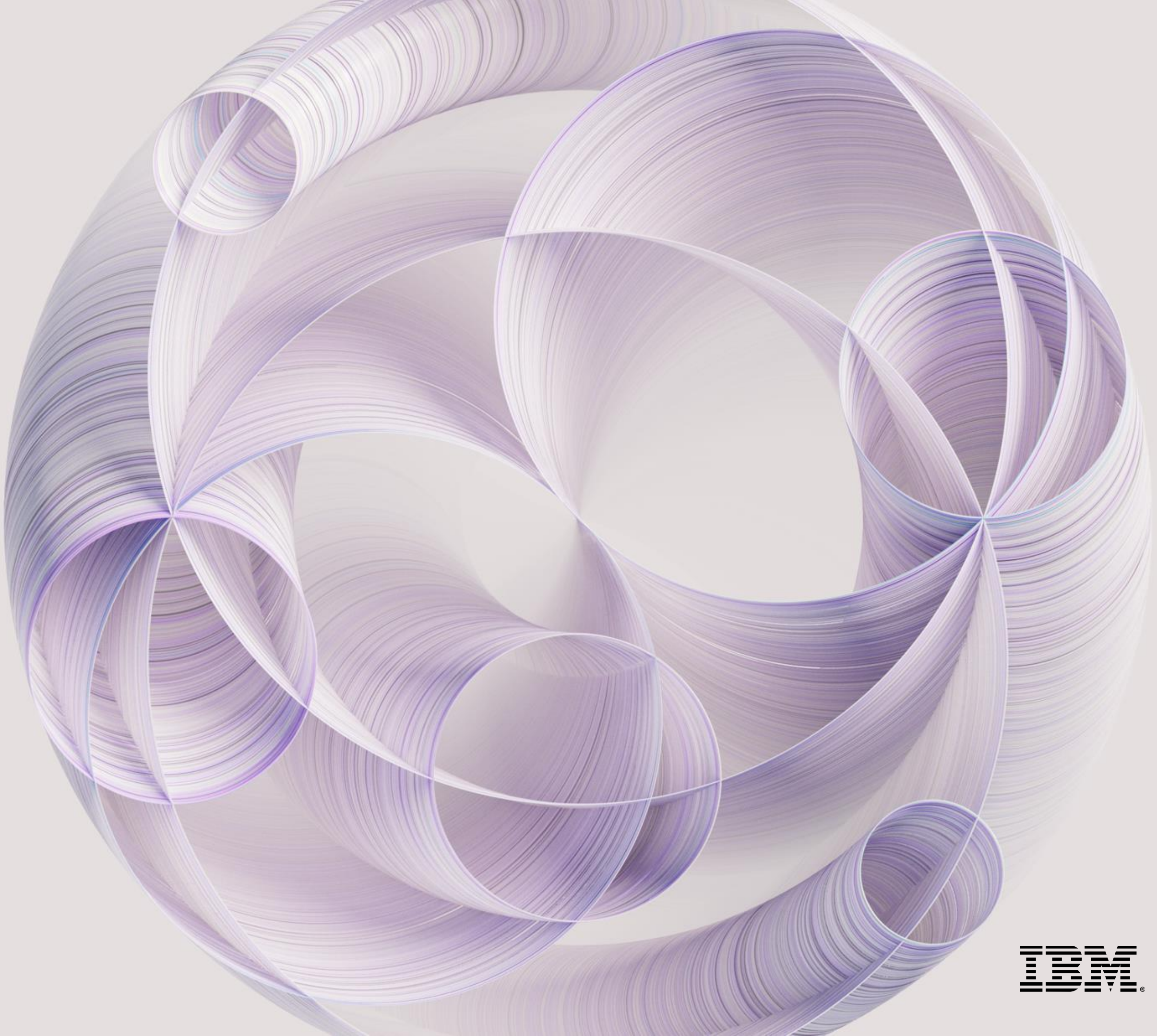
Onboarding Workshop

watsonx

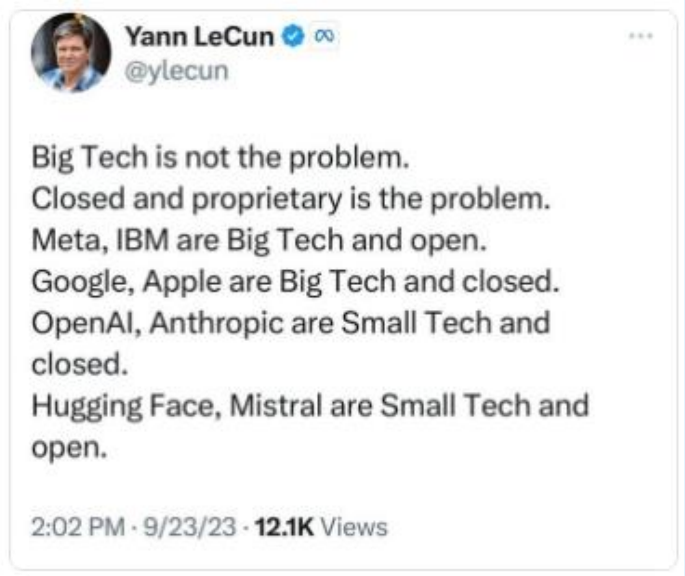


IBM

Agenda

- 12:00 PM - 12:10 PM Introductions
- 12:10 PM - 12:30 PM Account Onboarding / Logistics
- 12:30 PM - 1:00 PM watsonx Overview
- 1:00 PM - 2:00 PM watsonx.ai Labs

Lab Materials



Yann LeCun
@ylecun

Big Tech is not the problem.
Closed and proprietary is the problem.
Meta, IBM are Big Tech and open.
Google, Apple are Big Tech and closed.
OpenAI, Anthropic are Small Tech and closed.
Hugging Face, Mistral are Small Tech and open.

2:02 PM · 9/23/23 · **12.1K** Views

The platform for AI and data

watsonx

Scale and accelerate the impact of AI across your business

watsonx.ai

Build, train, validate, tune and deploy AI models

A next generation enterprise studio for AI builders to build, train, validate, tune, and deploy both traditional machine learning and new generative AI capabilities powered by foundation models. It enables you to build AI applications in a fraction of the time with a fraction of the data.

watsonx.data

Scale AI workloads, for all your data, anywhere

Fit-for-purpose data store, built on an open lakehouse architecture, supported by querying, governance and open data formats to access and share data.

