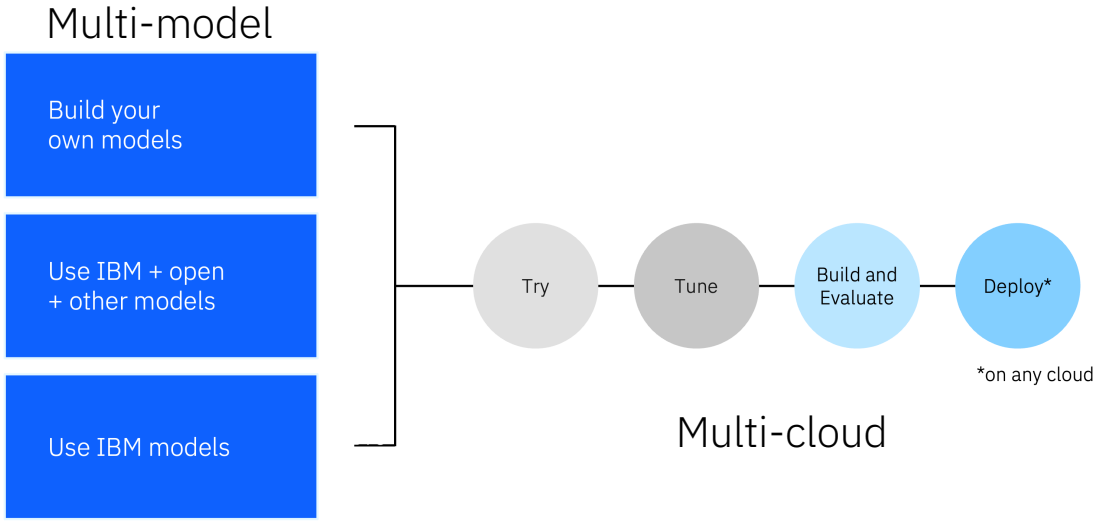
MLOps and Trustworthy AI with **watsonx**

watsonx

IBM

IBM’s Approach to AI for Business



TOTE Design Principles

Trusted	Transparent, responsible, and governed
Open	Based on the best open technologies available
Targeted	Designed for enterprise and targeted at business domains
Empowering	For value creators, not just users

Put AI to work with **watsonx**.

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

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

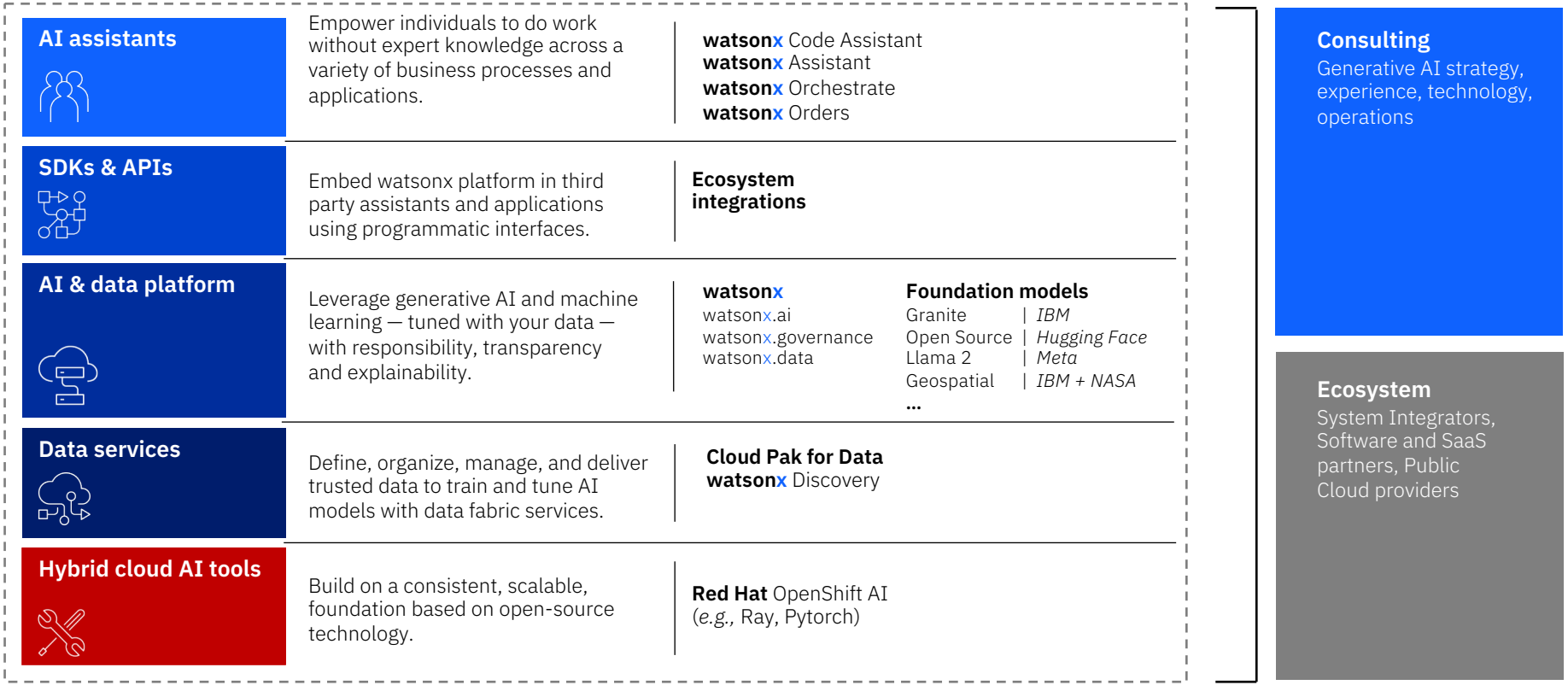
Increase developer productivity, reduce coding complexity, and accelerate developer onboarding. Purpose-built for targeted use cases, watsonx Code Assistant uses AI to support application modernization and IT automation.

