitlement |
| **Benefits** | • Pay only for the tokens that are consumed<br>• Indemnity offered for IBM and select models | • Predictable cost based on hourly model deployment<br>• Dedicated GPU capacity<br>• Ease of deployment<br>• Indemnity offered for IBM and select models | • Predictable cost based on hourly model deployment<br>• Dedicated GPU capacity | • Maximum flexibility in choosing which models to deploy and on what kind of GPUs<br>• Air-gapped environment for secure workloads |
| **Drawbacks** | • GPU capacity is shared; unable to take full advantage of the throughput | • Deprovision the model when not in use to optimize costs | • Requires user to import relevant model files<br>• Deprovision the model when not in use to optimize costs | • Infrastructure setup required |
| **Best Fit for** | • Experimentation<br>• Small companies with infrequent use | • Large production workloads<br>• Enterprises with SLA requirements | • Large production workloads<br>• Enterprises with SLA requirements | • Large production workloads<br>• Enterprises with SLA and security requirements |



© 2025 IBM Corporation

# Model Customization

# watsonx.ai Tuning Studio

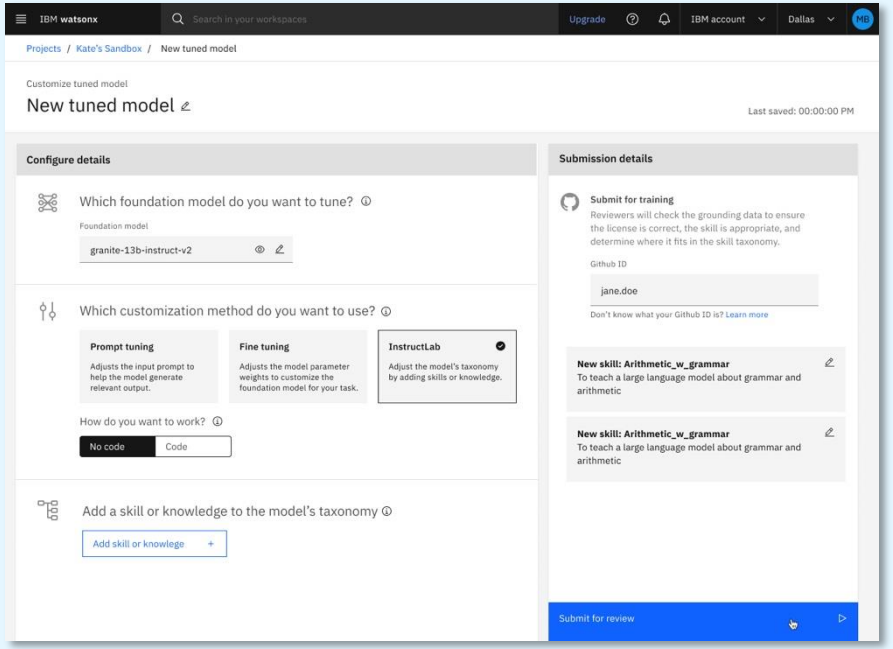
Improve foundation model performance through customization using a variety of methods:

- Fine Tuning
- LoRA/QLoRA (PEFT)

Easy to use, step by step interface to accelerate the setup time.

Customized models can be deployed within watsonx.ai and inferenced through the Prompt Lab or API.

Learn more in documentation




© 2025 IBM Corporation

# Tuning Studio: Fine Tuning

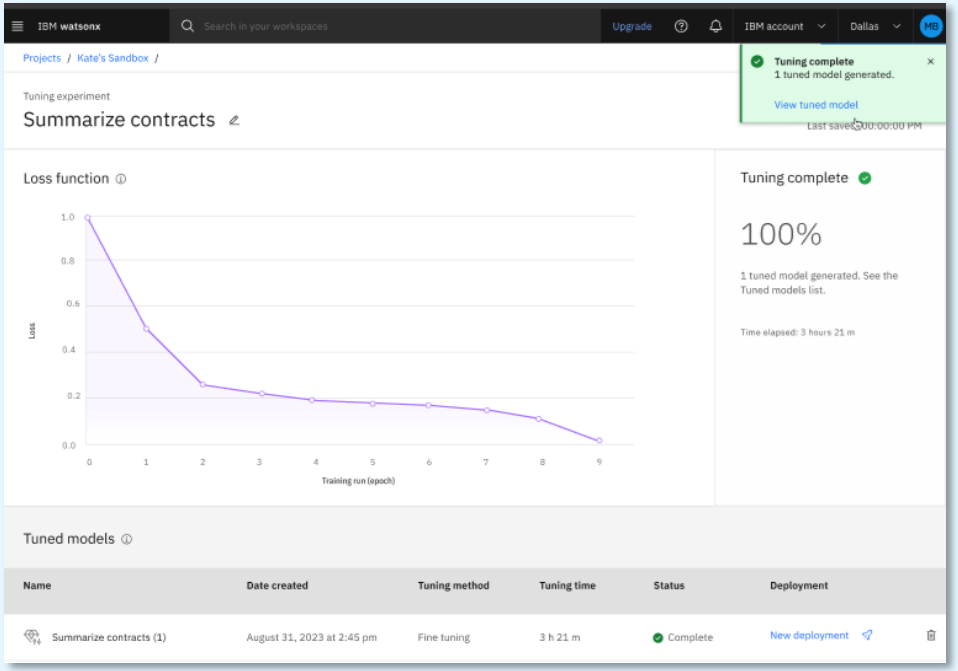
Software version only

Unlock the full potential of foundation models by customizing it for a specific task, use case or business domain with structured data.

Key benefits:

- Improved performance and accuracy
- Reduced Bias from original training data
- Easy to use interface that supports structured training data in files or connected data sources
- Adjust parameters and view performance metrics

Learn more in documentation




© 2025 IBM Corporation

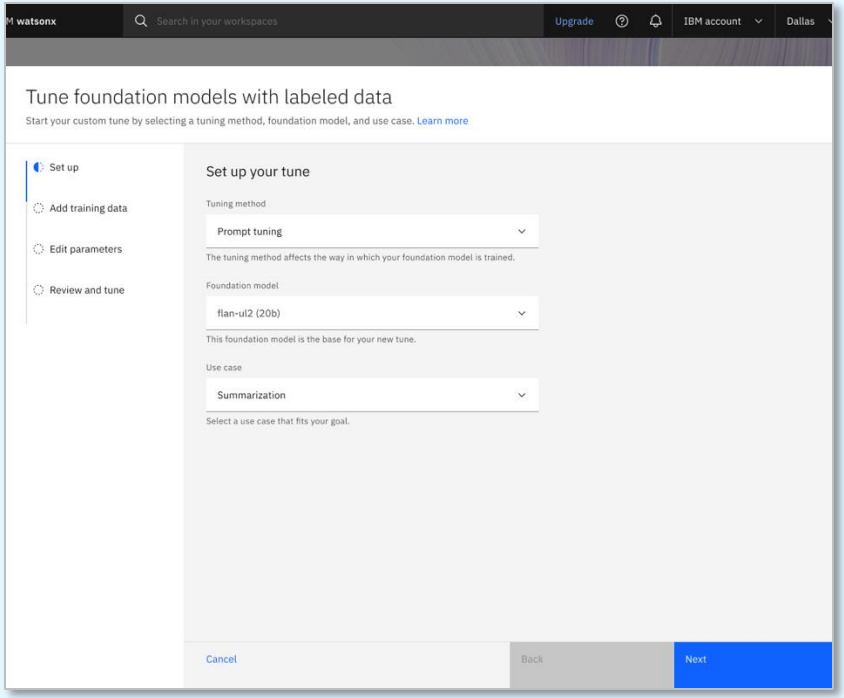
# Tuning Studio: LoRA/QLoRA Tuning

Leverage LoRA and QLoRA to further train a model with domain data. This method drastically reduces memory, computation, and time compared to full fine-tuning.