watsonx.governance

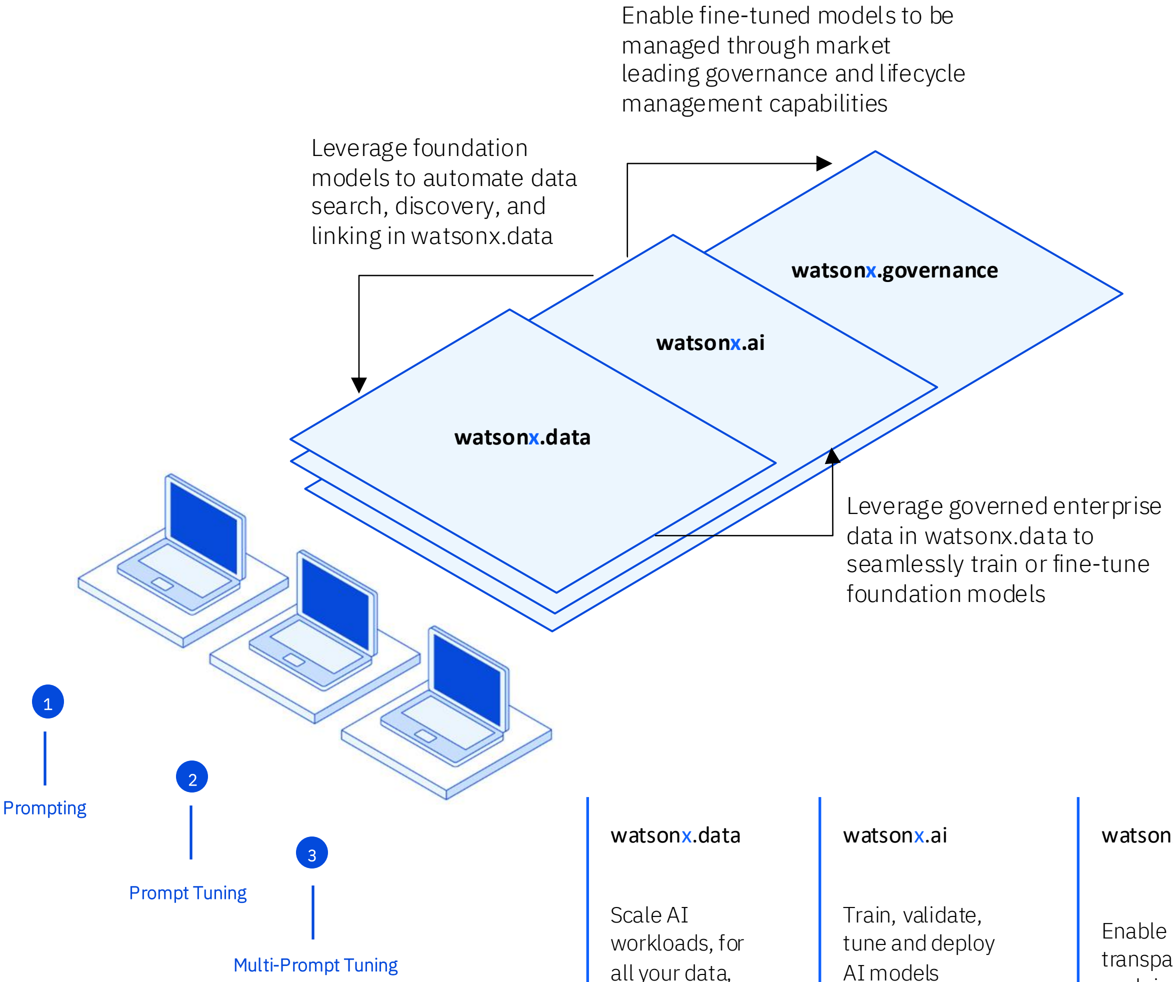
Accelerate responsible, transparent and explainable AI workflows

End-to-end toolkit for AI governance across the entire model lifecycle to accelerate responsible, transparent, and explainable AI workflows

The platform for AI and data |

watsonx

Scale and accelerate the impact of AI with trusted data.



Enable fine-tuned models to be managed through market leading governance and lifecycle management capabilities

Leverage foundation models to automate data search, discovery, and linking in watsonx.data

Leverage governed enterprise data in watsonx.data to seamlessly train or fine-tune foundation models

watsonx.data

Scale AI workloads, for all your data, anywhere

watsonx.ai

Train, validate, tune and deploy AI models

watsonx.governance

Enable responsible, transparent and explainable AI workflows

What IBM offers

AI assistants

watsonx

watsonx Orchestrate

Harness the power of AI and automation to free up individuals from tedious tasks

Enable employees to quickly offload time-consuming work to tackle more of the work only they can do. Business users can delegate common and complex tasks such as creating a job description, pulling a report in Salesforce or SAP SuccessFactors, sourcing candidates, and more using natural language.

40%

improvement in HR productivity[1]

watsonx Assistant

Build better virtual agents, to deliver consistent and intelligent customer care

Understand customers in the right context, and provide fast, consistent, and accurate answers, and self-service support across any application, device, or channel. The intuitive build experience empowers everyone in the organization to build and deploy AI-powered virtual agents without writing a line of code.

>90%

customer inquiries handled by AI assistant[2]

watsonx Code Assistant

Accelerate development, application modernization, and assist with IT Operations

Increase developer productivity, reduce coding complexity, and accelerate developer onboarding. Purpose-built for targeted use cases, watsonx Code Assistant uses AI to support application modernization and IT automation.

60%

software development content automatically generated by AI[3]

[1] IBM HR use case
[2] Vodafone Case Study in partnership with IBM and Genesys
[3] IBM CIO case study based on limited internal test

IBM’s generative AI technology and expertise

AI assistants	Empower individuals to do work without expert knowledge across a variety of business processes and applications.	**watsonx** Code Assistant **watsonx** Assistant **watsonx** Orchestrate **watsonx** Orders
SDKs & APIs	Embed watsonx platform in third party assistants and applications using programmatic interfaces.	**Ecosystem integrations**
AI & data platform	Leverage generative AI and machine learning — tuned with your data — with responsibility, transparency and explainability.	**watsonx** watsonx.ai watsonx.governance watsonx.data **Foundation models** Granite | *IBM* Open Source | *Hugging Face* Llama 2 | *Meta* Geospatial | *IBM + NASA* …
Data services	Define, organize, manage, and deliver trusted data to train and tune AI models with data fabric services.	**Cloud Pak for Data** **watsonx** Discovery
Hybrid cloud AI tools	Build on a consistent, scalable, foundation based on open-source technology.	**Red Hat** OpenShift AI (*e.g.*, Ray, Pytorch)

Consulting
Generative AI strategy, experience, technology, operations

Ecosystem
System Integrators, Software and SaaS partners, Public Cloud providers

watsonx

Model strategy →

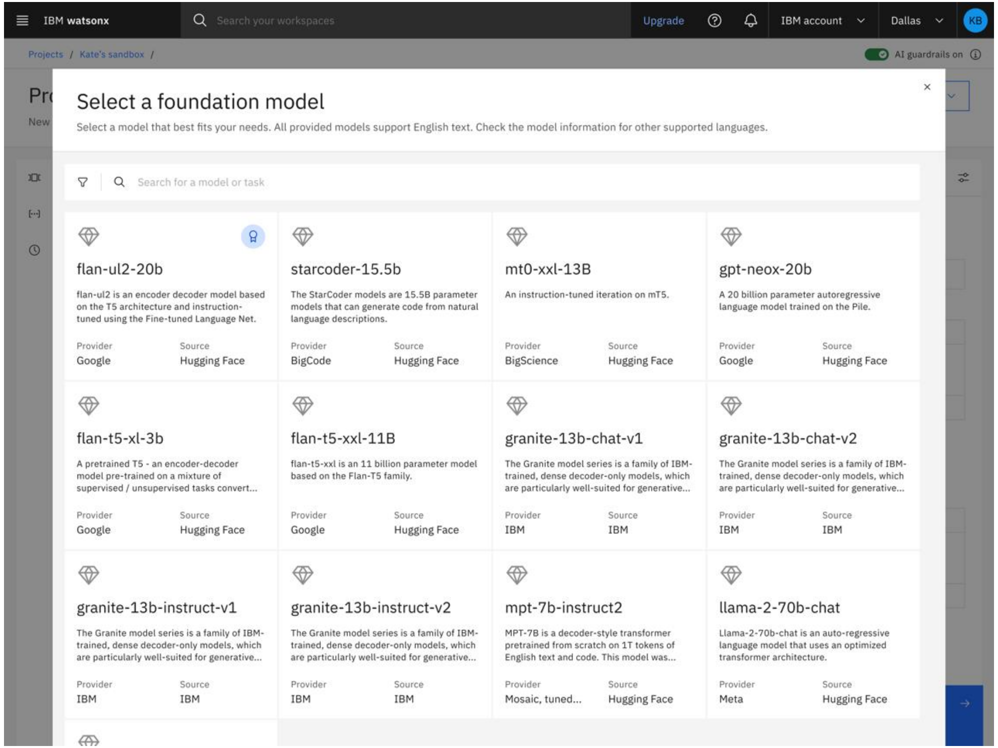
Multi-model

One model doesn’t fit all use cases. We offer IBM-developed, open-source, third party, and BYOM.

Bigger is not always better. Specialized models can outperform general-purpose models with lower infrastructure requirements.

Hybrid, multi-cloud

Hybrid deployments. We provide the flexibility to deploy models on the platform of choice.