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



watsonx

Model strategy →

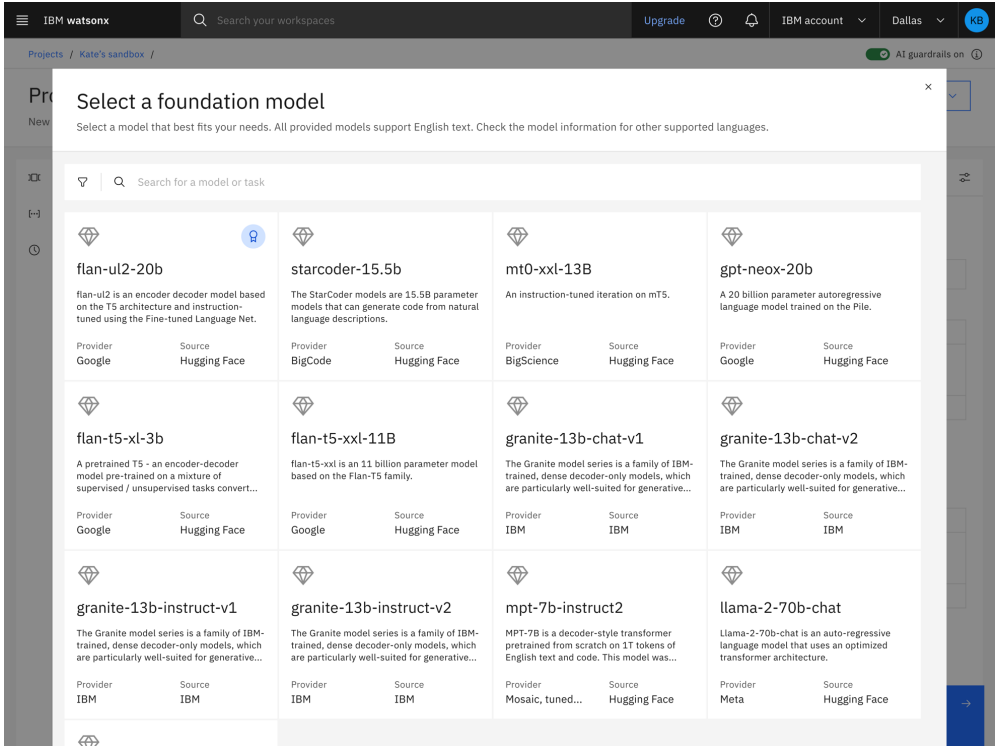
Multi-model

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

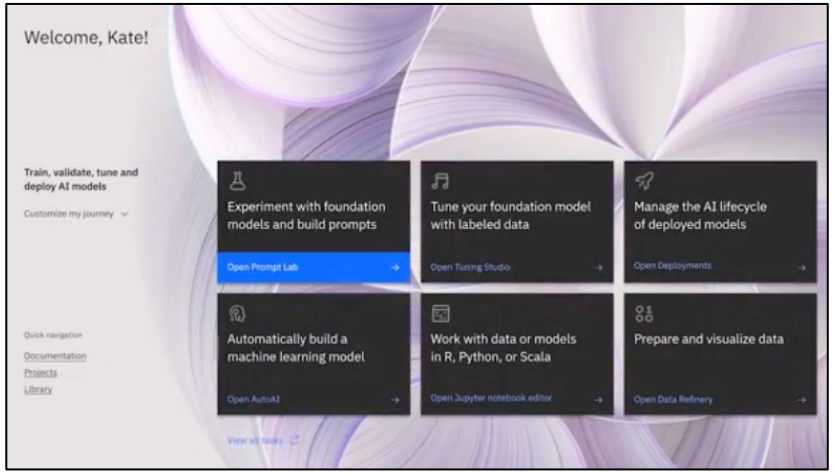
Hybrid deployments. We provide the flexibility to deploy models on the platform of choice.



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

watsonx.ai

Build, train, validate, tune, and deploy AI models



A next generation enterprise studio for AI builders to train, validate, tune, and deploy generative AI, foundation models, and machine learning capabilities.

The watsonx.ai components include:

- **Foundation Model Library** with IBM and open-source models
- **Prompt Lab** to experiment with foundation models and build prompts for various use cases and tasks
- **Tuning Studio** to tune your foundation models with labeled data
- **Data Science and MLOps** to build machine learning models automatically with model training, development, visual modeling, and synthetic data generation

