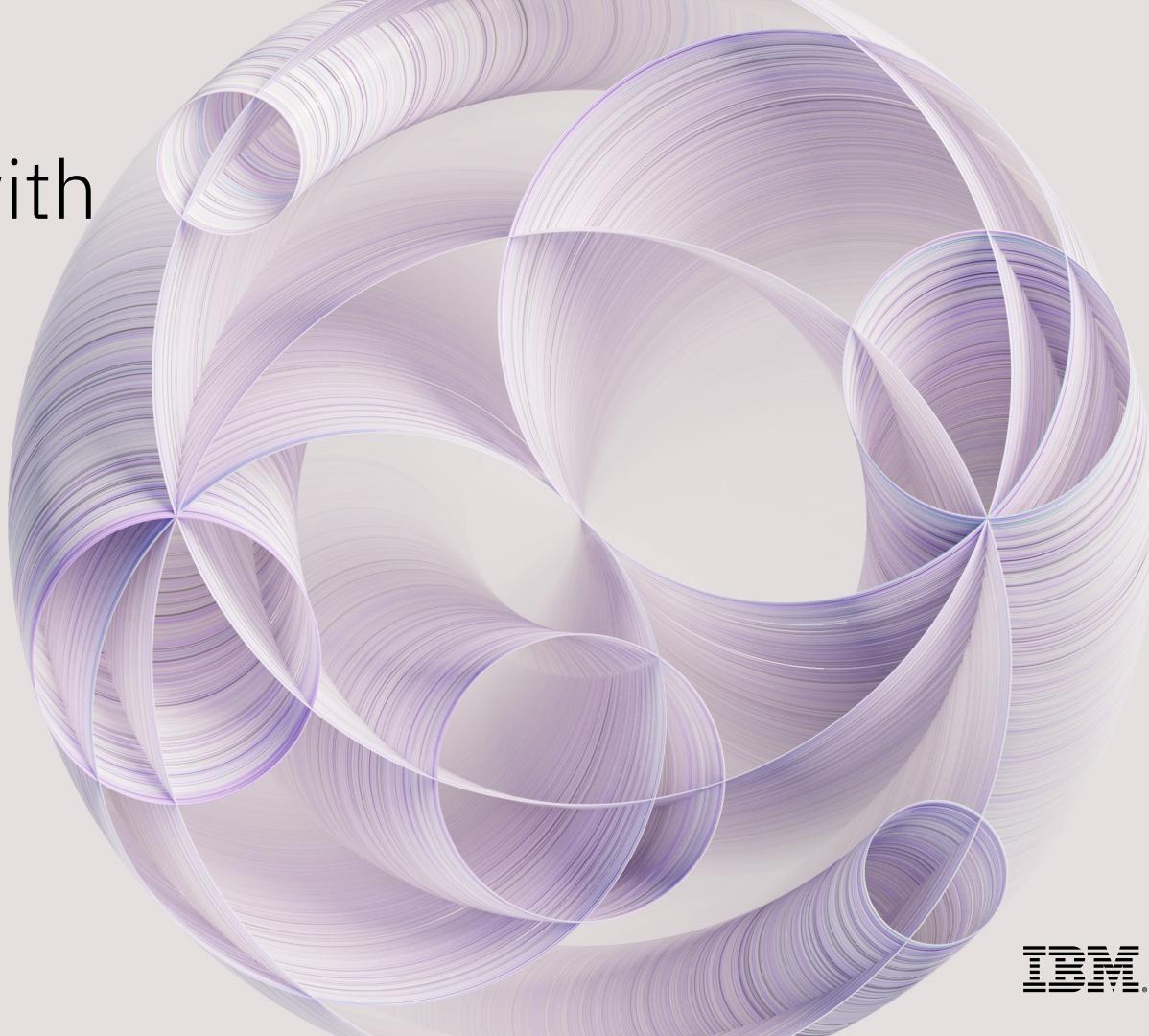


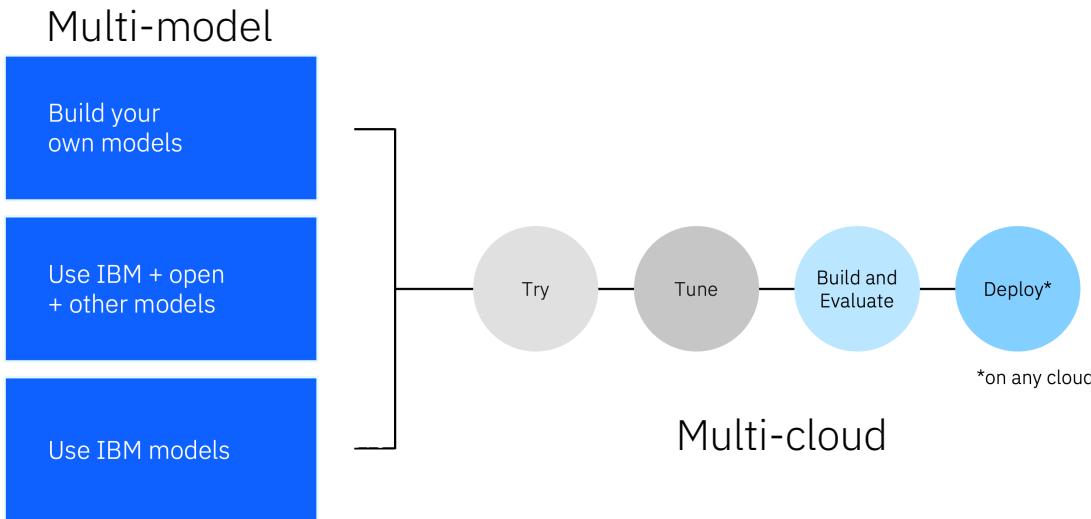
MLOps and Trustworthy AI with **watsonx**



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 productivity¹

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 assistant²

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 AI³

¹ IBM HR use case

²Vodafone 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



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.

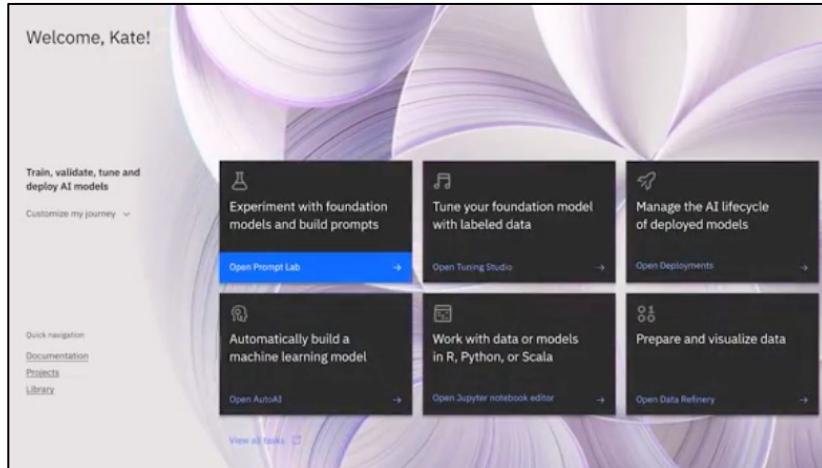
The screenshot shows the IBM WatsonX web interface. At the top, there's a navigation bar with links for 'Upgrade', 'IBM account', 'Dallas', and 'KB'. Below the navigation is a search bar labeled 'Search your workspaces'. The main content area is titled 'Select a foundation model' with the sub-instruction 'Select a model that best fits your needs. All provided models support English text. Check the model information for other supported languages.' A search bar 'Search for a model or task' is also present. The page displays a grid of eight foundation models, each with a diamond icon, name, provider, source, and a brief description. The models listed are:

Model	Provider	Source	Description
flan-ul2-20b	Provider Google	Source Hugging Face	flan-ul2 is an encoder-decoder model based on the T5 architecture and instruction-tuned using the Fine-tuned Language Net.
starcoder-15.5b	Provider BigCode	Source Hugging Face	The StarCoder models are 15.5B parameter models that can generate code from natural language descriptions.
mt0-xxl-13B	Provider BigScience	Source Hugging Face	An instruction-tuned iteration on mT5.
gpt-neox-20b	Provider Google	Source Hugging Face	A 20 billion parameter autoregressive language model trained on the Pile.
flan-t5-xl-3b	Provider Google	Source Hugging Face	A pretrained T5 - an encoder-decoder model pre-trained on a mixture of supervised / unsupervised tasks convert...
flan-t5-xxl-11B	Provider Google	Source Hugging Face	flan-t5-xxl is an 11 billion parameter model based on the Flan-T5 family.
granite-13b-chat-v1	Provider IBM	Source IBM	The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for generative...
granite-13b-chat-v2	Provider IBM	Source IBM	The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for generative...
granite-13b-instruct-v1	Provider IBM	Source IBM	The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for generative...
granite-13b-instruct-v2	Provider IBM	Source IBM	The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for generative...
mpt-7b-instruct2	Provider Mosaic, tuned...	Source Hugging Face	MPT-7B is a decoder-style transformer pretrained from scratch on 1T tokens of English text and code. This model was...
llama-2-70b-chat	Provider Meta	Source Hugging Face	Llama-2-70b-chat is an auto-regressive language model that uses an optimized transformer architecture.

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

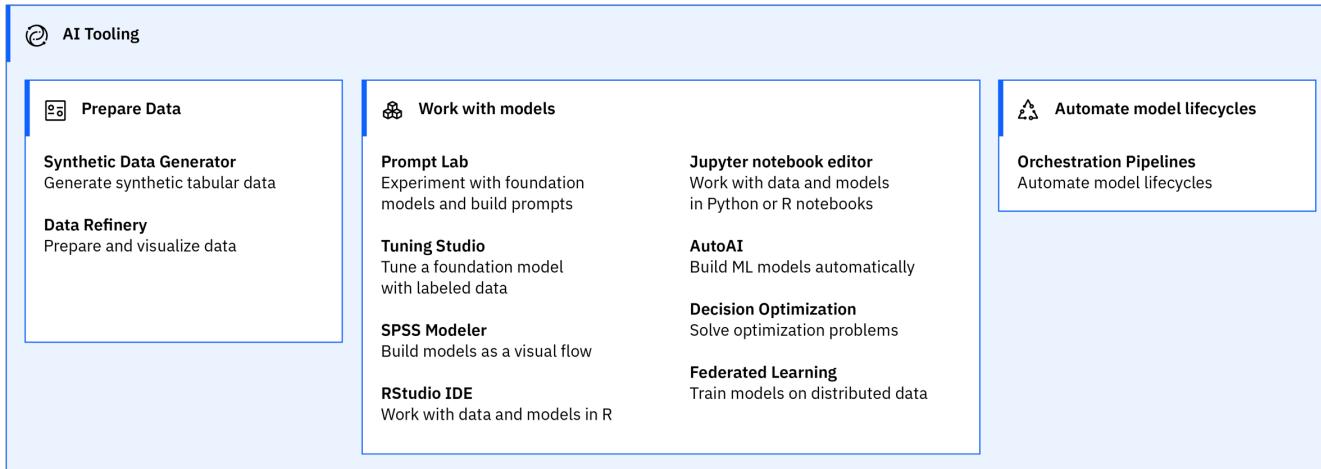
Prevent the model from generating repeating phrases