Key benefits:

- Efficient, low-cost way of adapting an AI foundation model to new tasks
- Tune the model by changing only a subset of model parameters (not the entire base model)
- Achieve comparable results to full fine-tuning depending on base model size and data used
- Minimize the risk of catastrophic forgetting

This method is a subset of Parameter Efficient Fine Tuning (PEFT). Learn more in documentation.

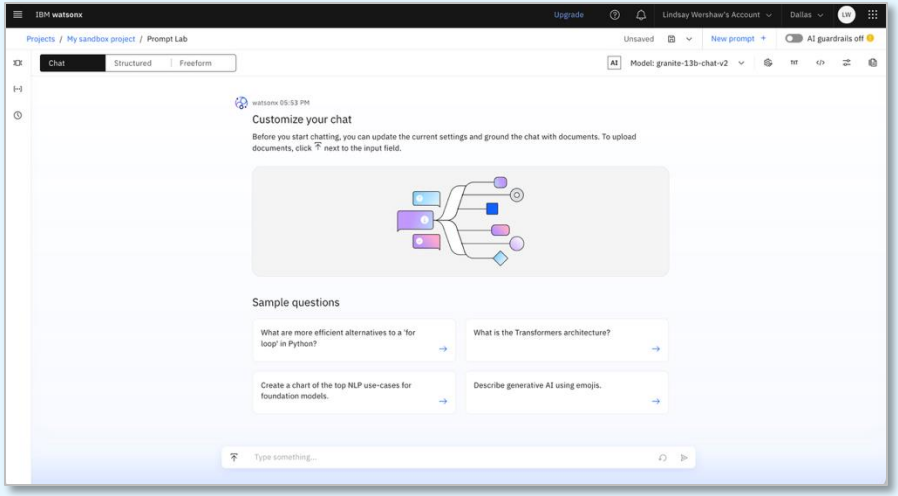


# Prompt Lab

# watsonx.ai Prompt Lab

Experiment with foundation models through an interactive user interface or an API.

- Easy to use chat, structured and free form prompt building interfaces
- Experiment with zero-shot, one-shot, or few-shot prompting to get the best results
- Adjustable AI guardrails
- Includes prompt examples for various use cases and tasks
- Save and share prompts for team collaboration
- Wide selection of foundation models to meet any task requirement
- Adjust model parameters to optimize results such as sampling, min/max tokens, stop sequences, repetition penalties, and more
- Export prompts and settings to a notebook to jump start development




Learn more in documentation

© 2025 IBM Corporation

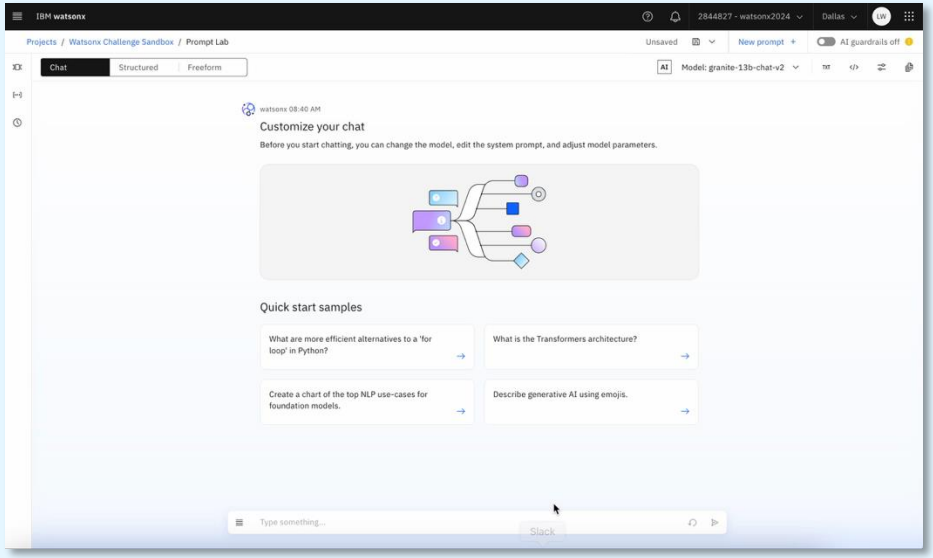
# Chat with Documents

The Chat with Documents feature, within the Prompt Lab, provides an easy way to support RAG use cases.

Key benefits:

- Dramatically improve model output by grounding with relevant content
- Fast prototyping of your RAG use cases (e.g. test different LLMs, iterate on document index/collections, change embedding models, set chunk size and overlap, and more)
- Accelerate development by exporting the prompt as a python function deployed as a REST API
- Save prompts and settings as a template for team collaboration.

Learn more in documentation




© 2025 IBM Corporation

# Chat with Images

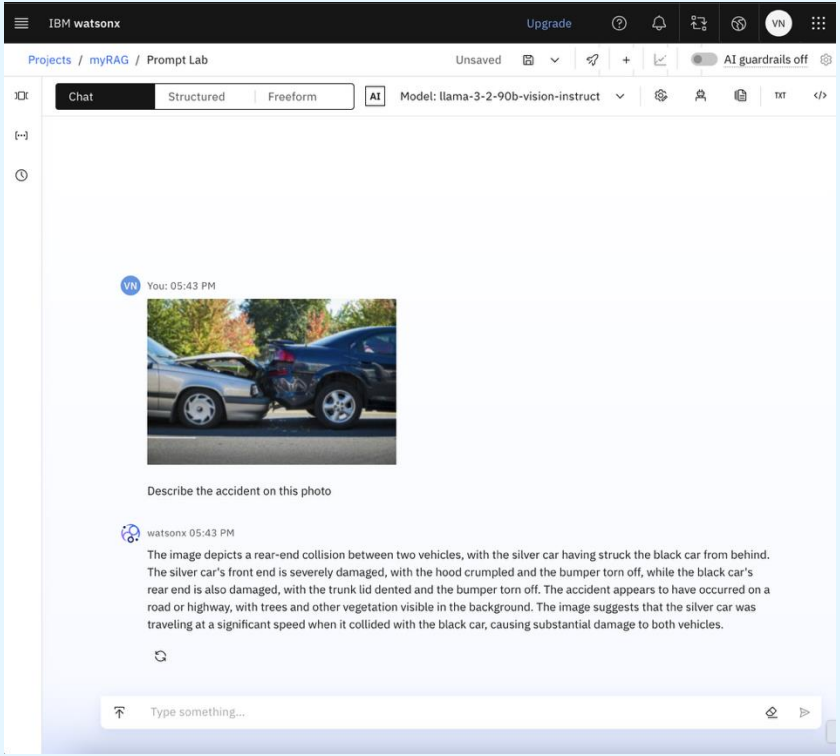
The Chat with Images feature, within the Prompt Lab, provides an easy way to convert visual information into text.

Example use cases:

- Automate the generation of alternative text for images to meet accessibility requirements
- Summarize photos to support use cases such as insurance claims (and many others)
- Convert images from a document into text before the document is used as grounding information for a RAG use case.

Chat with Images is supported by **Llama 3.2** (11b and 90b vision models).

Learn more in the documentation and release blog.