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

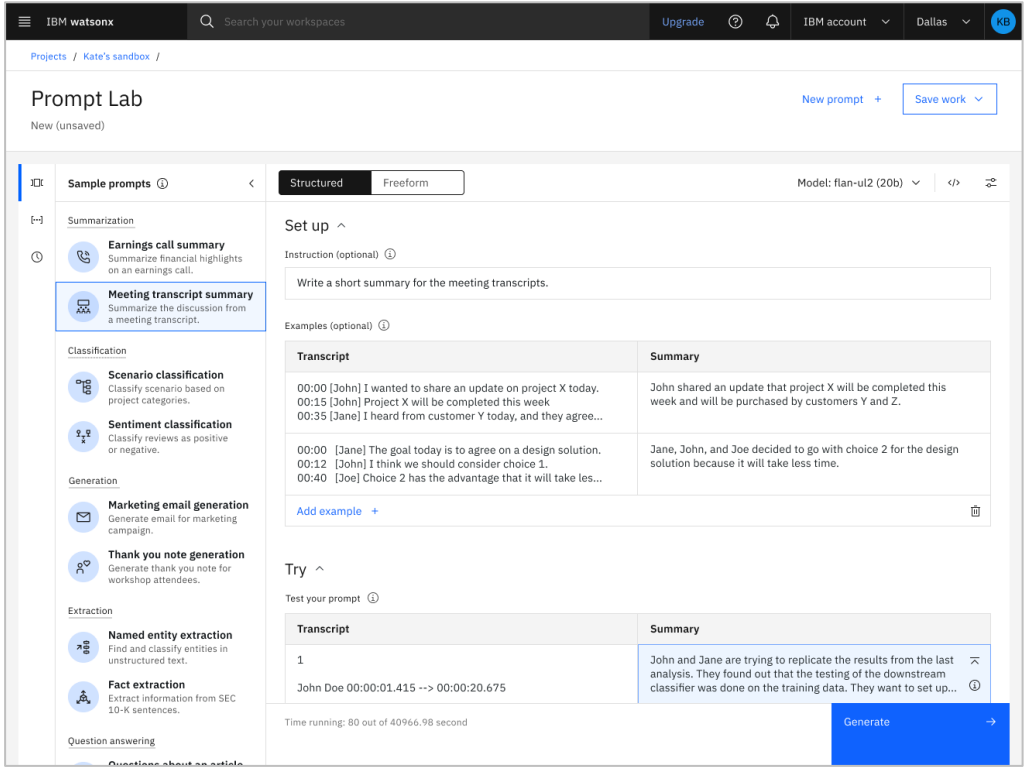
Experiment with prompt engineering

Choice of foundation models to use based on task requirements

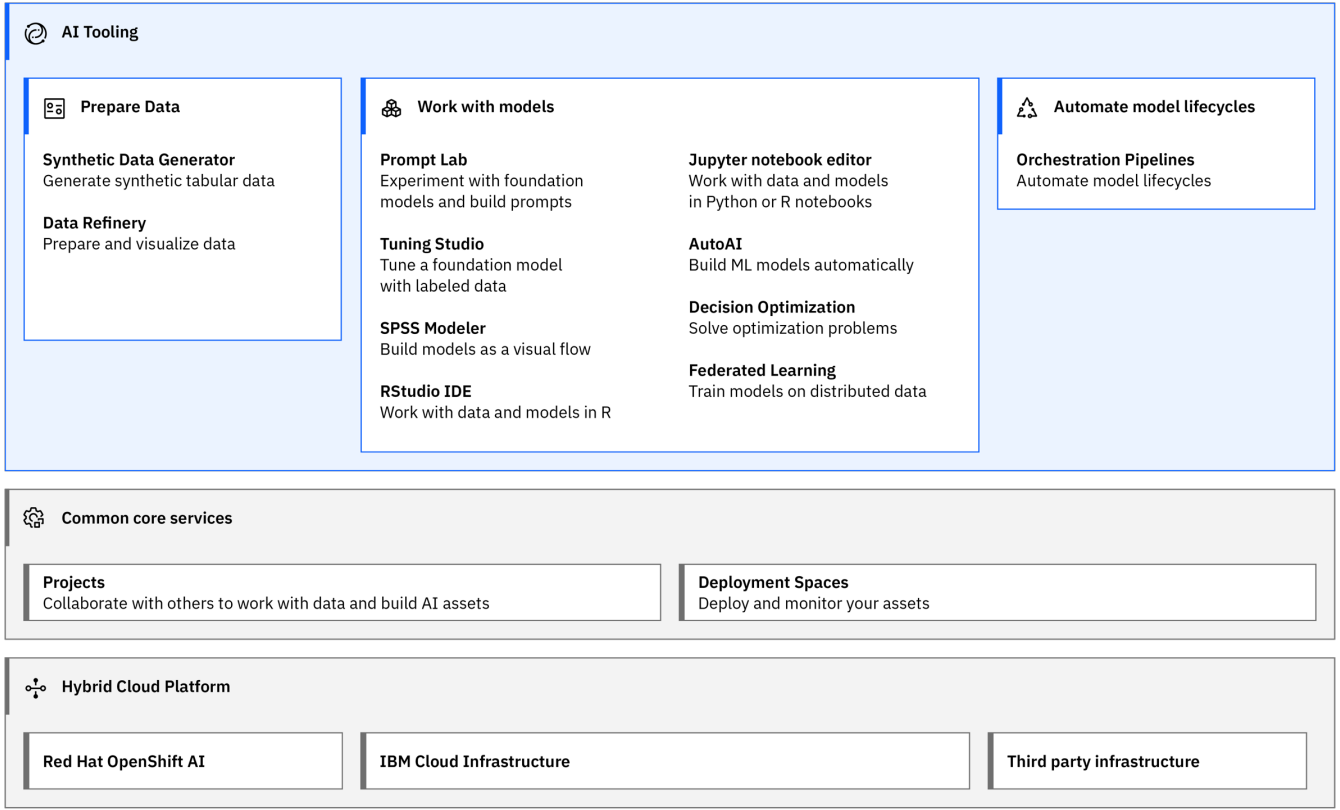
Prevent the model from generating repeating phrases

Number of min and max new tokens in the response

Stop sequences – specifies sequences whose appearances should stop the model



IBM watsonx.ai architecture



Common core services

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

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

Initial tuning method at GA: 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.

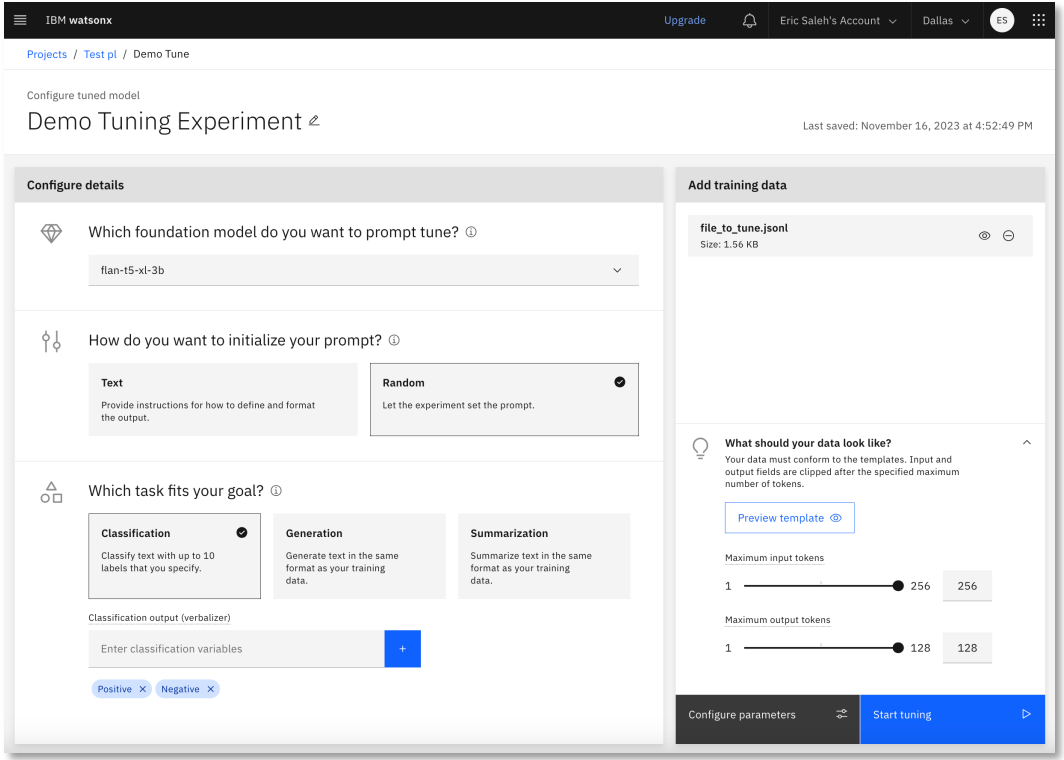
FMs eligible for prompt-tuning:

- flan-t5-xl-3b, granite-13b-chat and llama-2-13b-chat
- SaaS (Dallas, Tokyo, Frankfurt DC)
- Additional FMs currently in-development

Pricing:

- 43 capacity-unit-hours (CUH) rate per hour of active tuning
- Inference Resource-Unit price for deployed tuned model depends on FM inferencing class (learn more)

Product documentation



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

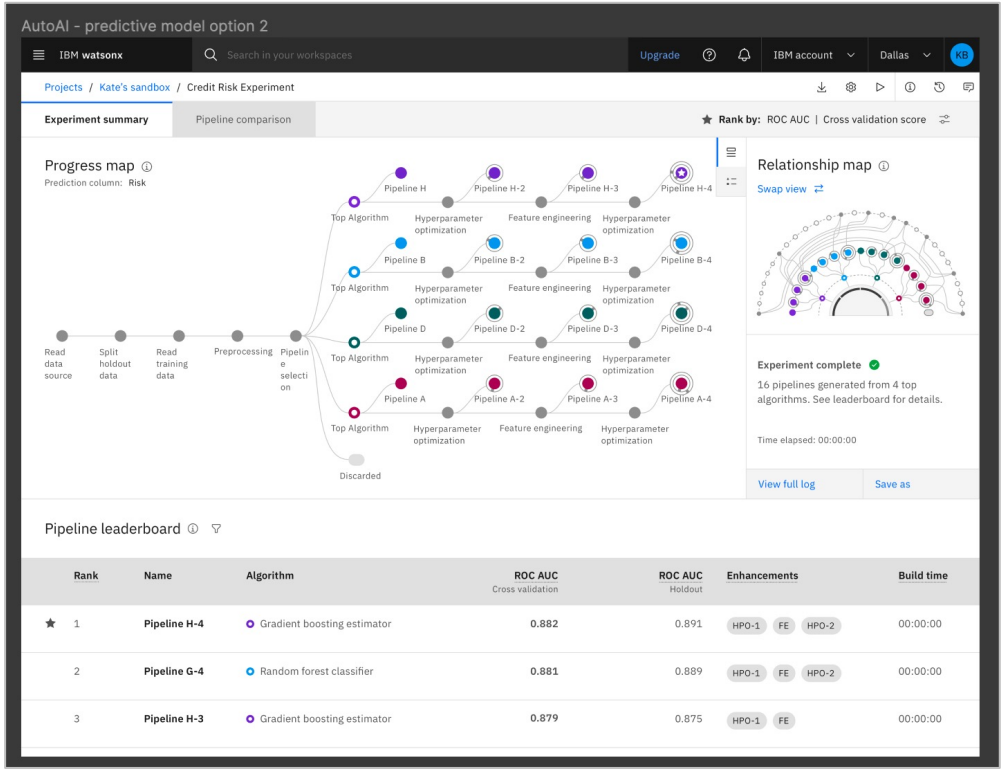
AutoAI, including preparing data for machine learning and generating and ranking candidate model pipelines

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



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.

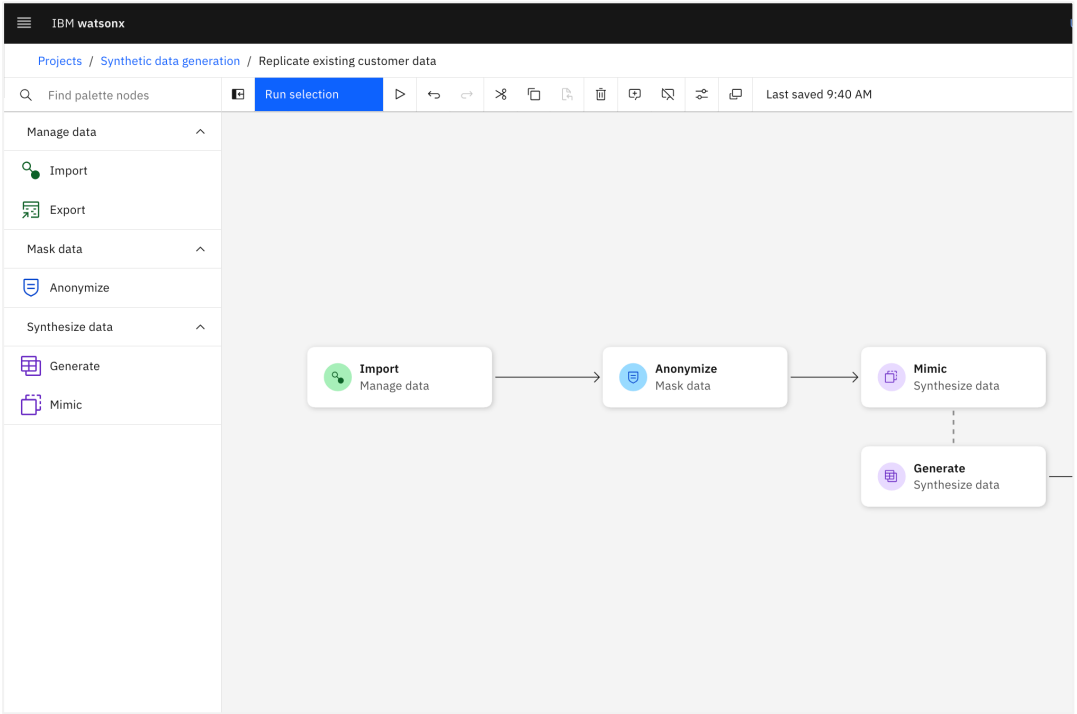
Create synthetic data using your existing data in a database or by uploading a file. If no data exists or can't be accessed, you can design your own data schema.