Number of min and max new tokens in the response

Stop sequences – specifies sequences whose appearances should stop the model

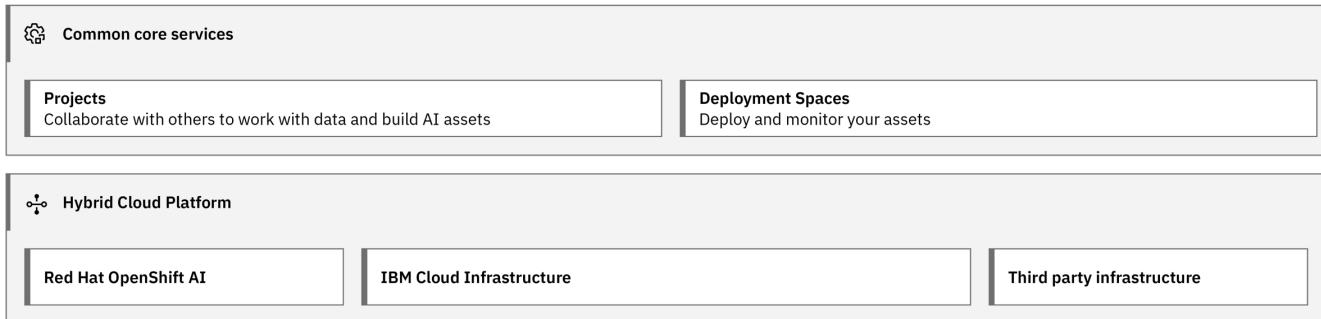
The screenshot shows the IBM WatsonX Prompt Lab interface. At the top, there's a navigation bar with 'IBM watsonx', a search bar, and various account and location options. The main area is titled 'Prompt Lab' and shows a workspace named 'Kate's sandbox'. In the 'Sample prompts' section, several options are listed, including 'Summarization' (with a 'Summarize financial highlights on an earnings call.' example), 'Meeting transcript summary' (selected, with a 'Summarize the discussion from a meeting transcript.' example), 'Scenario classification' (with a 'Classify scenario based on project categories.' example), 'Sentiment classification' (with a 'Classify reviews as positive or negative.' example), 'Marketing email generation' (with a 'Generate email for marketing campaign.' example), 'Thank you note generation' (with a 'Generate thank you note for workshop attendees.' example), 'Named entity extraction' (with a 'Find and classify entities in unstructured text.' example), 'Fact extraction' (with a 'Extract information from SEC 10-K sentences.' example), and 'Question answering' (with a 'Answer questions about an article.' example). Below this, the 'Set up' section has an 'Instruction (optional)' field containing 'Write a short summary for the meeting transcripts.' and an 'Examples (optional)' section showing two rows of transcript and summary pairs. The first row is for a meeting transcript, and the second row is for a marketing email. At the bottom, there's a 'Try' section with a 'Test your prompt' input field containing '1' and a summary output: 'John Doe 00:00:01.415 --> 00:00:20.675'. A 'Generate' button is at the bottom right.

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](#)

The screenshot shows the IBM Watson Tuning Studio interface. At the top, there's a navigation bar with 'IBM watsonx', 'Upgrade', 'Eric Saleh's Account', 'Dallas', and a search bar. Below the navigation is a header for 'Demo Tuning Experiment' with a save timestamp of 'November 16, 2023 at 4:52:49 PM'. The main area is divided into several sections: 'Configure details' (specifying the foundation model as 'flan-t5-xl-3b' and initialization method as 'Random'), 'Add training data' (a file named 'file_to_tune.jsonl' of size 1.56 KB), 'Configure parameters' (with sliders for 'Maximum input tokens' set to 256 and 'Maximum output tokens' set to 128), and a summary section 'What should your data look like?' which includes a 'Preview template' button and token count controls.