© 2025 IBM Corporation

# RAG Solutions

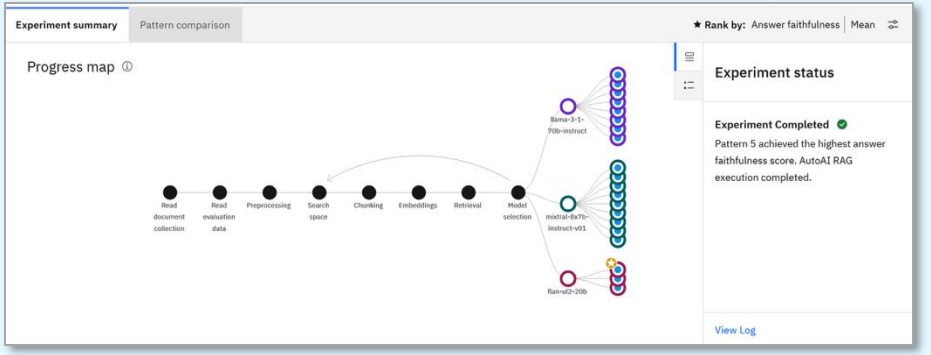
# watsonx.ai AutoAI RAG

RAG prototypes are easy to build but hard to productionalize, which can require a team of experts and months of effort. AutoAI RAG helps to resolve this.

Key benefits:

- Accelerates design and deployment of optimized RAG systems based on client data and use case
- Quickly run various experiments to evaluate a constrained set of configuration options (models, chunk size, and more)
- Re-evaluate and modify recommended configurations when something changes (e.g. a new model version is released, quality of model responses change).

Learn more in documentation.




AutoAI RAG tests a range of parameters (models, chunk size, etc.) in a series of experiments to automatically find the most optimal combination. It has 3 layers:

- Efficient RAG **hyper-parameter optimization** algorithm with end-to-end automation
- Best-of-breed RAG **evaluation metrics** and **benchmarking tools**
- **Parameterized RAG pipelines** for creating embeddings and for retrieval-based inference

© 2025 IBM Corporation

# watsonx.ai Text Extraction API

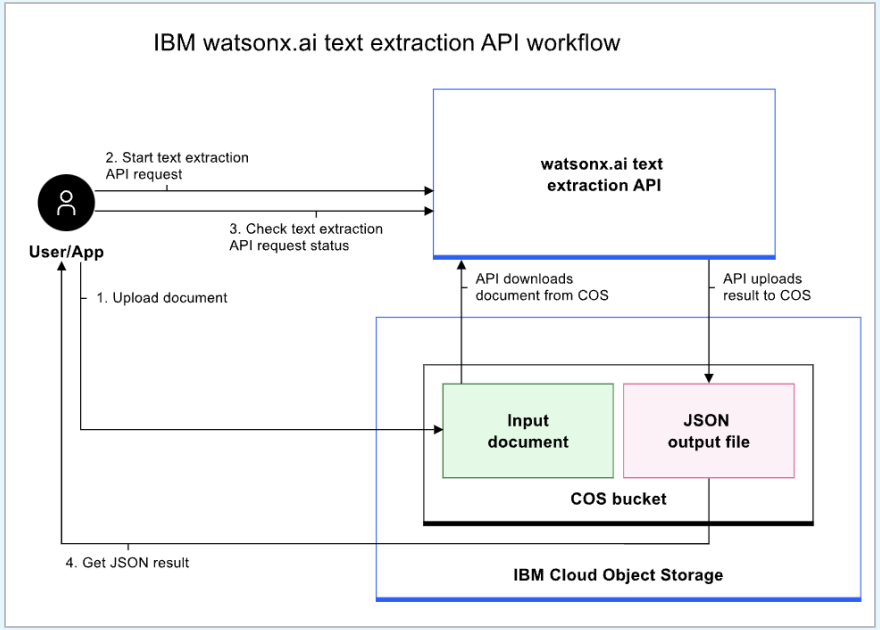
Extract content from documents which is essential for RAG use cases.

- Convert files with tables, diagrams and images into an AI-model friendly JSON file format
- Process the following file types: GIF, JPG, PDF, PNG, TIFF. Including scanned, hand-written documents
- Supports multiple languages

Technology:

- IBM's Natural Language Understanding (NLU) Service. See Watson Document Understanding.
- Optical Character Recognition (OCR) to extract text from images.

Learn more in the documentation.




© 2025 IBM Corporation

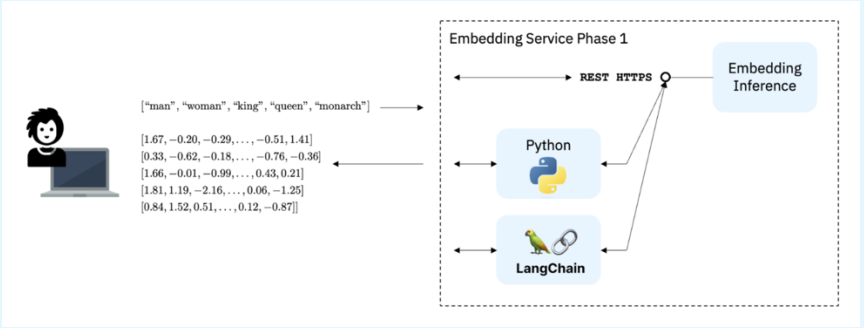
# watsonx.ai Embeddings API

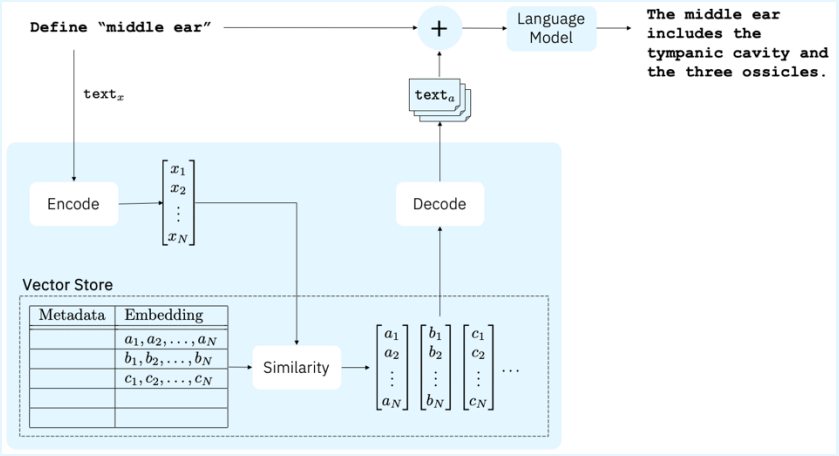
Convert text into dense vector embedding representations. Embeddings capture nuanced semantic and syntactic relationships between words and passages which are then stored in a vector database for retrieval.

Key benefits:

- Support RAG patterns with contextual grounding when utilizing embedding models for query and passage vectorization
- Improve RAG performance with semantically faithful representation of content, especially when compared to basic keyword-based alternatives in classic NLP modeling
- Efficient storage and compute profiles of embeddings make them easily infusible into generative AI applications

Learn more in documentation. See list of supported embedding models.




© 2025 IBM Corporation

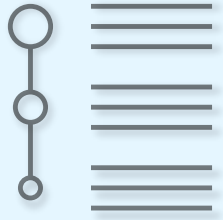
# watsonx.ai Text Rerank API

Add precision to RAG use cases by improving the order of retrieved information with advanced re-ranking algorithms.

The Text Rerank API helps ensure relevant embeddings are used as context within a RAG pipeline, significantly enhancing performance.

- Reorder document passages (from most-to-least likely to answer) based on their similarity to a specified query
- Powered by the ms-marco-minilm-l-12-v2 reranker model (see details)
- Fully integrated with the watsonx.ai REST API

Learn more in documentation.