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.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

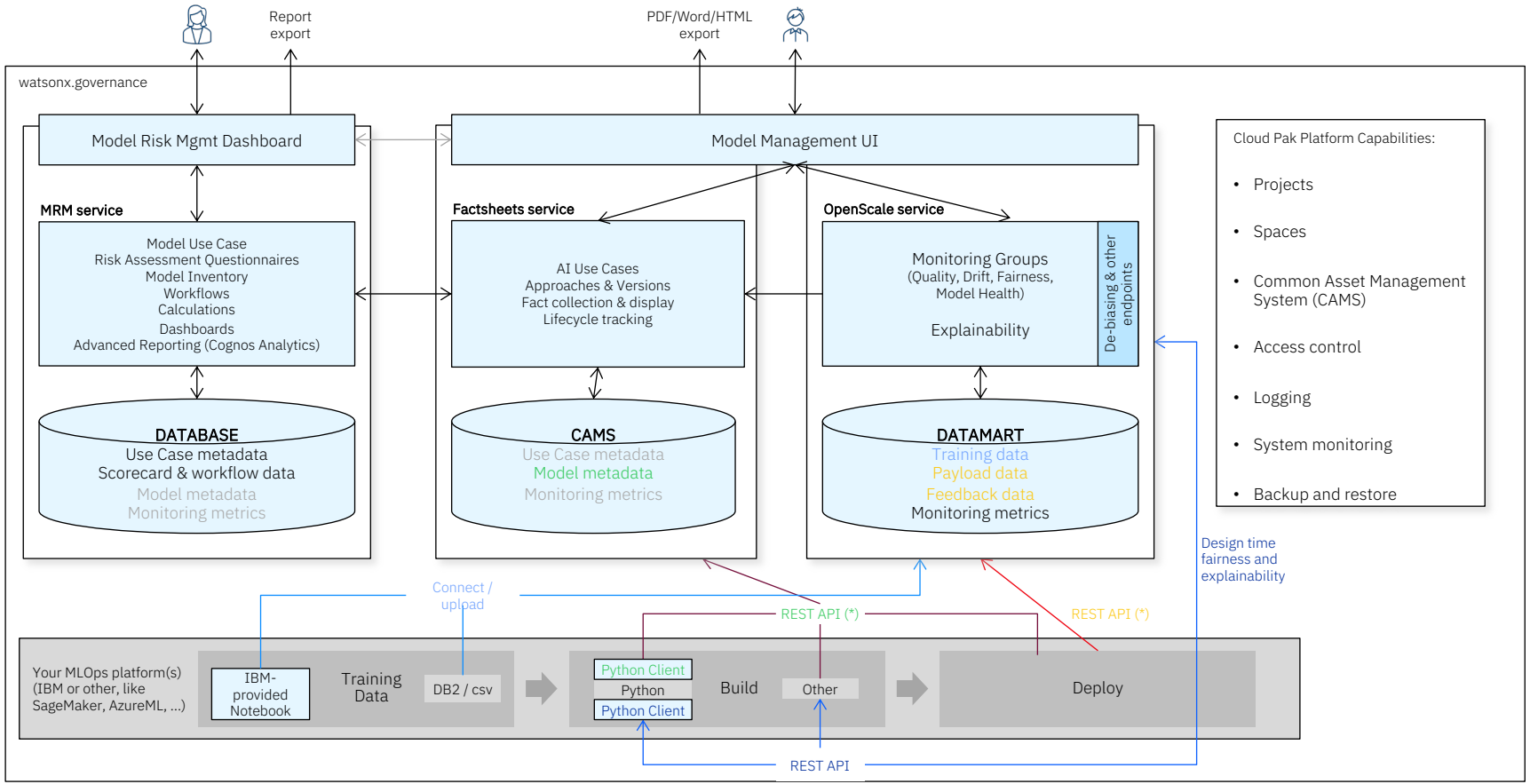
Adhere to regulatory compliance by translating growing regulations into enforceable policies.

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.

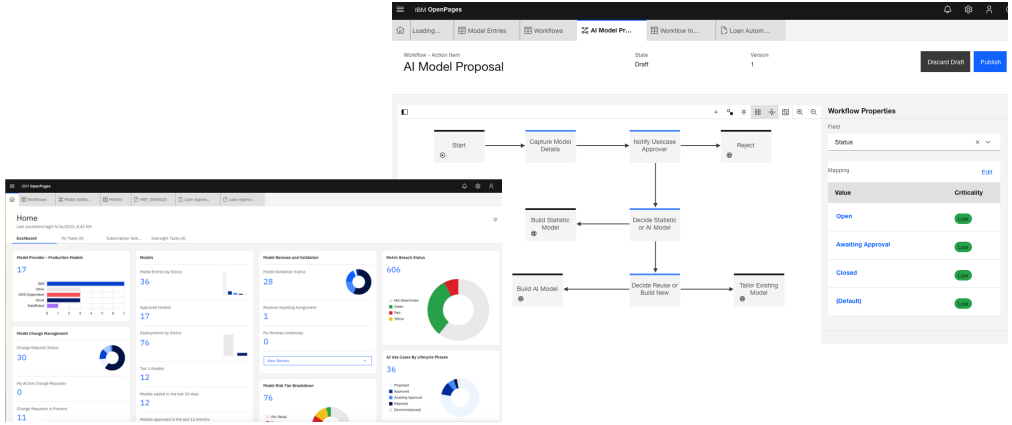
Automatic metadata recording
and data transformation/lineage capture though Python notebooks.