Select a foundation model

Select a model that best fits your needs. All provided models support English text. Check the model information for other supported languages.

flan-ul2-20b
flan-ul2 is an encoder decoder model based on the T5 architecture and instruction-tuned using the Fine-tuned Language Net.
Provider: Google — Source: Hugging Face

starcoder-15.5b
The StarCoder models are 15.5B parameter models that can generate code from natural language descriptions.
Provider: BigCode — Source: Hugging Face

mt0-xxl-13B
An instruction-tuned iteration on mT5.
Provider: BigScience — Source: Hugging Face

gpt-neox-20b
A 20 billion parameter autoregressive language model trained on the Pile.
Provider: Google — Source: Hugging Face

flan-t5-xl-3b
A pretrained T5 - an encoder-decoder model pre-trained on a mixture of supervised / unsupervised tasks convert...
Provider: Google — Source: Hugging Face

flan-t5-xxl-11B
flan-t5-xxl is an 11 billion parameter model based on the Flan-T5 family.
Provider: Google — Source: Hugging Face

granite-13b-chat-v1
The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for generative...
Provider: IBM — Source: IBM

granite-13b-chat-v2
The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for generative...
Provider: IBM — Source: IBM

granite-13b-instruct-v1
The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for generative...
Provider: IBM — Source: IBM

granite-13b-instruct-v2
The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for generative...
Provider: IBM — Source: IBM

mpt-7b-instruct2
MPT-7B is a decoder-style transformer pretrained from scratch on 1T tokens of English text and code. This model was...
Provider: Mosaic, tuned... — Source: Hugging Face

llama-2-70b-chat
Llama-2-70b-chat is an auto-regressive language model that uses an optimized transformer architecture.
Provider: Meta — Source: Hugging Face

granite.20b.code is delivered through watsonx Code Assistant

What is IBM Granite ?

- Granite is IBM's flagship series of LLM foundation models based on decoder-only transformer architecture.
- Granite language models are trained on trusted enterprise data spanning internet, academic, code, legal and finance.

By 2027 more than 50% of the Gen AI models that enterprises use will be domain-specific — specific to either an industry or business function — up from approximately 1% in 2023.

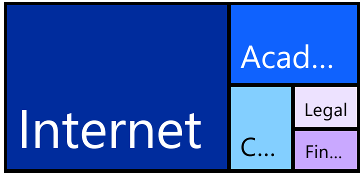
Gartner Report, Predicts 2024: The Future of Generative AI Technologies.

Trusted, Performant, Cost-effective AI foundation models purpose built for enterprises.

granite-13b-v2 (English LLM)

-chat-v2.1, -instruct-v2

13B parameters in size
2.5T tokens of data

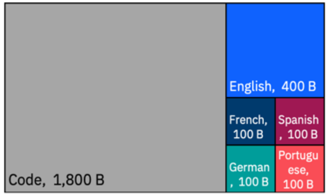


(v1 breakdown)

- Chat derivative model is optimized for dialogue use cases and works well with virtual agent and chat applications.
- Instruct derivative model was designed to perform well on natural language tasks and can be customized for specific industries and domains via prompt-tuning.

granite-20b-multilingual

20B parameters in size
2.6 T tokens of Data



granite-7b-lab

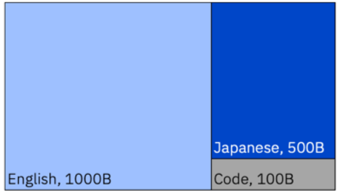
Open-source

7B parameters in size

- Tuned using IBM's large-scale alignment of chatbots(LAB).

granite-8b-japanese

8B parameters in size
1.6T tokens of Data

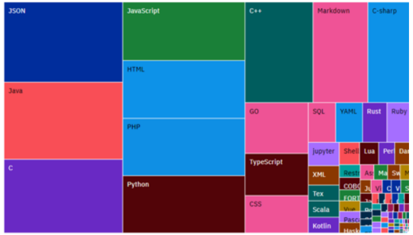


granite-code

Open-source

3B, 8B, 20B, 34B parameters in size

- A family of models trained in 116 programming languages



Granite-code-20b

IBM open-source models : https://huggingface.co/ibm-granite

watsonx

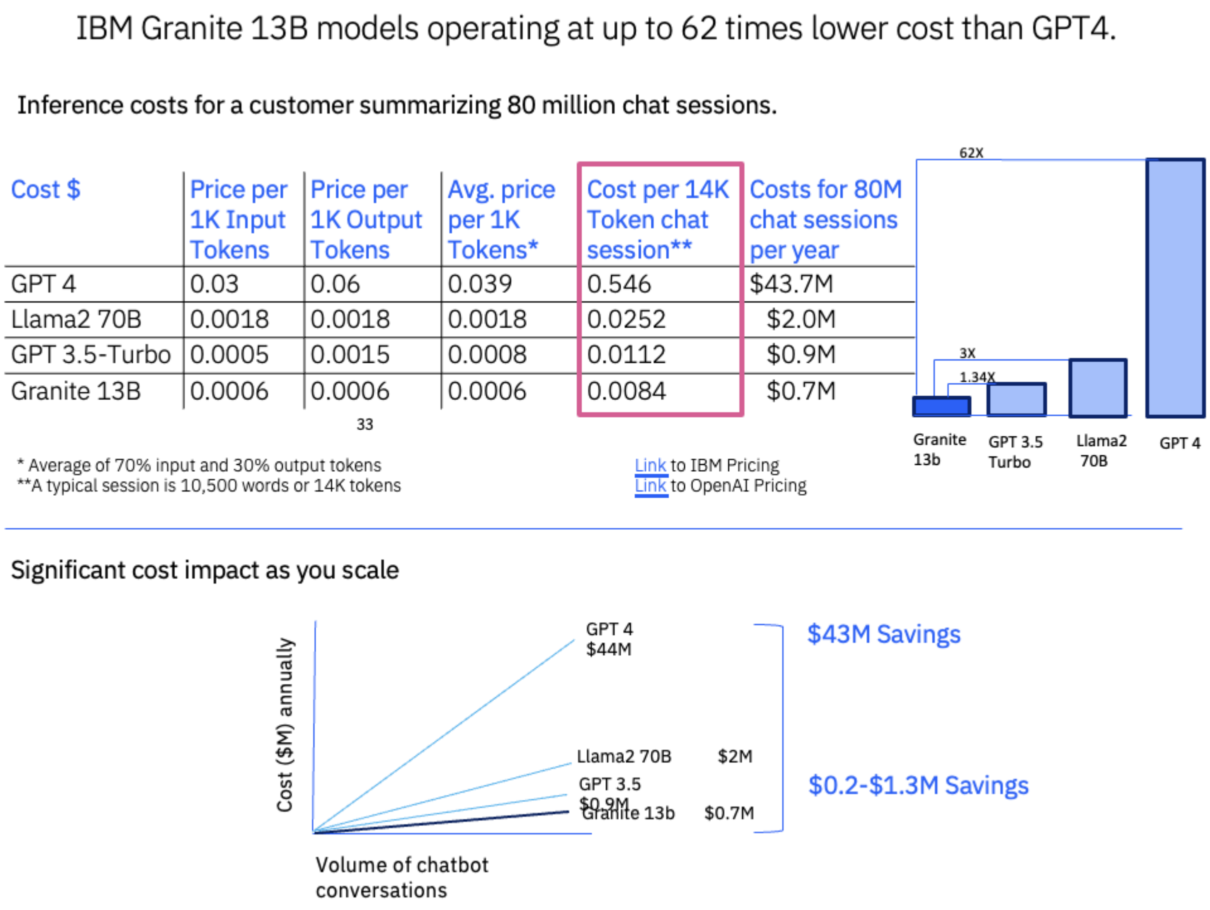
Operating Costs →

Inferencing Costs (aaS)

Price Differentiation: Model costs per token in managed infrastructure can vary significantly by model type and service provider.

Infrastructure Scale

Capacity and Performance. Model performance can also differ greatly in infrastructure requirements and speed of inferencing.



IBM Granite 13B models operating at up to 62 times lower cost than GPT4.

Inference costs for a customer summarizing 80 million chat sessions.