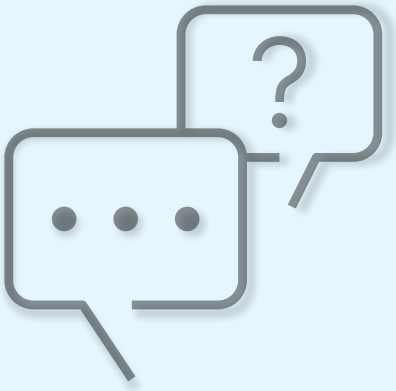
© 2025 IBM Corporation

# watsonx.ai Chat API

Create dynamic and adaptive conversational interfaces including chat-based application that require grounding documents, images, and tool calling for agent-driven applications.

- Build multi-user conversational workflows that use foundation models to generate answers.
- Easily identify different message types, such as a system prompt, user inputs, and foundation model outputs, including user-specific follow-up questions and answers.
- Supports granite, llama and mistral models (see supported list)
- Fully integrated with the watsonx.ai API & SDK

Learn more about adding chat functions and building agents in documentation. See the developer hub for examples.



© 2025 IBM Corporation

# Agents

# Three areas of agentic innovation

## Pre-built agents

**watsonx Orchestrate**
Accelerate AI agent deployment.

Get started quickly with prebuilt AI agents powered with business logic and seamless integration to the tools that power your business.

---

## Custom-built agents

**watsonx.ai**
Build custom-designed agents.

Design, deploy and manage AI agents with ease using pro- and low-code options.

---

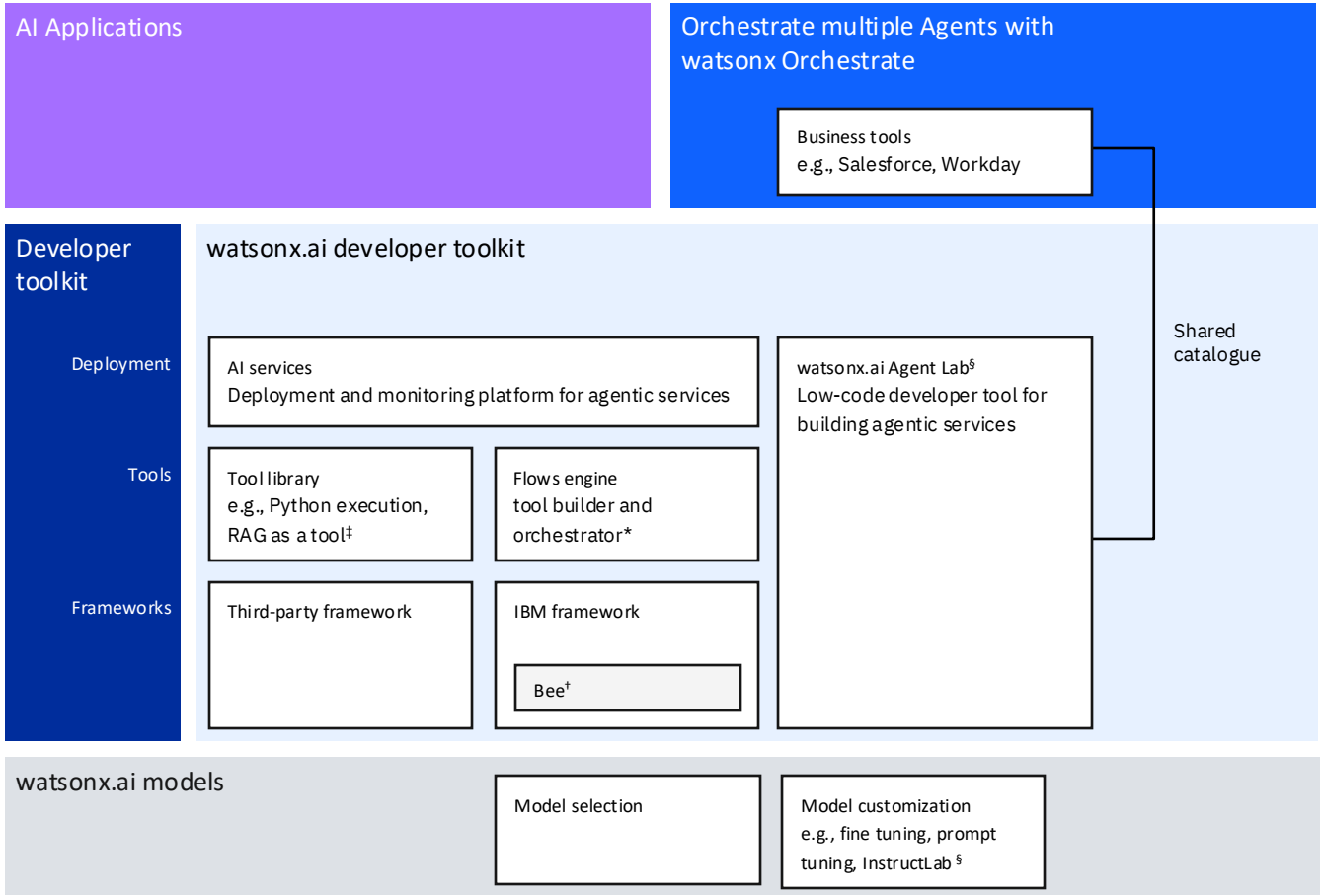
## Multi-agent orchestration

**watsonx Orchestrate**
Manage all agents in one place.

Easily deploy and manage any agent for any task within a simple and unified user experience optimized to scale.