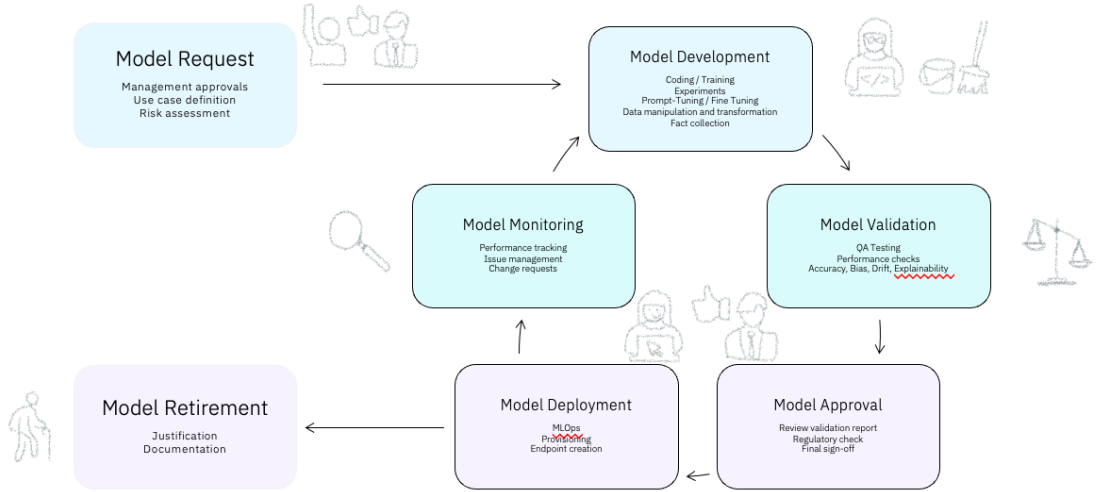
watsonx.governance – functional architecture (software)



watsonx.governance

Model Risk Governance



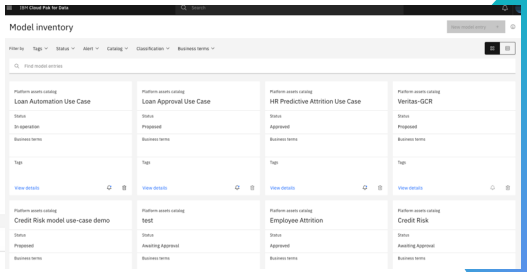


- Consolidated view of models from multiple platforms
- View development status, model performance and alerts or emerging issues
- Monitor and trigger workflows for model validation, retraining and performance issues

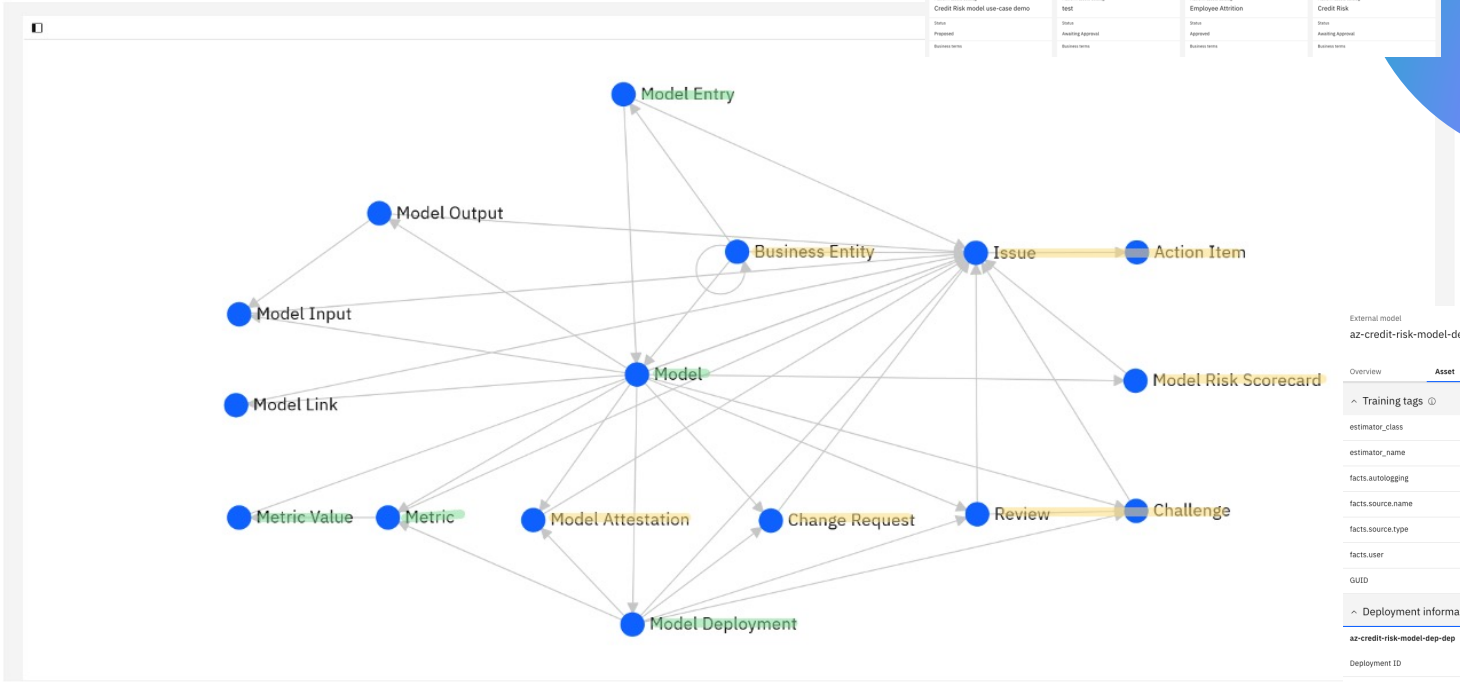
watsonx.governance

AI Governance object model

Much more than just "a model"