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

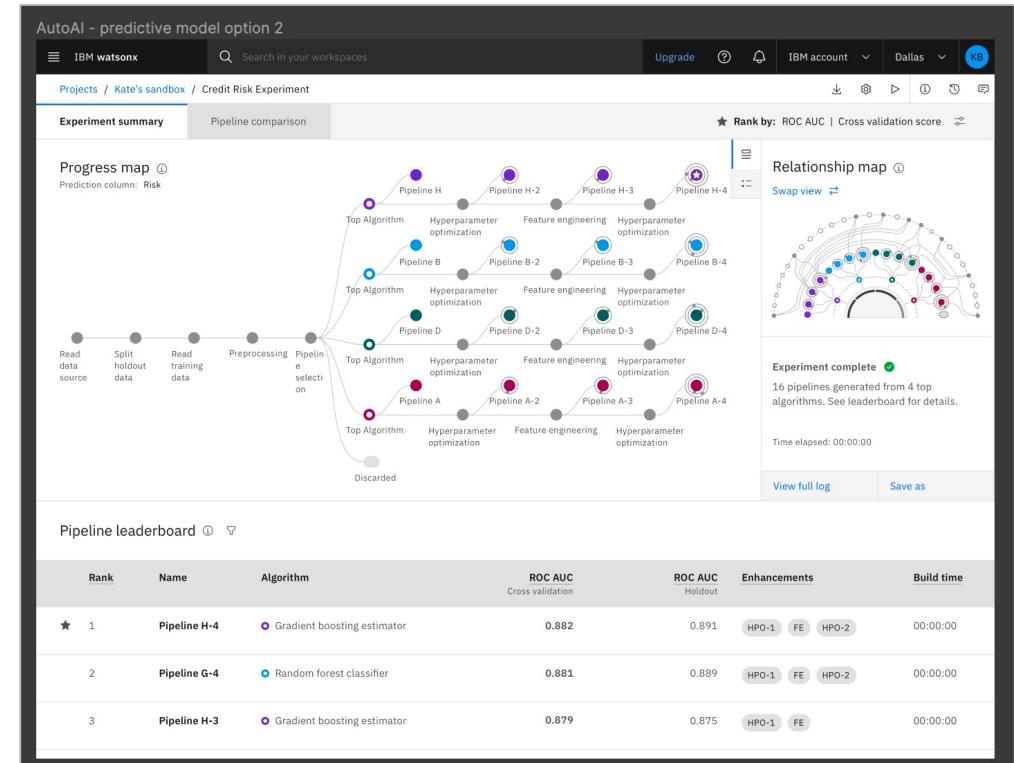
AutoAI, including preparing data for machine learning and generating and ranking candidate model pipelines

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.

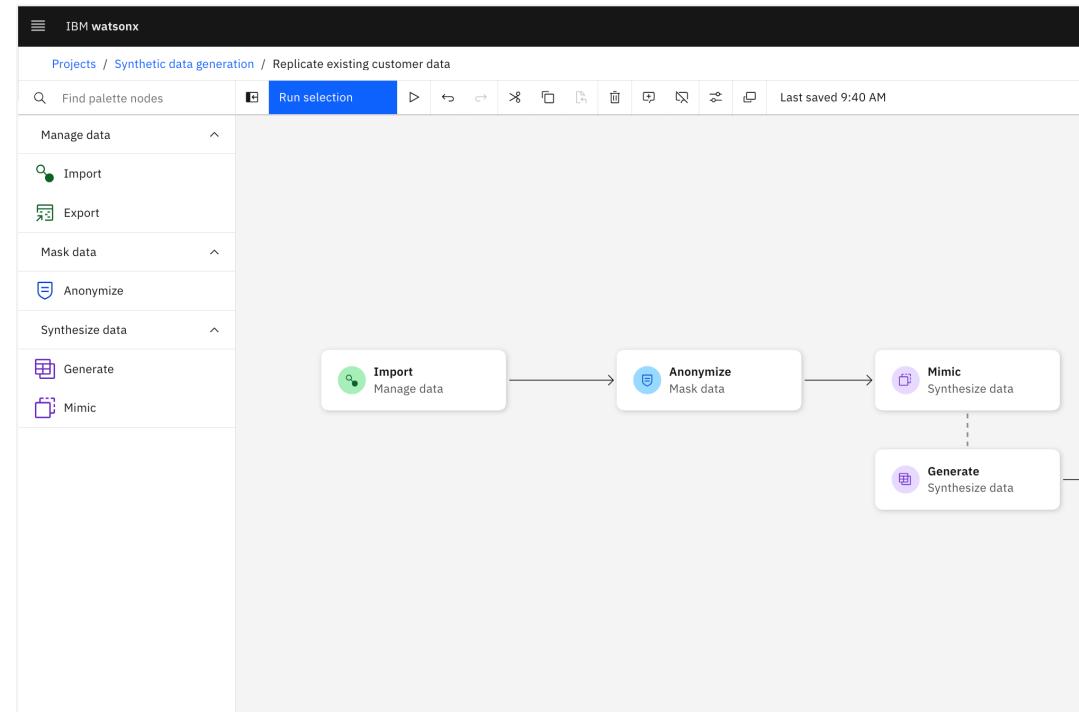
Create synthetic data using your existing data in a database or by uploading a file. If no data exists or can't be accessed, you can design your own data schema.