# Architecture
## Custom-built agents

One size does not fit all for building agents



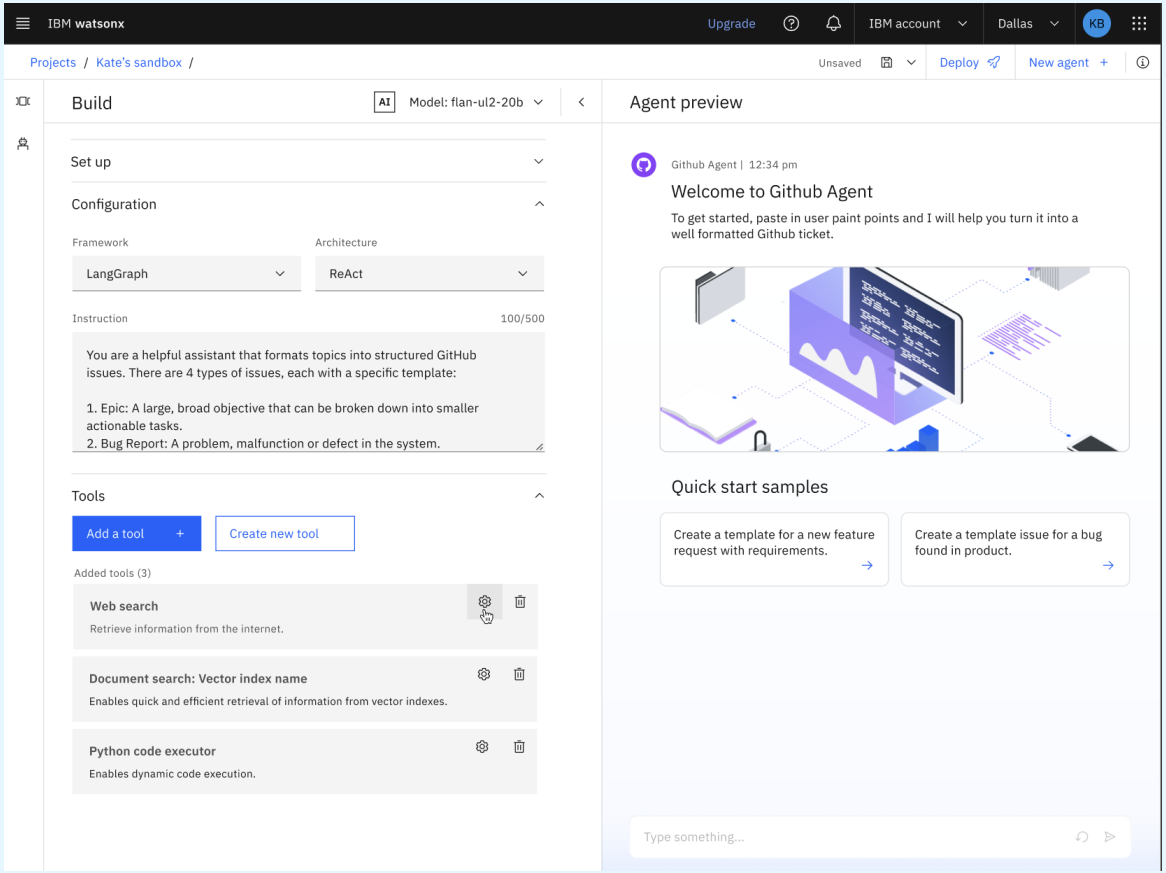

* Stand-alone in tech preview; to be integrated into watsonx.ai
† Experimental, open-source project
‡ In discovery
§ Coming soon

Agent Lab currently in Tech Preview

# watsonx.ai Agents

What's available today?

- Support for Chat API, Tool Calling, JSON returns together with providing the foundation for agentic support
- AI Services that provide the foundation for deployment of python based agentic solutions
- End-to-end template for the development and deployment of a custom Langraph agents
- Introduction of Agent Lab (Beta), enabling a low-code experience to build agentic solution with 1-click deployment as a production-ready API endpoint



# watsonx.ai SDKs

**Build agentic services** using popular open-source agentic frameworks through industry-standard API and SDK support.

© 2025 IBM Corporation

# Lightweight Engine

Software only

# watsonx.ai Lightweight Engine

Reduced footprint, inference-only installation of watsonx.ai for client-managed software

## Foundation Models

Access a variety of fit-for-purpose foundation models: IBM, open source, third party and imported custom foundation models.

## Deploy Anywhere

Run watsonx.ai Lightweight Engine in any environment.

## Reduced footprint

Run watsonx.ai Lightweight Engine with a smaller footprint.

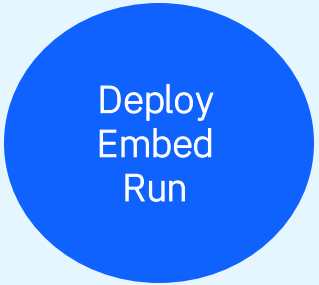
## Developer Experience

Leverage enterprise-grade APIs for a consistent developer experience across watsonx.ai full service and the lightweight engine.

watsonx.ai

Full service, enterprise-grade AI toolkit for building and deploying end-to-end AI solutions.

watsonx.ai lightweight engine




lightweight engine to embed into external gen AI applications

© 2025 IBM Corporation

# watsonx.ai Lightweight Engine Features

## Included

**Features**

- ✓ Foundation models (IBM, third-party, open-source)
- ✓ Imported custom foundation models
- ✓ watsonx.ai REST API
- ✓ watsonx.ai Text Embeddings API
- ✓ watsonx guardrails (HAP/PII filters)

**Enterprise Capabilities**

- ✓ Automated install
- ✓ Horizontal Scaling
- ✓ Backup and restore
- ✓ User Authentication
- ✓ Logging
- ✓ Certification Manager

## Excluded, watsonx.ai full-service only

**Features**

- ✗ Tuning Studio
- ✗ Prompt Lab
- ✗ Instruct Lab
- ✗ Prompt Tuning via watsonx.ai REST API
- ✗ Fine-tuning via watsonx.ai REST API [1]
- ✗ Chat with Documents and Images

**Enterprise Capabilities**

- ✗ Projects and Spaces
- ✗ Access Control

1 – on watsonx.ai full-service roadmap

© 2025 IBM Corporation

# Analyzing data and working with models

# watsonx.ai Data Science and MLOps

Wide selection of integrated tooling to quickly build, deploy and manage machine learning models and AI applications

## Build models

Comprehensive tooling for Model training and development

- Low/Pro code data science tools in a single environment
- Auto AI automates several aspects of the model lifecycle
- open-source libraries and integrated notebook-based interfaces
- Support for popular languages such as Python and R
- Integrated tooling for data pipelines and data preparation

## Deploy models

Support for model deployment and management

- One-click deployment
- Continually deliver models into production
- Automated version control, retraining, code snippet generation for developers
- Build on OpenShift for reliability

## Monitor models with watsonx.governance

Automate governance to drive trust across the AI lifecycle

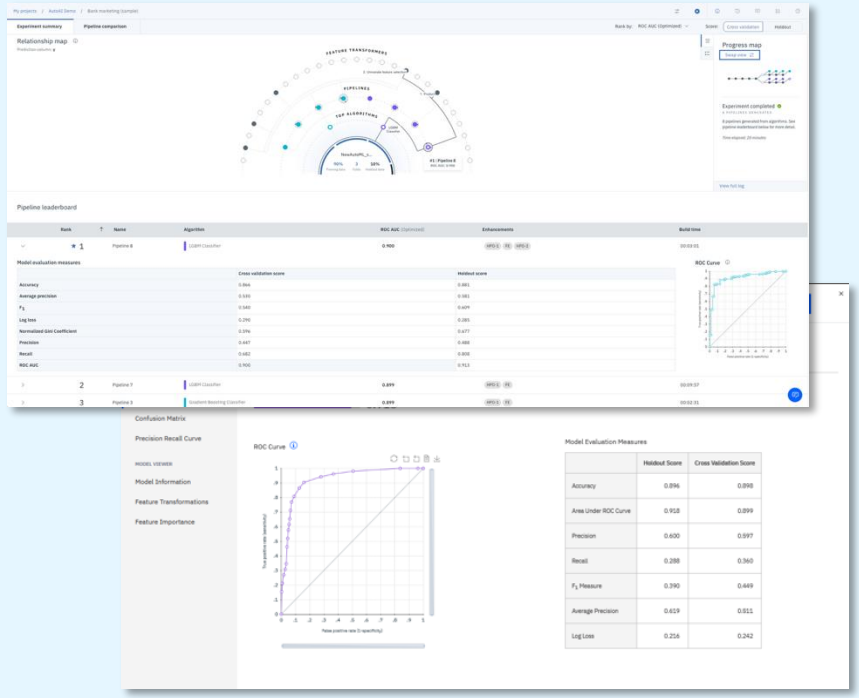
- Monitor drift over time, automatic retraining as required
- Provide model verification to stakeholders to ensure fairness, transparency and explainability over time
- Inform stakeholders of accuracy shifts before an issue occurs

# watsonx.ai AutoAI

Expedite the creation of machine learning models with AutoAI, an easy to use, no-code interface. With a few clicks, quickly build powerful models such as classification, regression, time series forecasting, and more.