Model Inventory and lifecycle Tracking



External model

az-credit-risk-model-dep

Overview | Asset | Access | Review

Training tags

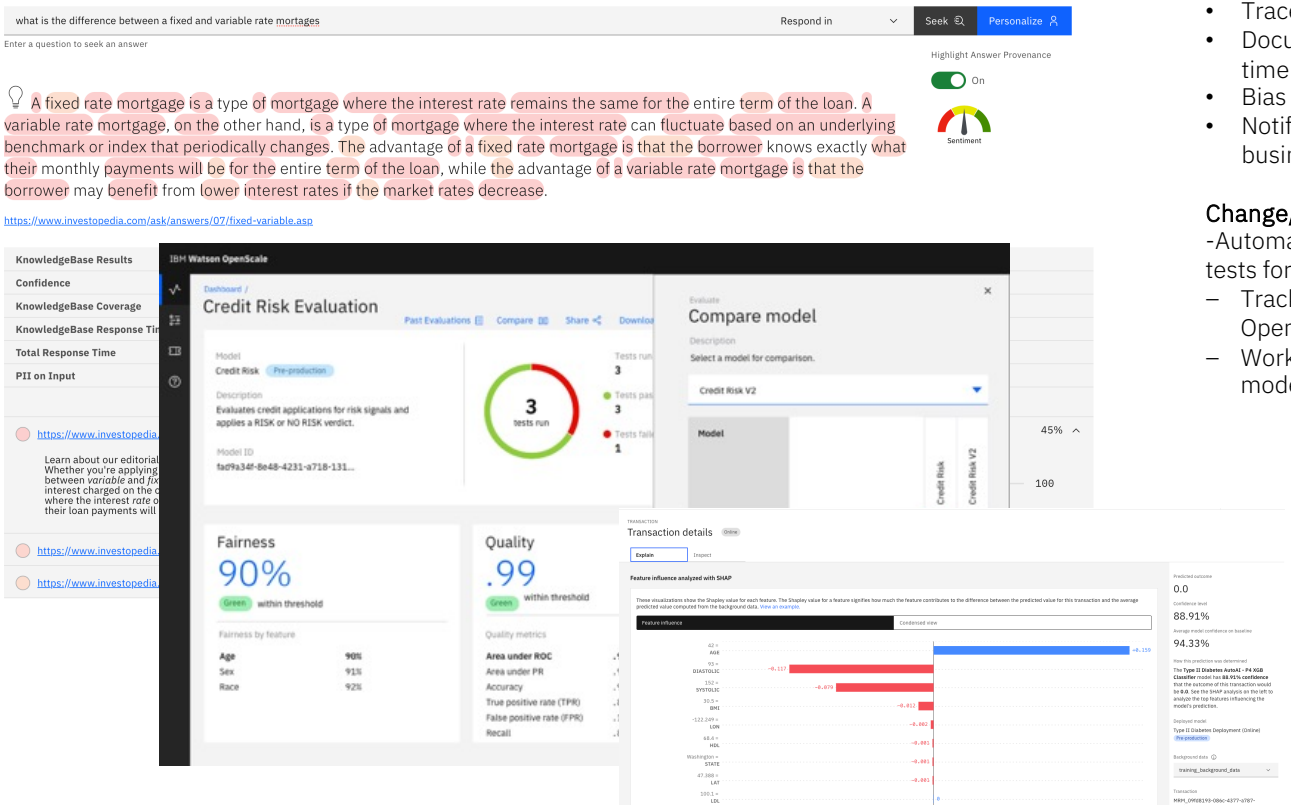
estimator_class	sklearn.pipeline.Pipeline
estimator_name	Pipeline
facts.autologging	sklearn
facts.source.name	/anaconda/envs/jupyter_env/lib/python3.8/site-packages/ipykernel_launcher.py
facts.source.type	LOCAL
facts.user	azureuser
GUID	9160bb102f8245ed946106206b246a0f

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



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

Model Management

Evaluate and monitor hallucination and answer quality for Retrieval Augmented Generation (RAG)

Value-add for:

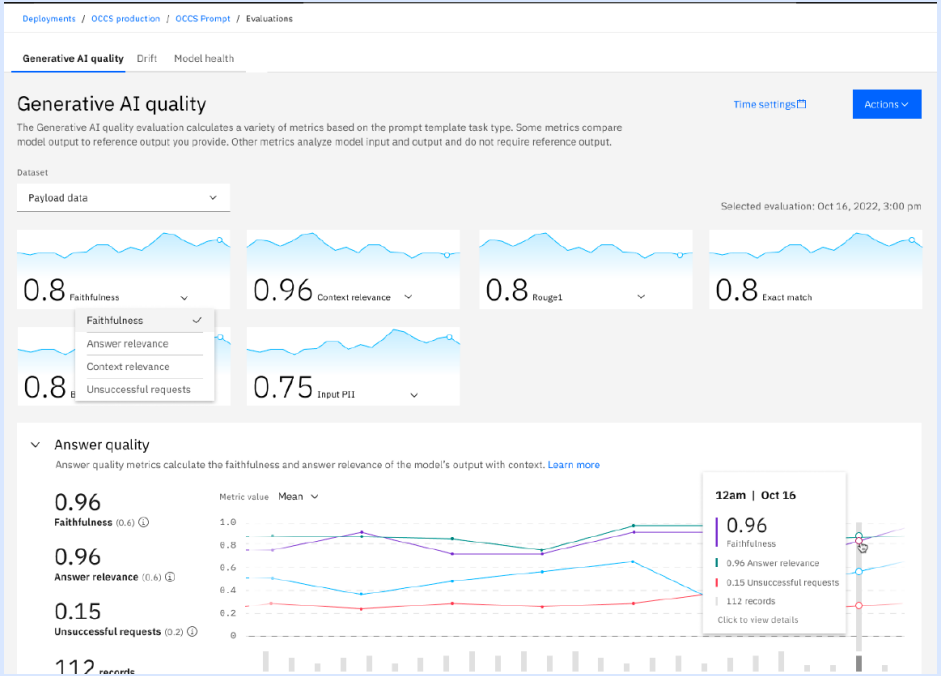
AI Engineers and users that are developing, validating, and deploying RAG to understand the quality of responses

Key features in model management:

- Quantify quality through these metrics:
  - Faithfulness
  - Answer relevance
  - Unsuccessful request
  - Keyword inclusion*
  - Answer coverage*
  - Spelling robustness* of question
- Visualize trends over time
- Assess metrics for individual responses with visualization to easily understand / assess response