Cost $	Price per 1K Input Tokens	Price per 1K Output Tokens	Avg. price per 1K Tokens*	Cost per 14K Token chat session**	Costs for 80M chat sessions per year
GPT 4	0.03	0.06	0.039	0.546	$43.7M
Llama2 70B	0.0018	0.0018	0.0018	0.0252	$2.0M
GPT 3.5-Turbo	0.0005	0.0015	0.0008	0.0112	$0.9M
Granite 13B	0.0006	0.0006	0.0006	0.0084	$0.7M

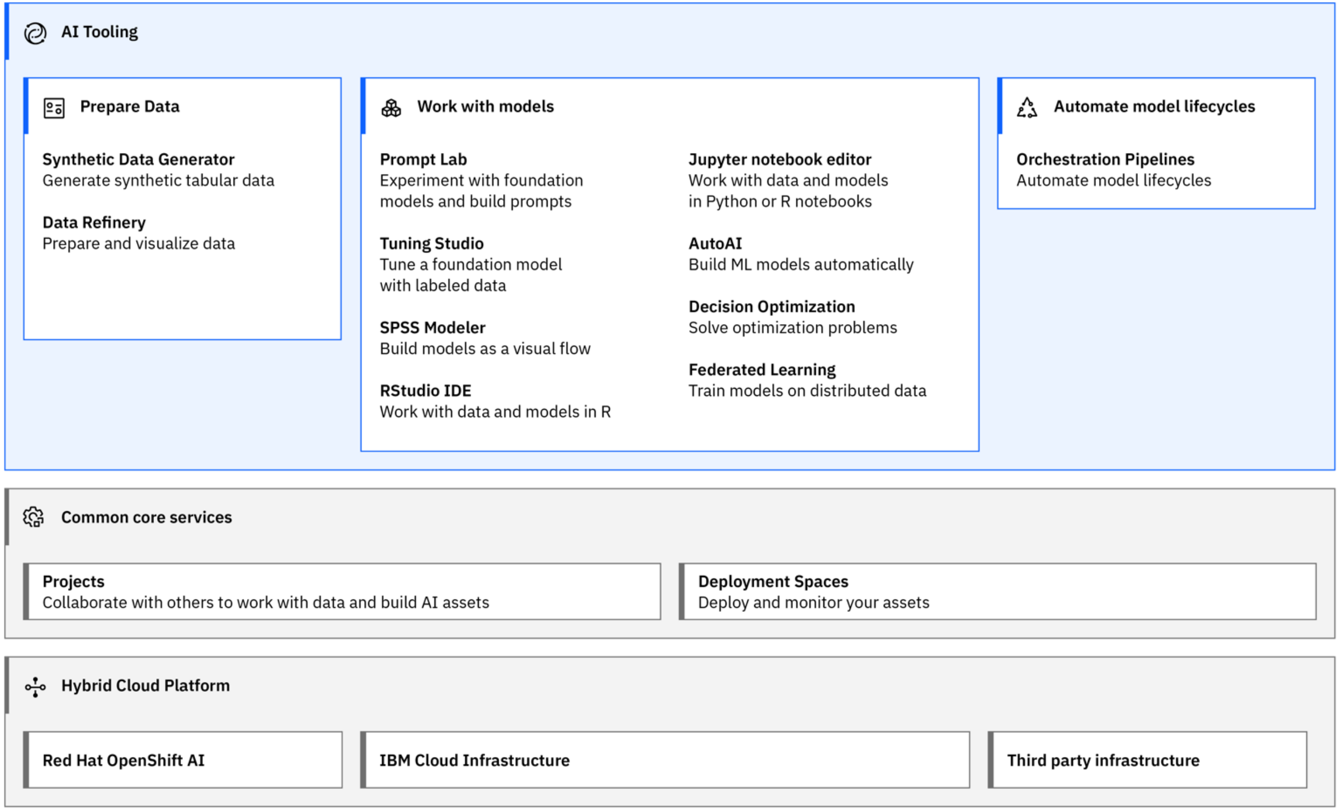
33

* Average of 70% input and 30% output tokens
**A typical session is 10,500 words or 14K tokens

Link to IBM Pricing
Link to OpenAI Pricing

Significant cost impact as you scale

IBM watsonx.ai architecture



AI Tooling

Prepare Data

Synthetic Data Generator
Generate synthetic tabular data

Data Refinery
Prepare and visualize data

Work with models

Prompt Lab
Experiment with foundation models and build prompts

Tuning Studio
Tune a foundation model with labeled data

SPSS Modeler
Build models as a visual flow

RStudio IDE
Work with data and models in R

Jupyter notebook editor
Work with data and models in Python or R notebooks

AutoAI
Build ML models automatically

Decision Optimization
Solve optimization problems

Federated Learning
Train models on distributed data

Automate model lifecycles

Orchestration Pipelines
Automate model lifecycles

Common core services

Projects
Collaborate with others to work with data and build AI assets

Deployment Spaces
Deploy and monitor your assets

Hybrid Cloud Platform

Red Hat OpenShift AI

IBM Cloud Infrastructure

Third party infrastructure

Common core services

- Collaborative projects
- Deployment spaces
- Jobs
- Notifications
- Common connectivity
- Access and Authentication
- Resource management
- Central asset management system

watsonx.ai: Prompt Lab

Experiment with foundation models and build prompts

Interactive prompt builder

Includes prompt examples for various use cases and tasks

Experiment with different prompts, save and reuse older prompts, use different models and vary different parameters

Experiment with zero-shot, one-shot, or few-shot prompting to get the best results

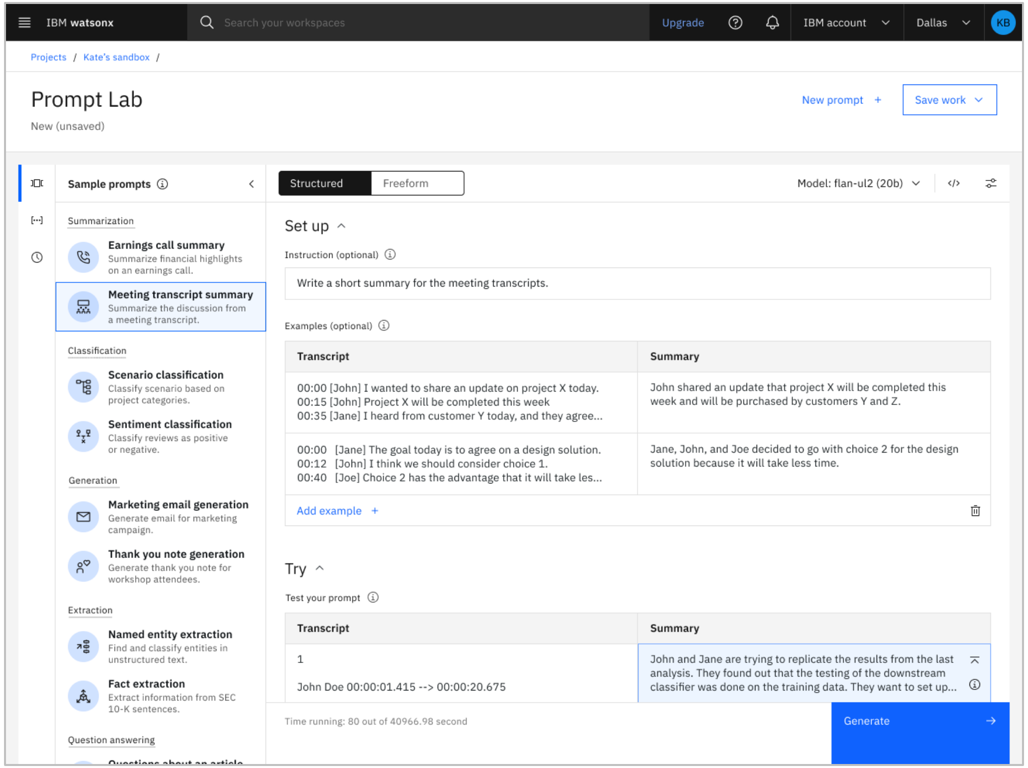
Experiment with prompt engineering

Choice of foundation models to use based on task requirements

Prevent the model from generating repeating phrases

Number of min and max new tokens in the response

Stop sequences – specifies sequences whose appearances should stop the model



Prompt Lab

New (unsaved)

Summarization

Earnings call summary
Summarize financial highlights on an earnings call.

Meeting transcript summary
Summarize the discussion from a meeting transcript.

Classification

Scenario classification
Classify scenario based on project categories.

Sentiment classification
Classify reviews as positive or negative.

Generation

Marketing email generation
Generate email for marketing campaign.