- Enhance the efficiency of the limited resources in a Data Science team by automating modeling.
- Quickly identify the best features, transforms, and models, as well as train the models, tune the hyperparameters, and rank the best-performing pipelines.
- Save an experiment as a notebook to understand the transformations applied to build the model or make changes.
- One click deployment of models, online or in batch mode

Learn more in documentation




© 2025 IBM Corporation

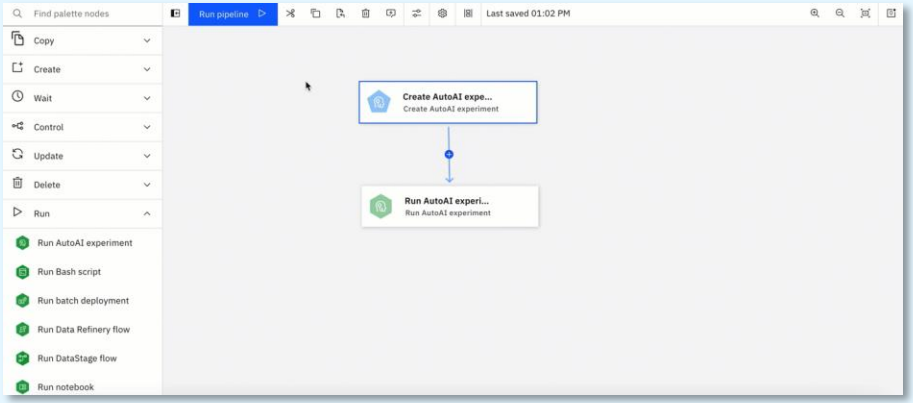
# watsonx.ai Pipelines

Create end-to-end flow of assets from creation through deployment on a graphical canvas.

Key benefits:

- Increase efficiency and time savings with automated pipelines
- Combine tasks from different tools
  - Notebooks
  - Data refinery
  - Model creation
  - Model deployments
- Automate model training with new data sets
- Automatically deploy better challenger models

Learn more in documentation




© 2025 IBM Corporation

# watsonx.ai Synthetic Data Generator

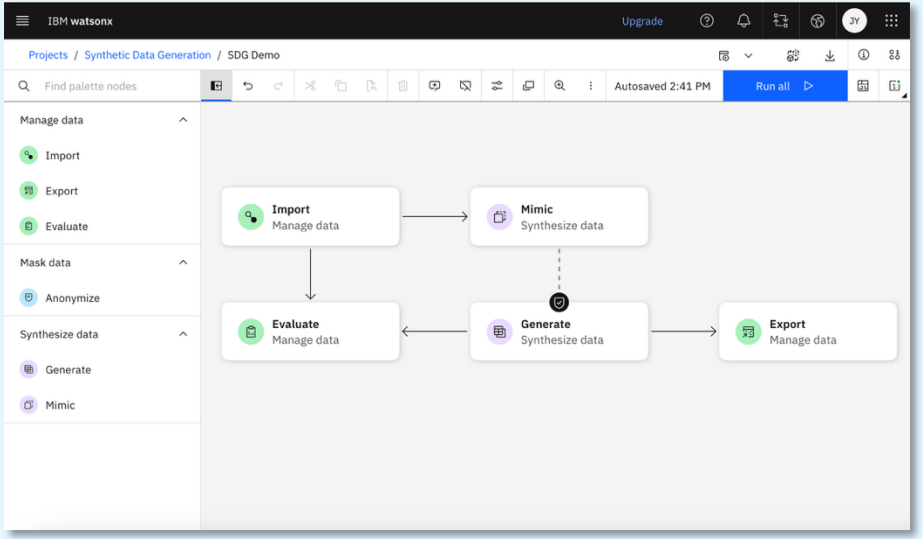
Generate synthetic tabular data to address your data gaps

## Create synthetic data at scale

- Unlock your valuable insights by using synthetic data.
- Create synthetic data using your existing data in a database or by uploading a file. Optionally, design your own data schema.
- Address data gaps and create synthetic edge cases to expedite classical AI model training.

## Select your model & privacy needs

- Depending on your cost, fidelity, application, or data needs, you can select from multiple IBM models* to create your synthetic tabular data.
- When using existing data, IBM models apply differential privacy to minimize your privacy risk and give you control over the level of privacy protection required for your organization.




Learn more in documentation

© 2025 IBM Corporation

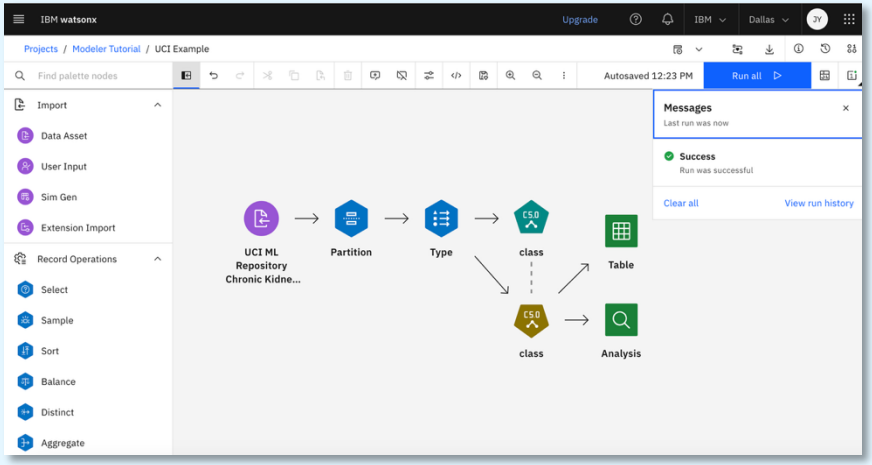
# watsonx.ai Visual Modeling

The integrated SPSS Modeler brings a comprehensive set of nodes for data preparation, machine learning including auto modeling, data visualization, and push to production options.

Key benefits:

- Quickly build machine learning models using powerful drag and drop data analysis and visualization
- Drive direct insights from raw texts, uncover hidden value and patterns from text for prediction
- Deploy SPSS streams to into production

Learn more in documentation




© 2025 IBM Corporation

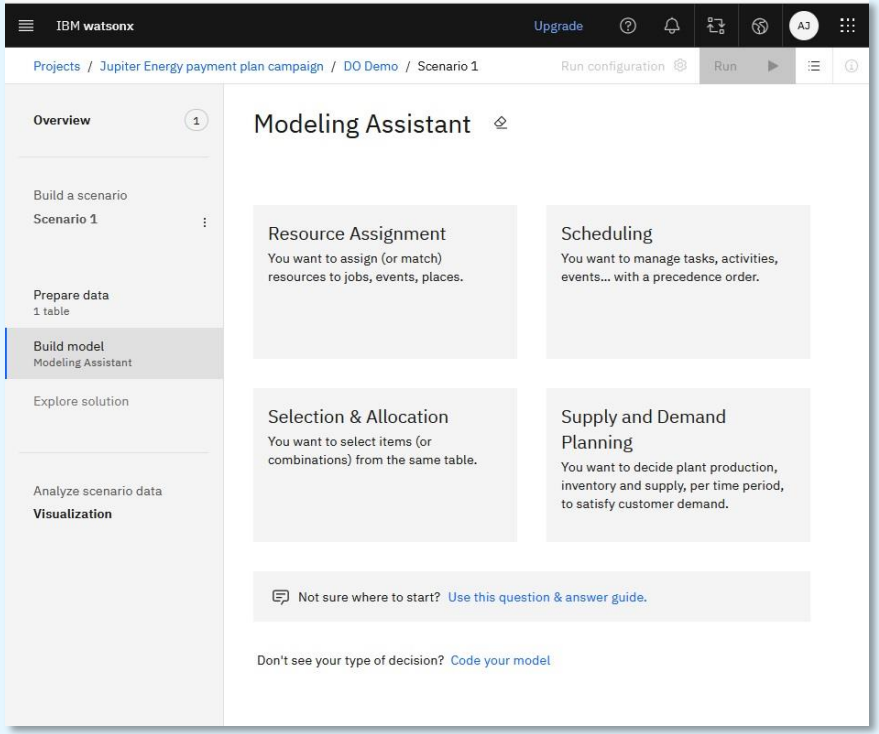
# watsonx.ai Decision Optimization

Leverage the power of prescriptive analytics with Decision Optimization (DO). Evaluate a problem holistically, identify constraints, and determine the best plan to accomplish your goal (e.g. max profit, min costs, optimal scheduling, etc.)