*For 2.0, supported for development-time. Runtime support in later releases

Note: above is not intended to be an exhaustive list



Please note: UI support for RAG in 2.0.1 (API support in 2.0.0)

Metrics for evaluating Large Language Models

Text Summarization Metrics

- ROUGE
- SARI
- WIKI_SPLIT
- BLEURT
- METEOR
- Sentence Similarity - Jaccard Similarity
- Sentence Similarity - Cosine Similarity

Content Generation, Q&A Evaluation Metrics

- BLEU
- exact_match
- Perplexity**
- rl_reliability**

Text Classification Metrics

- Accuracy
- Precision
- Recall
- ROC AUC
- F1 Score
- Brier Score
- GLUE metrics
- Matthews Correlation Coefficient
- Label Skew

Watson NLP

- HAP Detection**
- PII Detection**

** Uses Metrics Computation Model

Entity Extraction Metrics

- Seq eval

Language Translation Evaluation Metrics

- Character
- charcut_mt
- Chrf
- google_bleu
- super_glue
- TER
- nist_mt
- Poseval
- sacrebleu
- XTREME-S

Reference-Free Metrics from Haifa Research Labs

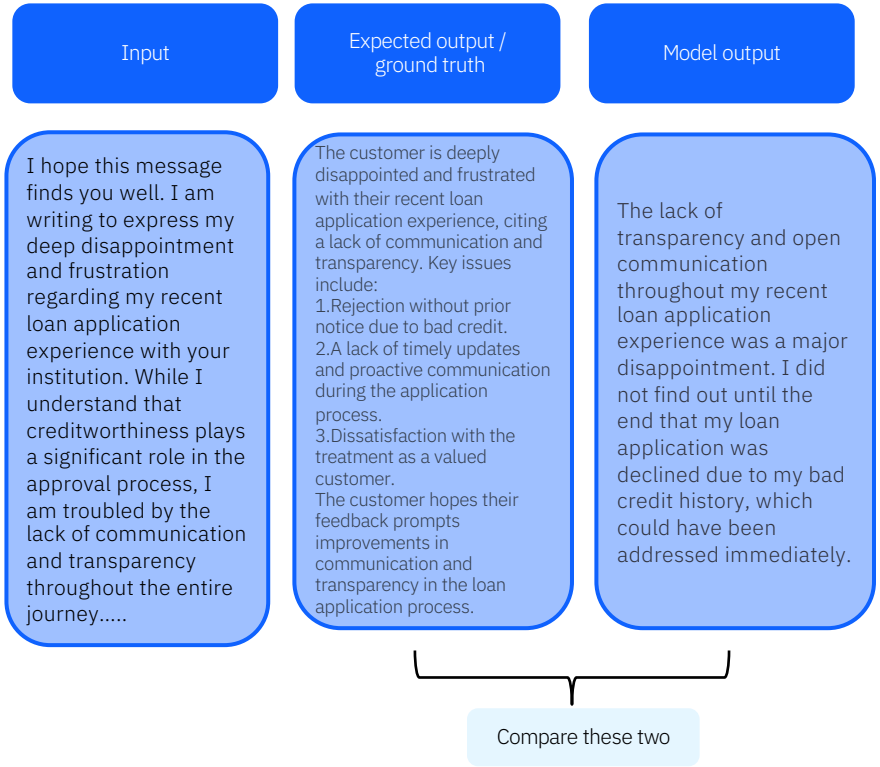
- Levenshtein distance based Diversity metrics
- Textstat toolkit based flesch metrics to determine readability, complexity, and grade level.
- blanchelp**
- Shannon**
- BartScore - Exploring
- Rquge - Exploring

Under Exploration from SVL Research Labs

- Stigma Detection**
- Social Bias/Values Detection**
- Faithfulness / Hallucination**

Summarization Example

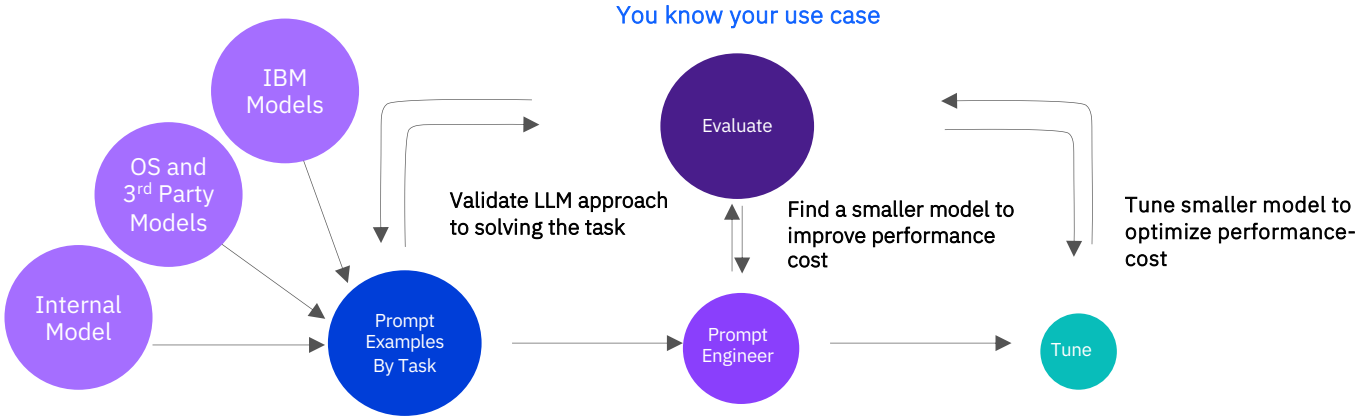
Read the customer review and write a short summary.



Text Summarization Metrics

- ROUGE
- SARI
- WIKI_SPLIT
- BLEURT
- METEOR
- Sentence Similarity - Jaccard Similarity
- Sentence Similarity - Cosine Similarity

Evaluation workflow



Proof-of-concept

Prove the use case using a large model with minimal labeled data. Evaluate.

Tuning

Use a medium-size model. Prompt engineer or prompt-tune with additional labeled data. Evaluate.

Selection

- Smallest model
- Additional data
- Fine-tune/prompt-tune
- Evaluate

Winning

Using the most accurate and cost-effective model

Thank you