Thank you note generation
Generate thank you note for workshop attendees.

Extraction

Named entity extraction
Find and classify entities in unstructured text.

Fact extraction
Extract information from SEC 10-K sentences.

Question answering

Set up

Instruction (optional)

Write a short summary for the meeting transcripts.

Examples (optional)

Transcript	Summary
00:00 [John] I wanted to share an update on project X today. 00:15 [John] Project X will be completed this week 00:35 [Jane] I heard from customer Y today, and they agree...	John shared an update that project X will be completed this week and will be purchased by customers Y and Z.
00:00 [Jane] The goal today is to agree on a design solution. 00:12 [John] I think we should consider choice 1. 00:40 [Joe] Choice 2 has the advantage that it will take les...	Jane, John, and Joe decided to go with choice 2 for the design solution because it will take less time.

Add example +

Try

Test your prompt

Transcript	Summary
1 John Doe 00:00:01.415 --> 00:00:20.675	John and Jane are trying to replicate the results from the last analysis. They found out that the testing of the downstream classifier was done on the training data. They want to set up...

Time running: 80 out of 40966.98 second

watsonx.ai: Tuning Studio

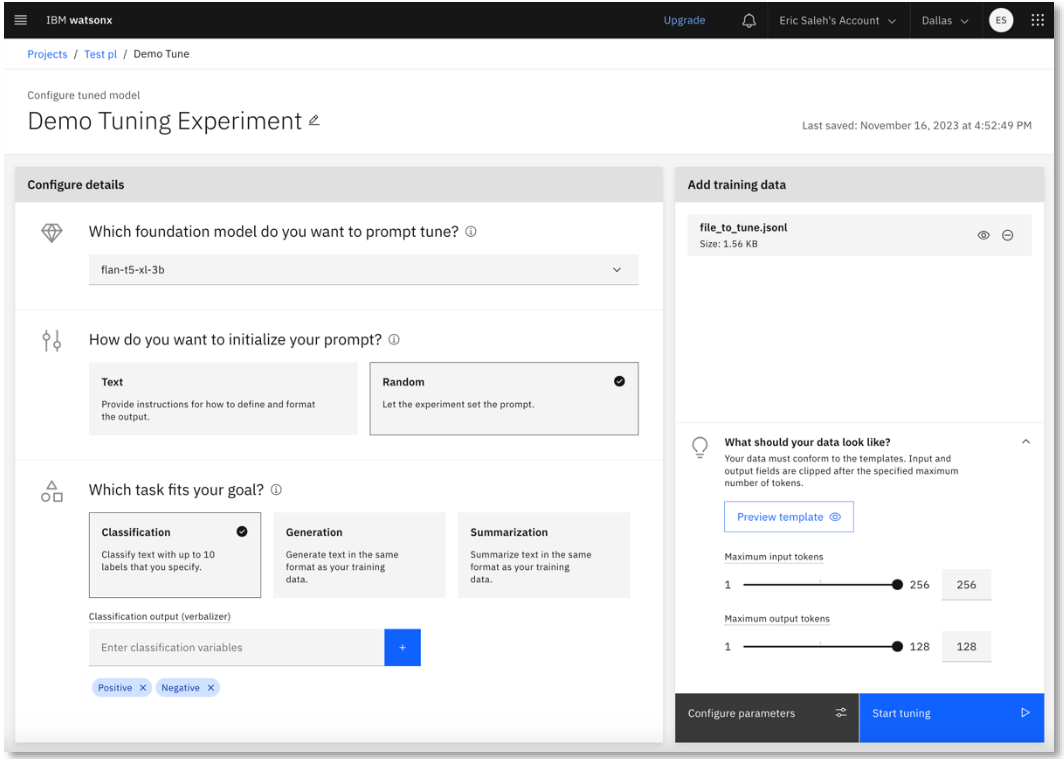
Tune your foundation models with labeled data

Summary:

- Tool for performing PEFT and fine-tuning training techniques to optimize FM task performance
- Tuned model can be deployed and inferenced via the API or Prompt Lab

Prompt-tuning:

- How it works: creates an optimized sequence of values (called a soft-prompt vector) to add as a prefix to FM prompt to improve task performance
- Technical origins: The Power of Scale for Parameter-Efficient Prompt Tuning
- Subset of PEFT, similar to P-Tuning, LoRA, etc.



Configure tuned model

Demo Tuning Experiment

Last saved: November 16, 2023 at 4:52:49 PM

Configure details

Which foundation model do you want to prompt tune?

flan-t5-xl-3b

How do you want to initialize your prompt?

Text: Provide instructions for how to define and format the output.

Random (selected): Let the experiment set the prompt.

Which task fits your goal?

Classification (selected): Classify text with up to 10 labels that you specify.

Generation: Generate text in the same format as your training data.

Summarization: Summarize text in the same format as your training data.

Classification output (verbalizer)

Enter classification variables +

Positive ✕ Negative ✕

Add training data

file_to_tune.jsonl
Size: 1.56 KB

What should your data look like?

Your data must conform to the templates. Input and output fields are clipped after the specified maximum number of tokens.

Preview template

Maximum input tokens: 1 — 256 | 256

Maximum output tokens: 1 — 128 | 128

Product documentation

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.

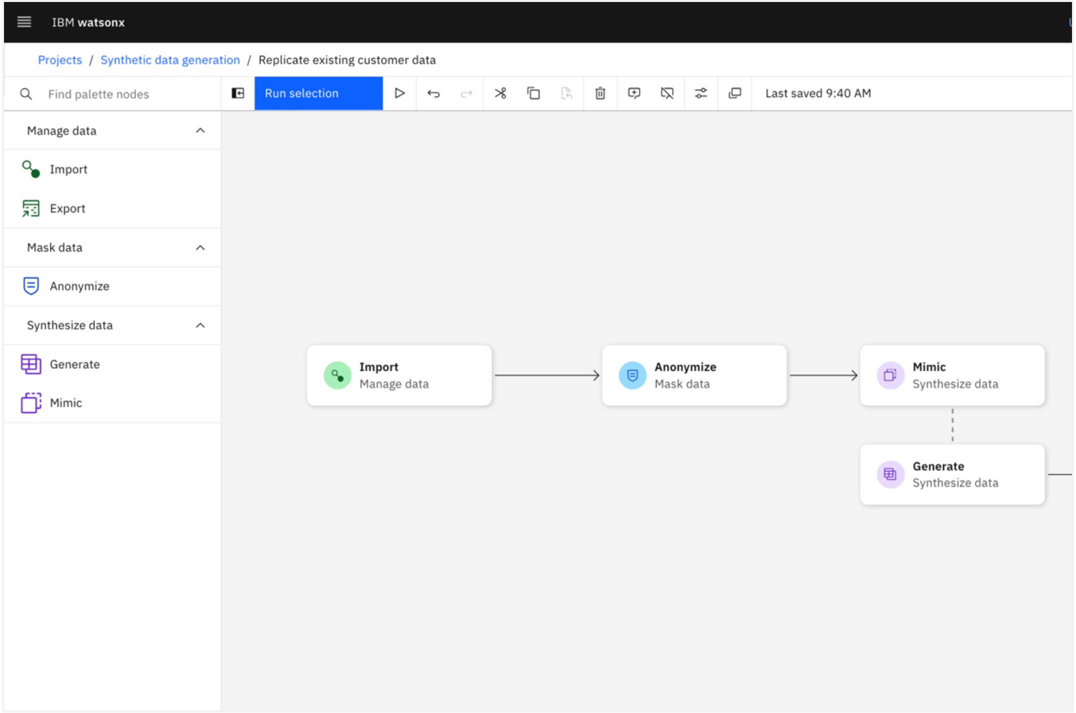
Create synthetic data using your existing data in a database or by uploading a file. If no data exists or can’t be accessed, you can design your own data schema.

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.