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.

Comprehensive

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

Risk Management

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

Open

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

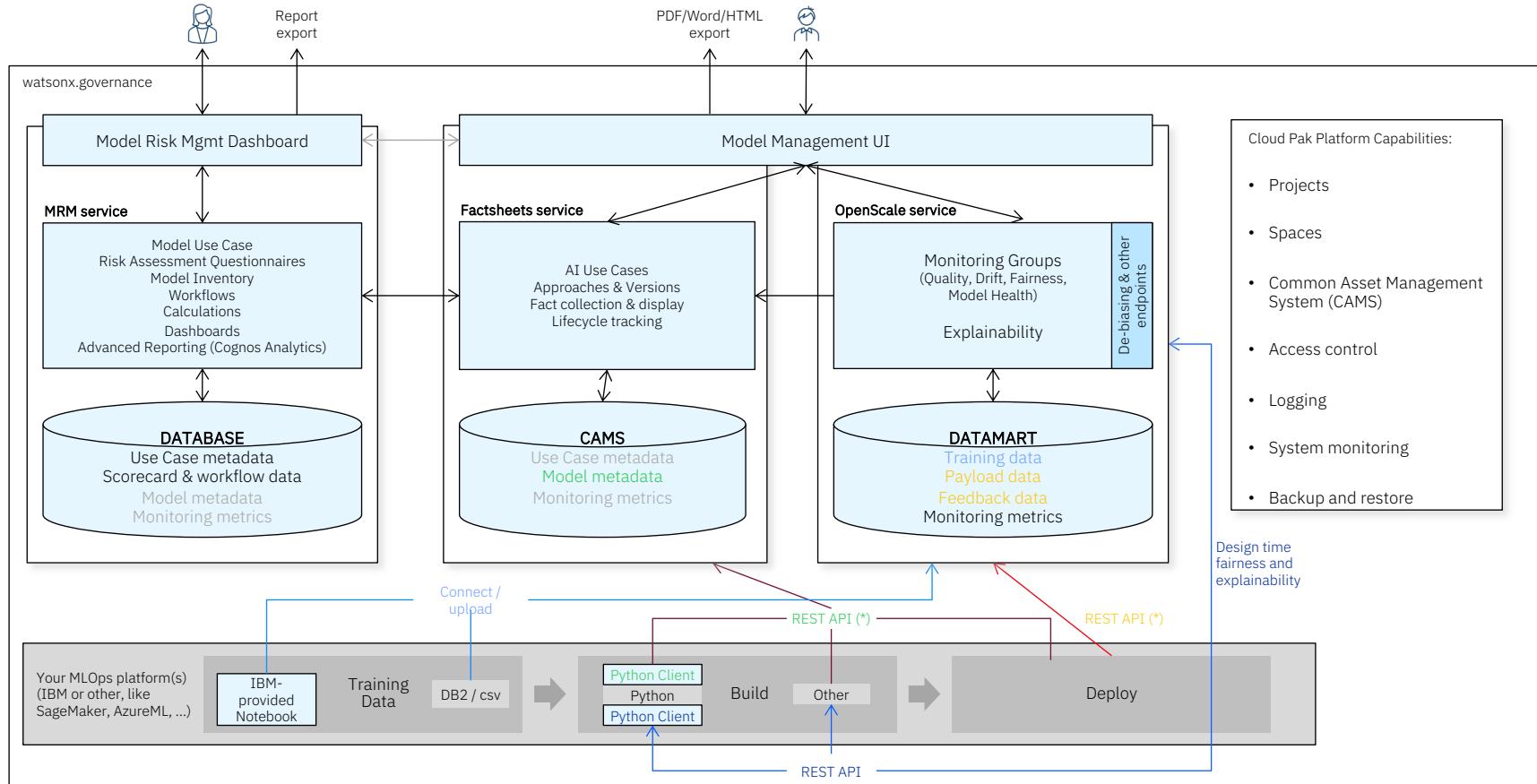
Regulatory Compliance

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

Automatic metadata recording

and data transformation/lineage capture though Python notebooks.

watsonx.governance – functional architecture (software)