Key benefits:

- Solve optimization models by using DO's highly powerful mathematical and constraint programming engines
- Improve productivity by validating optimization models more quickly and easily using visual dashboards
- Create innovative solutions by combining optimization technology with data science techniques such as machine learning
- Operationalize your projects by deploying optimization models in production to drive business impact around real use cases

Learn more in documentation





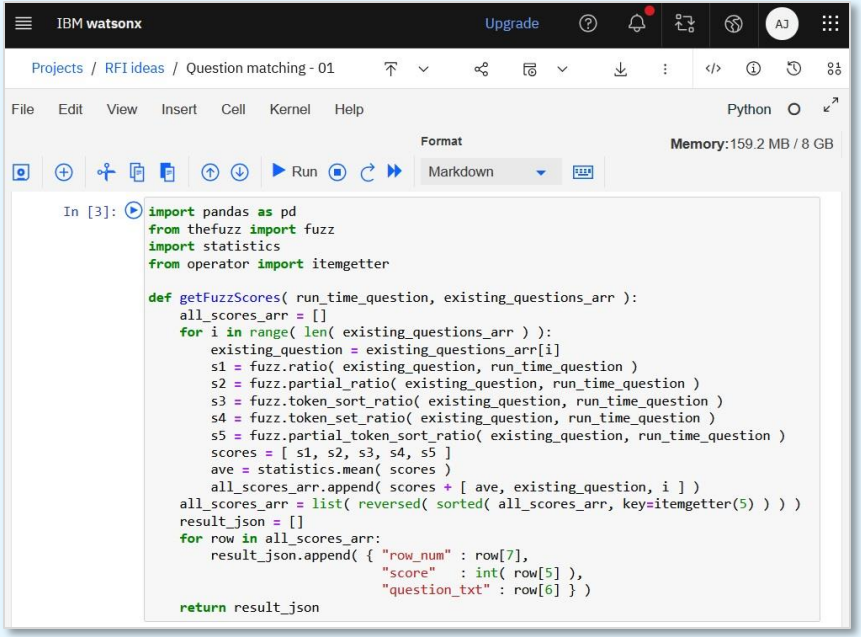
© 2025 IBM Corporation

# watsonx.ai Pro Code Tooling

Create, edit and execute Python and R code using Jupyter notebooks and scripts in code editors, for example the notebook editor or an the Rstudio integrated development environment (IDE)

Key benefits:

- Preinstalled IBM and opensource libraries
- Import your own libraries
- Connect notebooks to pipelines, publish to a catalog or Github for sharing and collaboration
- Automate repeatable processes through jobs to run the notebooks




Learn more in documentation


© 2025 IBM Corporation

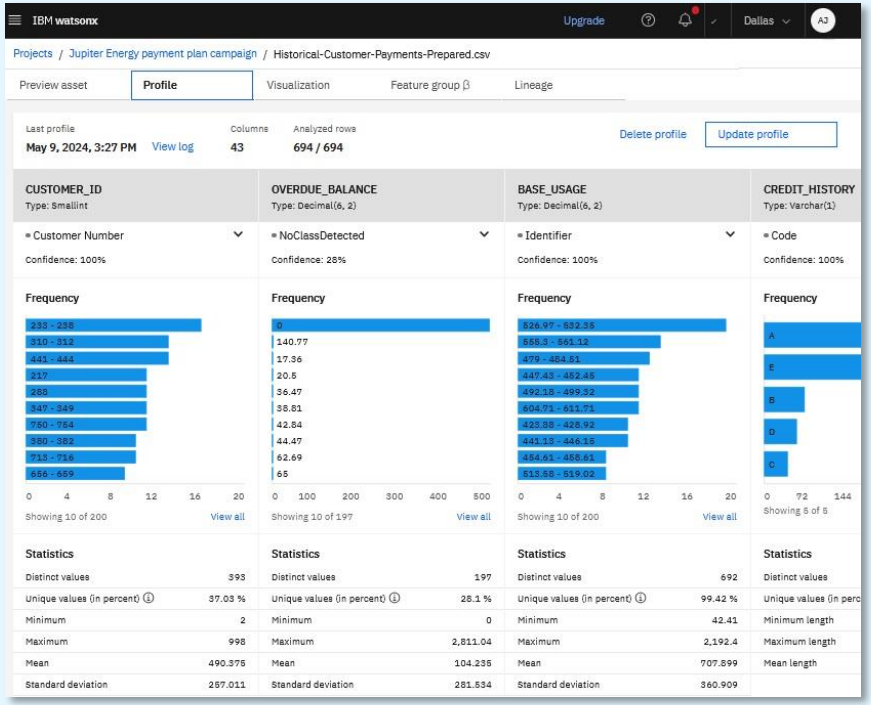
# watsonx.ai Data Refinery

Conveniently analyze and prepare data for model training and all your AI application use cases within a single environment.

Key benefits:

- Retrieve or view tabular data from over 60 built-in data connectors
- Quickly analyze and visualize data with built-in data profiling and charting tools
- Cleanse and shape data with graphical editing with over 100 built-in operations
- Save time with scheduled flows and jobs for repeatable operations

Learn more in documentation





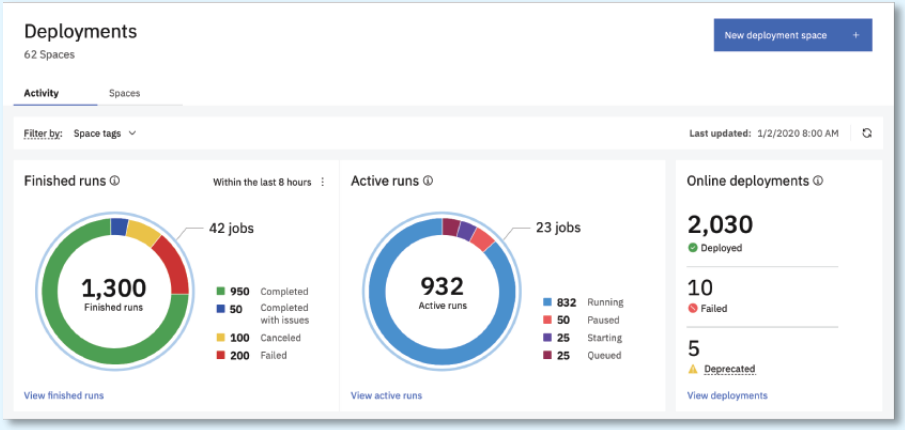
© 2025 IBM Corporation

# watsonx.ai Model Deployment

Expedite deployment of machine learning models, scripts, functions, and prompt templates for generative AI models.