*Evaluation metrics available in Q3 2024

watsonx.ai: Data Science and MLOps

Build machine learning models automatically in the studio

Model training and development

Build experiments quickly and enhance training by optimizing pipelines and identifying the right combination of data

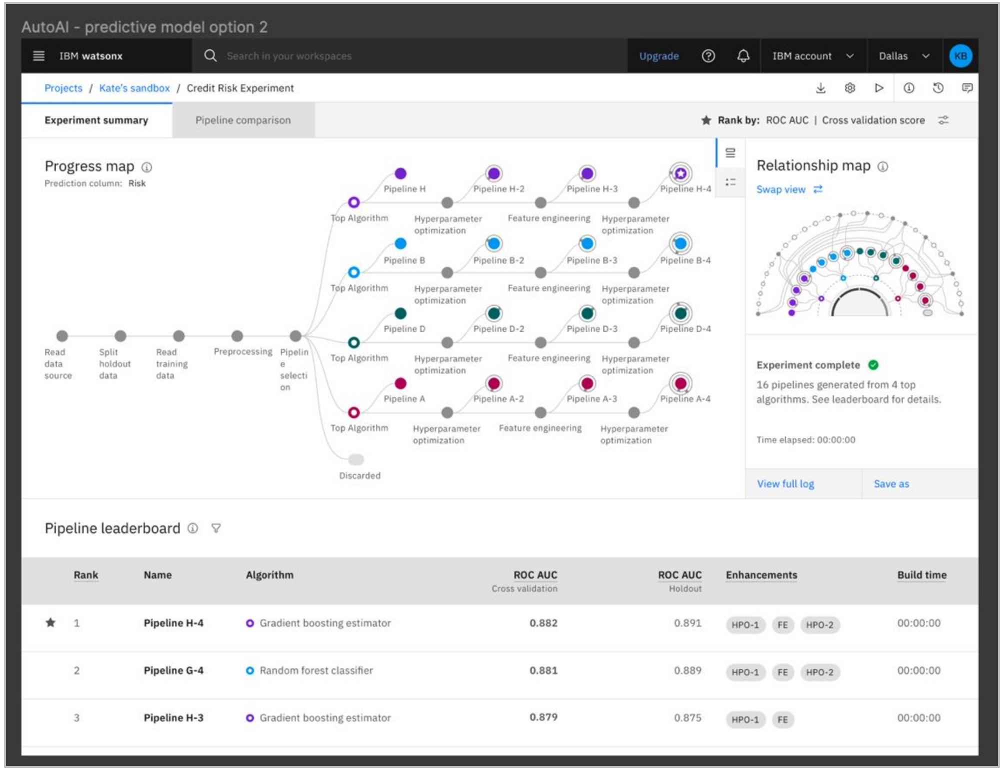
AutoAI, including preparing data for machine learning and generating and ranking candidate model pipelines

Use predictions to optimize decisions, create and edit models in Python, in OPL or with natural language

Integrated visual modeling

Prepare data quickly and develop models visually to help visualize and analyze enterprise data to identify patterns and trends, explore opportunities, and make informed, insightful business decisions

- Uncover correlations
- Insight for hypotheses
- Find relationships and connections within the data



Experiment complete

16 pipelines generated from 4 top algorithms. See leaderboard for details.

Time elapsed: 00:00:00

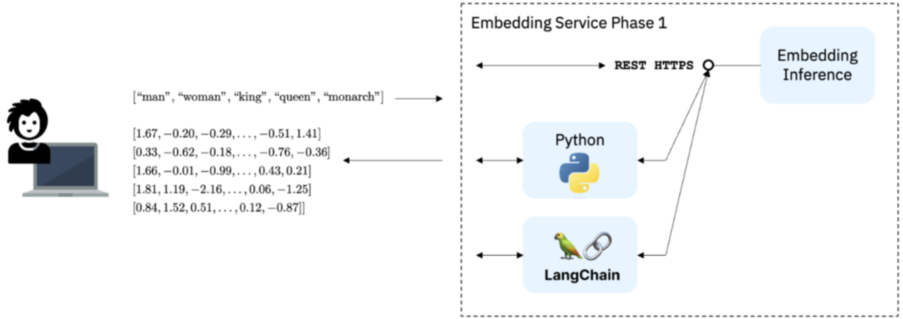
Pipeline leaderboard

Rank	Name	Algorithm	ROC AUC Cross validation	ROC AUC Holdout	Enhancements	Build time
★ 1	Pipeline H-4	Gradient boosting estimator	0.882	0.891	HPO-1 FE HPO-2	00:00:00
2	Pipeline G-4	Random forest classifier	0.881	0.889	HPO-1 FE HPO-2	00:00:00
3	Pipeline H-3	Gradient boosting estimator	0.879	0.875	HPO-1 FE	00:00:00

watsonx.ai Embeddings API

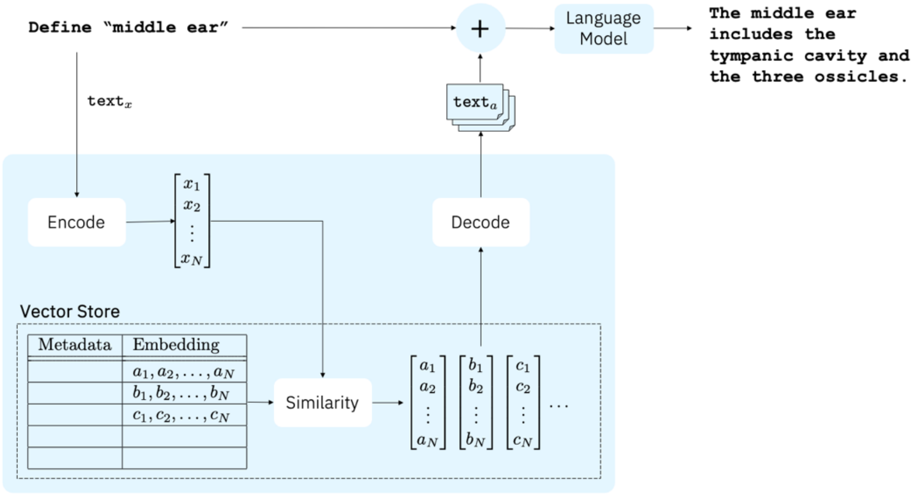
What does it do?

- Converts input text into embeddings, which are dense vector representations of the input text
- Embeddings capture nuanced semantic and syntactic relationships between words and passages in vector space



Customer value

- Embeddings provide a more semantically faithful representation of the supplied text, especially when compared to basic keyword-based alternatives in classic NLP modeling
- The efficient storage and compute profiles of embeddings make them easily infusible into generative AI application
- Retrieval Augmented Generation (RAG) patterns utilize embedding models for query and passage vectorization, enabling contextual grounding



Define "middle ear"
text_x
+
Language Model
The middle ear includes the tympanic cavity and the three ossicles.
text_a
Encode
$[x_1, x_2, \vdots, x_N]$
Decode
Vector Store
Similarity
$[a_1, a_2, \vdots, a_N]$ $[b_1, b_2, \vdots, b_N]$ $[c_1, c_2, \vdots, c_N]$ . . .

Metadata	Embedding
	$a_1, a_2, \ldots, a_N$
	$b_1, b_2, \ldots, b_N$
	$c_1, c_2, \ldots, c_N$

IBM differentiators

- Performance matching or exceeding market leaders in retrieval benchmarks
- Exclusively trained on legally approved and commercially viable data to enable enterprise usage
- Scalable, fully-managed, and integrated with the broader watsonx portfolio

watsonx.governance

Accelerate responsible, transparent and explainable AI

One unified, integrated AI Governance platform to govern generative AI and predictive ML

Lifecycle Governance

Govern across the AI lifecycle. Automate and consolidate tools, applications and platforms. Capture metadata at each stage and support models built and deployed in 3rd party tools.

Risk Management

Manage risk & protect reputation by automating workflows to ensure quality and better detect bias and drift.

Regulatory Compliance

Adhere to regulatory compliance by translating growing regulations into enforceable policies.

Comprehensive

Govern the end-to-end AI lifecycle with metadata capture at each stage

Open

Support governance of models built and deployed in 3rd party tools.