- Secured with authentication credentials
- Support popular frameworks
- Deploy models in online or batch mode
- Score data through REST API calls
- Automate deployments with jobs

Learn more in documentation

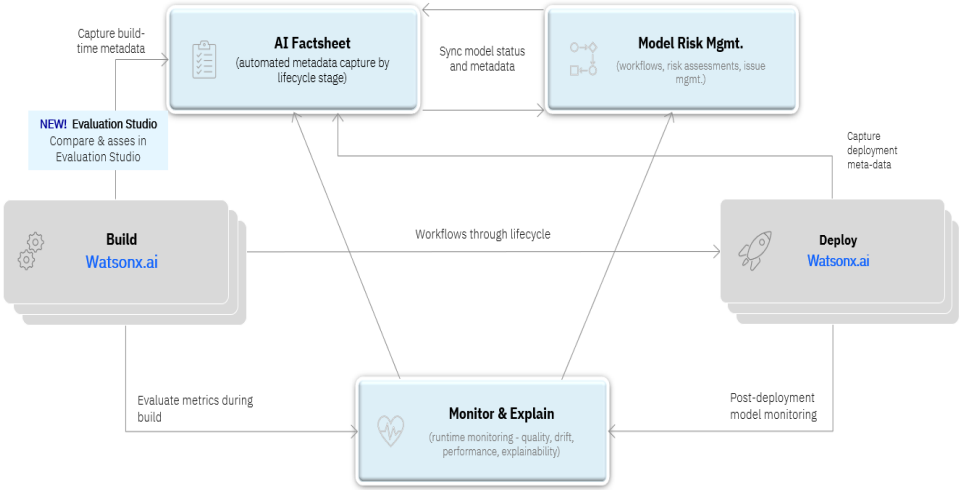



© 2025 IBM Corporation

# watsonx.governance and watsonx.ai

Watsonx.governance brings **observability**, **explainability**, and **risk management** - key elements to build and deploy LLM applications and ML models.

- **Easily track and organize** AI Assets in Projects and Spaces by use case and associated lifecycle stage
- Enable AI Engineers to **easily determine the right Prompt asset** through experiments in Evaluation Studio
- **Evaluate quality and performance** of multiple task types (summarization, context generation, Q&A and more) during build with out-of-the-box metrics and integrated experience in Prompt Lab
- **Runtime monitoring** for quality, drift, and performance (and bias for ML models); user is alerted when thresholds are breached
- **RAG Root Cause Analysis insights** help AI engineers understand LLM responses
- **Assess hallucination** through quantifiable metrics for RAG task-type; **easily see** the phrases in **context used to generate responses**
- Reduce overhead of manual documentation through **automated documentation** of Prompt Template metadata and training input from the Prompt Lab

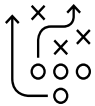



... and more!

- Out-of-the-box workflows through lifecycle
- Issue management integrated with use case
- Use case risk assessment
- Regulatory applicability assessments
- View AI use cases and deployments use across enterprise

# Four ways to get started with watsonx.ai today

## Looking forward to today's discussions



### Free trial

Get started quickly by accessing AI models, data, and tools to build AI applications.

### watsonx Developer Hub

Access a library of templates and guides to get started building gen AI applications.

### Request a client briefing or demo

Discussion and custom demonstration of IBM's generative AI watsonx point of view and capabilities. Understand where generative AI can be leveraged now for impact in your business.

2-4 hours onsite or virtual

### 5-year business value assessment

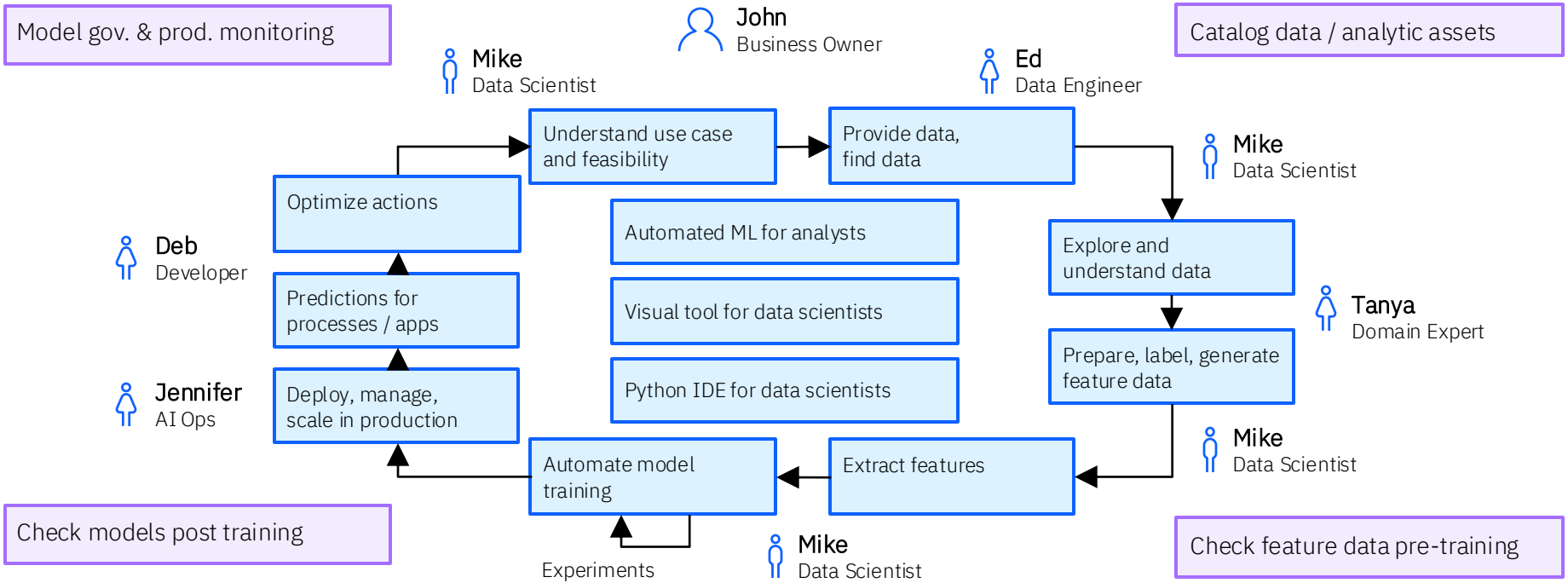
Engagement with an IBM multi-disciplinary team to jointly innovate and rapidly prove the business value of generative AI solutions using watsonx.

1-4 weeks

© 2024 IBM Corporation

# Backup

# Data science requires a team



# Model Customization Methods

| Criteria | Prompt tuning | Parameter Efficient fine tuning (PEFT) | Full fine tuning | InstructLab |
|---|---|---|---|---|
| Technique | Prompts or embeddings are tuned while model parameters remain fixed. | Adapters, LoRA, QLoRA - Only a subset of parameters (e.g., adapters, biases) are fine-tuned. | SFTTrainer - All model parameters are fine-tuned on the target task. | Multi-phase aligned full fine tuning. |
| Test data reusability | Test data should not be reused; prompts are central | Test data should not be reused; prompts can guide which parameters are fine-tuned and can be reused. | Test data should not be reused; prompts can be incorporated into training data but may become less relevant as the model learns. | Reusable taxonomy across models and their versions |
| Model performance | Moderate performance; works well when the task is similar to the pre-trained knowledge. | High performance with reduced risk of overfitting; may not reach full fine-tuning levels. | Higher performance with sufficient data and compute resources. | Potentially the highest performance with less reliance on human data |
| Compute resources | Low; minimal compute required as only prompts or a small number of parameters are tuned. | Moderate; requires less compute than full fine-tuning due to selective parameter updates. | High computational and memory requirements due to the full update of model parameters. | High computational and memory requirements due to the full update of model parameters. |
| Time taken for tuning | Short; typically, the quickest tuning process as only prompts are optimized. | Moderate; faster than full fine-tuning but still requires time for training the selected parameters. | Long; depends on model size and dataset but generally requires significant time. | Long; depends on model size and synthetic data but generally requires significant time. |
| Summary | Most suitable for quick adaptation tasks, especially when computational resources are limited, or the task is closely related to the pre-trained model. | Ideal for resource-constrained environments or when needing to fine-tune multiple tasks efficiently. | Best for scenarios where maximum accuracy and task-specific adaptation are critical, and resources are abundant. | Much more than just a tuning technique. Doesn't need deep technical expertise and data prep. Provides a collaborative platform, synthetic data generation and better accuracy of models |

© 2025 IBM Corporation