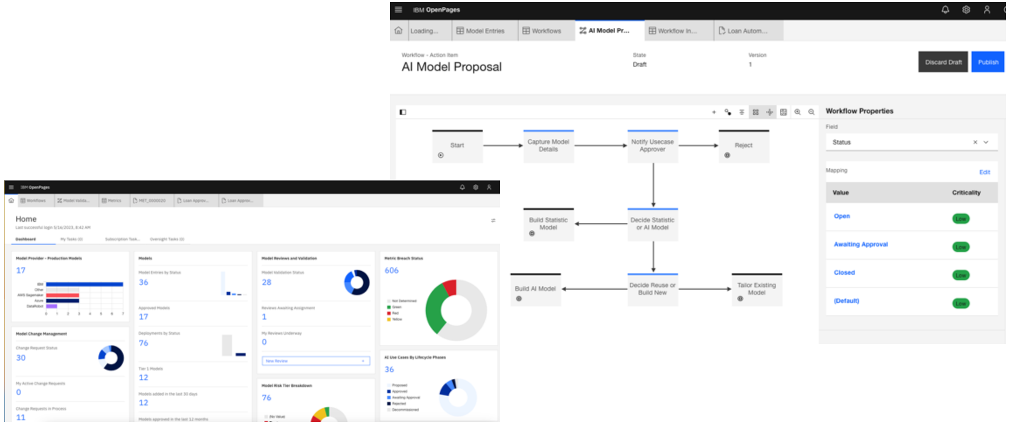
Automatic metadata recording

and data transformation/lineage capture though Python notebooks.

watsonx.governance

Model Risk Governance

- Consolidated view of models from multiple platforms
- View development status, model performance and alerts or emerging issues
- Monitor and trigger workflows for model validation, retraining and performance issues



Workflow - Action Item
AI Model Proposal

State: Draft
Version: 1

Workflow Properties

Field: Status

Mapping Edit

Value	Criticality
Open	Low
Awaiting Approval	Low
Closed	Low
(Default)	Low

Home
Last successful login 5/16/2023, 8:42 AM

Model Provider - Production Models: 17

Model Change Management
Change Request Status: 30
My Active Change Requests: 0
Change Requests in Process: 11

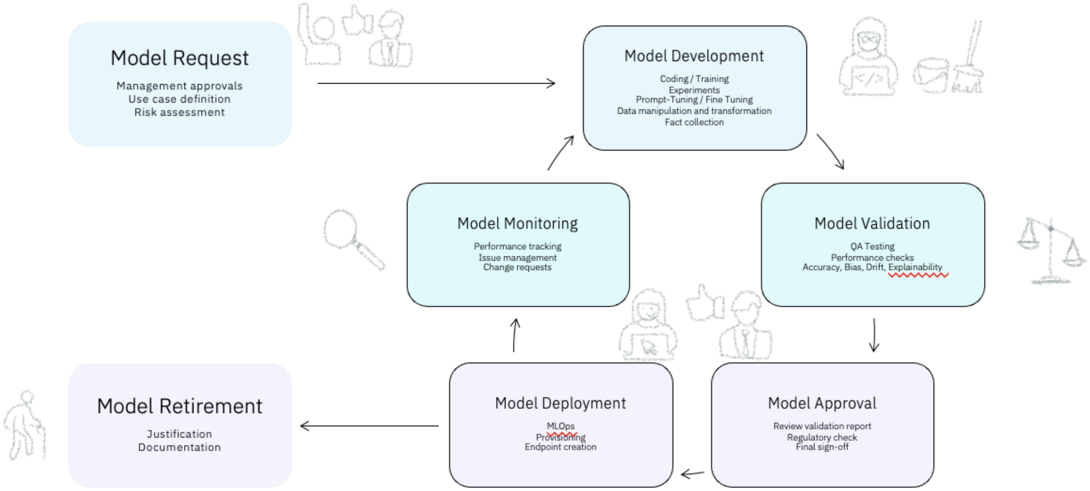
Models
Model Entries by Status: 36
Approved Models: 17
Deployments by Status: 76
Tier 1 Models: 12
Models added in the last 30 days: 12
Models approved in the last 12 months

Model Reviews and Validation
Model Validation Status: 28
Reviews Awaiting Assignment: 1
My Reviews Underway: 0

Model Risk Tier Breakdown: 76

Metric Breach Status: 606

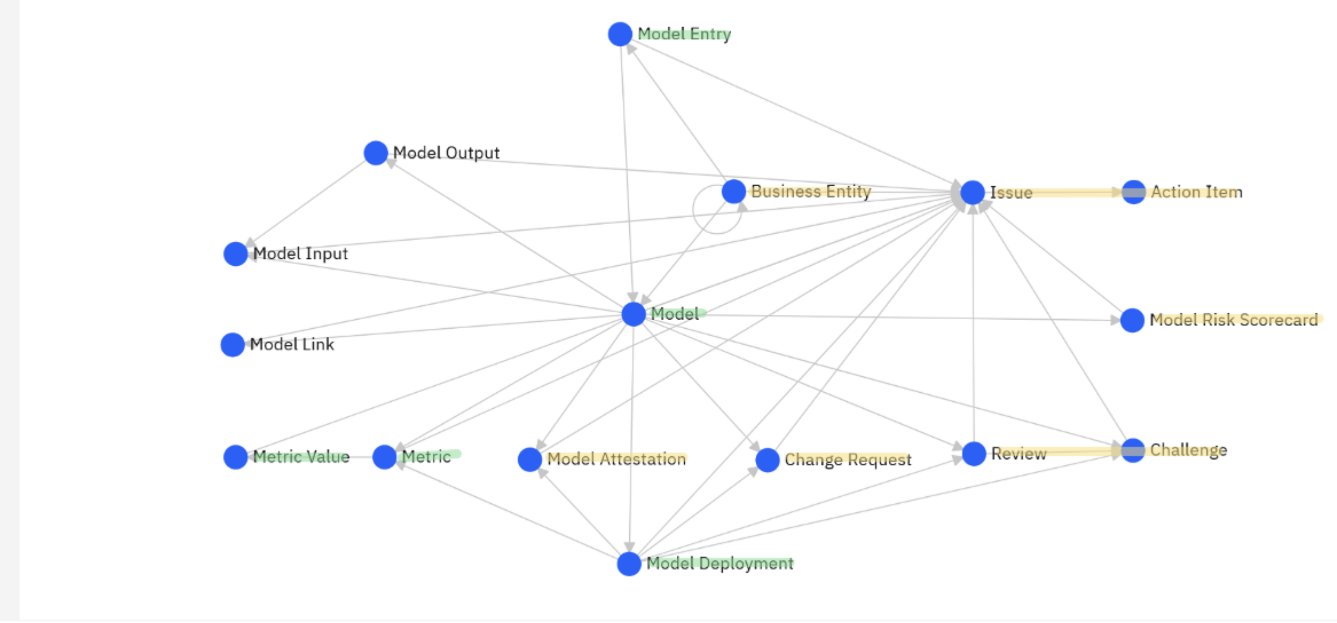
AI Use Cases By Lifecycle Phases: 36

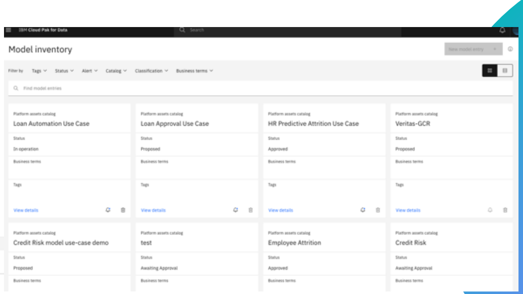


watsonx.governance

AI Governance object model

Much more than just "a model"





Model inventory

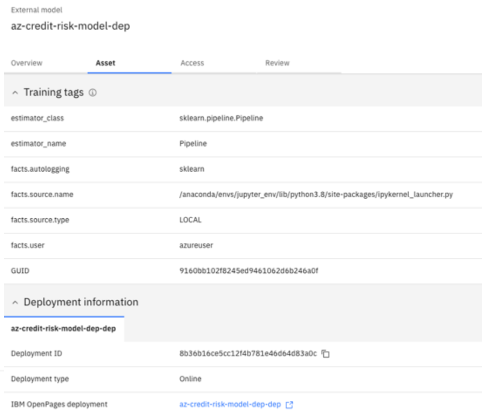
Filter by Tags Status Alert Catalog Classification Business terms

Find model entries

Platform assets catalog	Platform assets catalog	Platform assets catalog	Platform assets catalog
Loan Automation Use Case	Loan Approval Use Case	HR Predictive Attrition Use Case	Veritas-GCR
Status In operation	Status Proposed	Status Approved	Status Proposed
Business terms	Business terms	Business terms	Business terms
Tags	Tags	Tags	Tags
View details	View details	View details	View details

Platform assets catalog	Platform assets catalog	Platform assets catalog	Platform assets catalog
Credit Risk model use-case demo	test	Employee Attrition	Credit Risk
Status Proposed	Status Awaiting Approval	Status Approved	Status Awaiting Approval
Business terms	Business terms	Business terms	Business terms

Model Inventory and lifecycle Tracking



External model

az-credit-risk-model-dep

Training tags

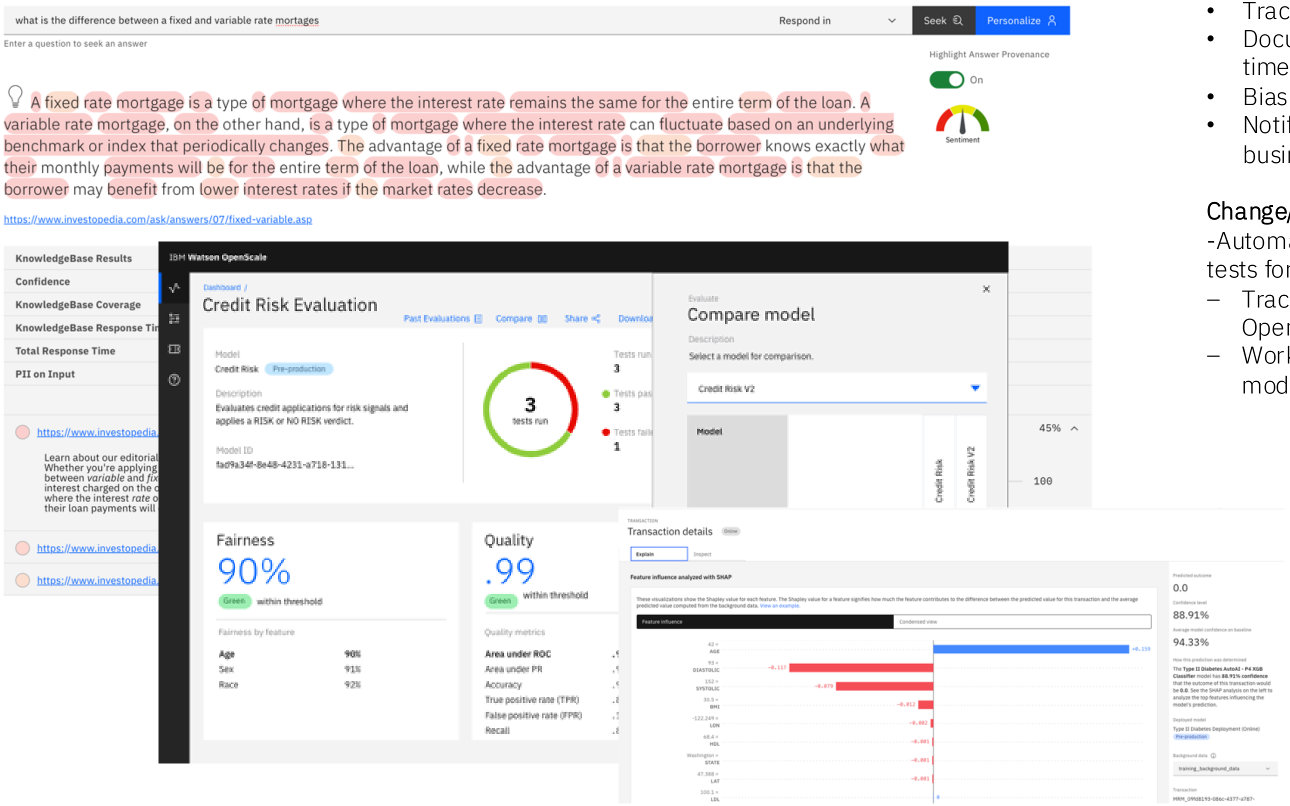
estimator_class	sklearn.pipeline.Pipeline
estimator_name	Pipeline
facts.autologging	sklearn
facts.source.name	/anaconda/envs/jupyter_env/lib/python3.8/site-packages/ipykernel_launcher.py
facts.source.type	LOCAL
facts.user	azureuser
GUID	9160bb102f8245ed9461062d6b246a0f

Deployment information

az-credit-risk-model-dep-dep

Deployment ID	8b36b16ce5cc12f4b781e46d64d83a0c
Deployment type	Online
IBM OpenPages deployment	az-credit-risk-model-dep-dep

watsonx.governance



what is the difference between a fixed and variable rate mortages

Respond in

Enter a question to seek an answer

A fixed rate mortgage is a type of mortgage where the interest rate remains the same for the entire term of the loan. A variable rate mortgage, on the other hand, is a type of mortgage where the interest rate can fluctuate based on an underlying benchmark or index that periodically changes. The advantage of a fixed rate mortgage is that the borrower knows exactly what their monthly payments will be for the entire term of the loan, while the advantage of a variable rate mortgage is that the borrower may benefit from lower interest rates if the market rates decrease.

https://www.investopedia.com/ask/answers/07/fixed-variable.asp

KnowledgeBase Results
Confidence
KnowledgeBase Coverage
KnowledgeBase Response Ti
Total Response Time
PII on Input

https://www.investopedia.

Learn about our editorial
Whether you're applying
between *variable* and *fix*
interest charged on the c
where the interest *rate* o
their loan payments will

https://www.investopedia.

https://www.investopedia.

IBM Watson OpenScale

Dashboard /

Credit Risk Evaluation

Model
Credit Risk Pre-production

Description
Evaluates credit applications for risk signals and applies a RISK or NO RISK verdict.

Model ID
fad9a34f-8e48-4231-a718-131...

3 tests run

Tests run
3

Tests pas
3

Tests faile
1

Fairness
90%
Green within threshold

Fairness by feature

Age	90%
Sex	91%
Race	92%

Quality
.99
Green within threshold

Quality metrics

Area under ROC
Area under PR
Accuracy
True positive rate (TPR)
False positive rate (FPR)
Recall

Evaluate

Compare model

Description
Select a model for comparison.

Credit Risk V2

TRANSACTION

Transaction details Online

Feature influence analyzed with SHAP

These visualizations show the Shapley value for each feature. The Shapley value for a feature signifies how much the feature contributes to the difference between the predicted value for this transaction and the average predicted value computed from the background data. View an example.

Predicted outcome
0.0

Confidence level
88.91%

Average model confidence on baseline
94.33%

How this prediction was determined
The **Type II Diabetes AutoAI - P4 XGB Classifier** model has **88.91% confidence** that the outcome of this transaction would be **0.0**. See the SHAP analysis on the left to analyze the top features influencing the model's prediction.

Deployed model
Type II Diabetes Deployment (Online)
Pre-production

Background data
training_background_data

Transaction
MRM_09fd8193-086c-4377-a787-

Performance Monitoring

- Ongoing health monitoring of AI Models during runtime
- Trace and explain AI predictions
- Document metrics and track metric values over time
- Bias detection and mitigation
- Notification of issues when quality thresholds or business KPIs are violated

Change/Issue Management

-Automatic deployment and execution of validation tests for AI Models

– Track issues and incidents related to models in OpenPages included Issue Management Solution
– Workflow to document and approve changes to models

Evaluation and Monitoring

Thank you

THIS DOCUMENT IS DISTRIBUTED “AS IS” WITHOUT ANY WARRANTY, EITHER EXPRESS OR IMPLIED. IN NO EVENT, SHALL IBM BE LIABLE FOR ANY DAMAGE ARISING FROM THE USE OF THIS INFORMATION, INCLUDING BUT NOT LIMITED TO, LOSS OF DATA, BUSINESS INTERRUPTION, LOSS OF PROFIT OR LOSS OF OPPORTUNITY.

Client examples are presented as illustrations of how those clients have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary.

Not all offerings are available in every country in which IBM operates.

IBM’s statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM’s sole discretion. Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

IBM