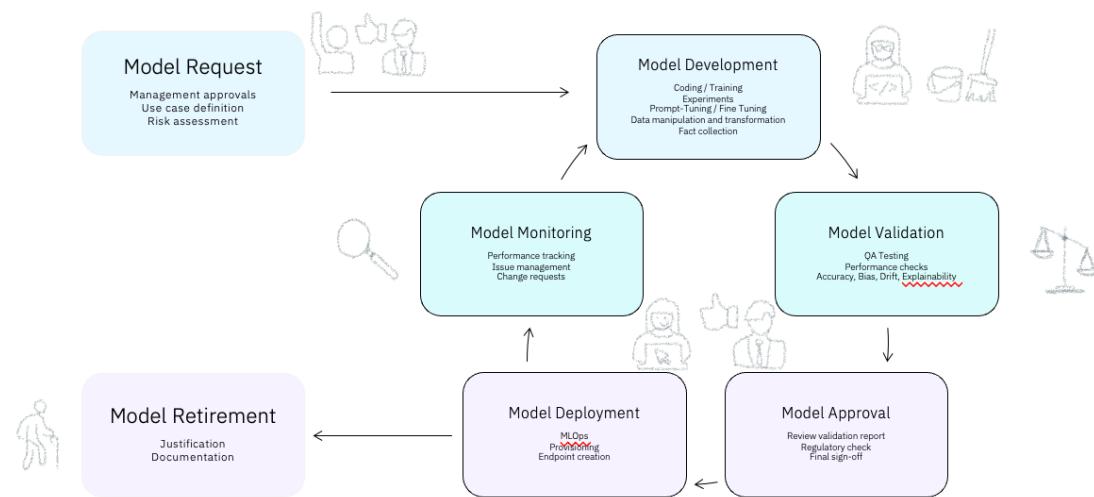


watsonx.governance

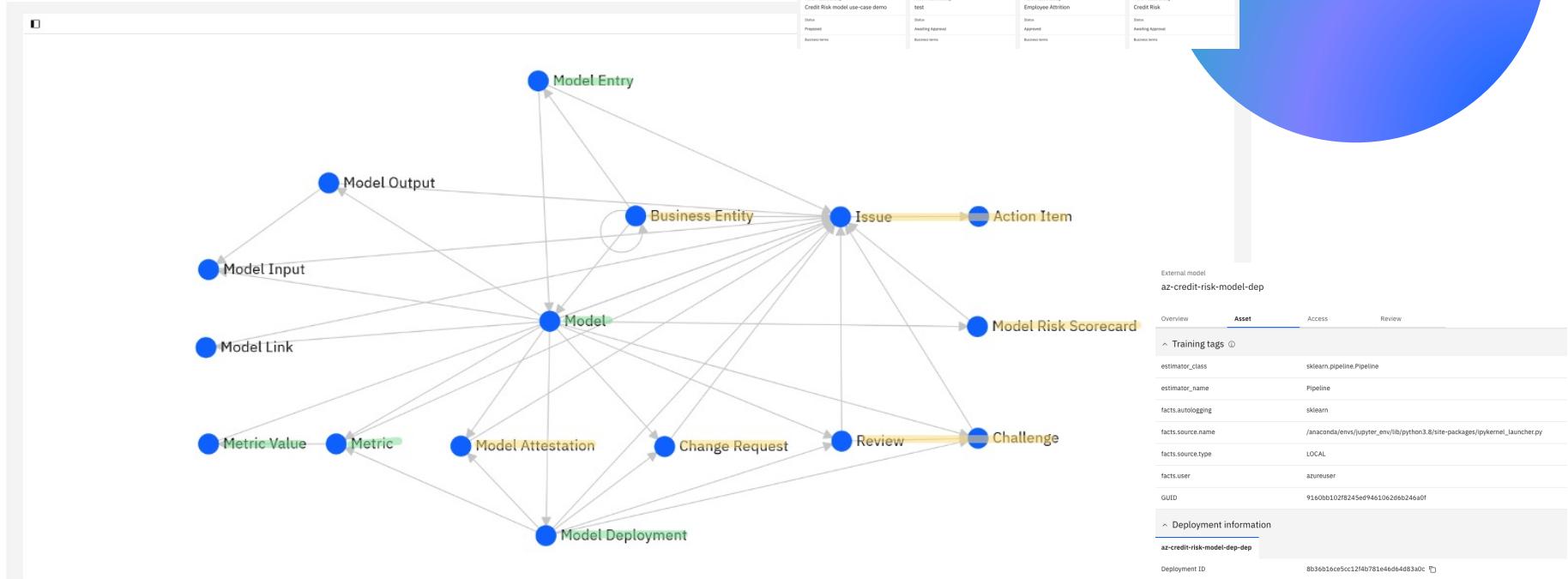


- Consolidated view of models from multiple platforms
- View development status, model performance and alerts or emerging issues
- Monitor and trigger workflows for model validation, retraining and performance issues

The screenshot shows the IBM OpenPages platform. At the top, there's a navigation bar with links like Home, Loading..., Model Entries, Workflows, AI Model Prop..., Workflow In..., Loan Autom..., Discard Draft, and Publish. Below the navigation is a specific workflow titled "AI Model Proposal" with steps: Start → Capture Model Details → Notify Usecase Approver → Reject. To the right of the workflow is a "Workflow Properties" table with columns for Field, Status, Mapping, Value, and Criticality. The table lists four rows: Open (Status: Open), Awaiting Approval (Status: Awaiting Approval), Closed (Status: Closed), and (Default) (Status: Open). On the left, there's a dashboard with sections for Model Overview, Model Requests, Model Reviews and Validation, Model Health Status, and Model Risk by Breakdown.



AI Governance object model *Much more than just “a model”*



what is the difference between a fixed and variable rate mortgages

Enter a question to seek an answer

Respond in Seek Personalize

Highlight Answer Provenance On Sentiment

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

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 fad934f-8e48-4231-a718-131...

Fairness 90% Green within threshold

Quality .99 Green within threshold

Fairness by feature

Age	90%
Sex	91%
Race	92%

Quality metrics

Area under ROC	.91
Area under PR	.91
Accuracy	.91
True positive rate (TPR)	.91
False positive rate (FPR)	.09
Recall	.91

TRANSACTION Transaction details

Feature influence analysis with SHAP

Predicted outcome 0.0

Confidence level 88.91%

Average model confidence on baseline 94.33%

How this prediction was determined

The Type II Diabetes model - P4 XGBoost model has been deployed online. This means that the outcome of this transaction would be predicted by running it through the model to analyze the top features influencing the model's prediction.

Type II Diabetes Deployment (Online) the prediction

Background data

Training background data

Transaction

MAPK_0998191-080c-4377-a7f7-

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:

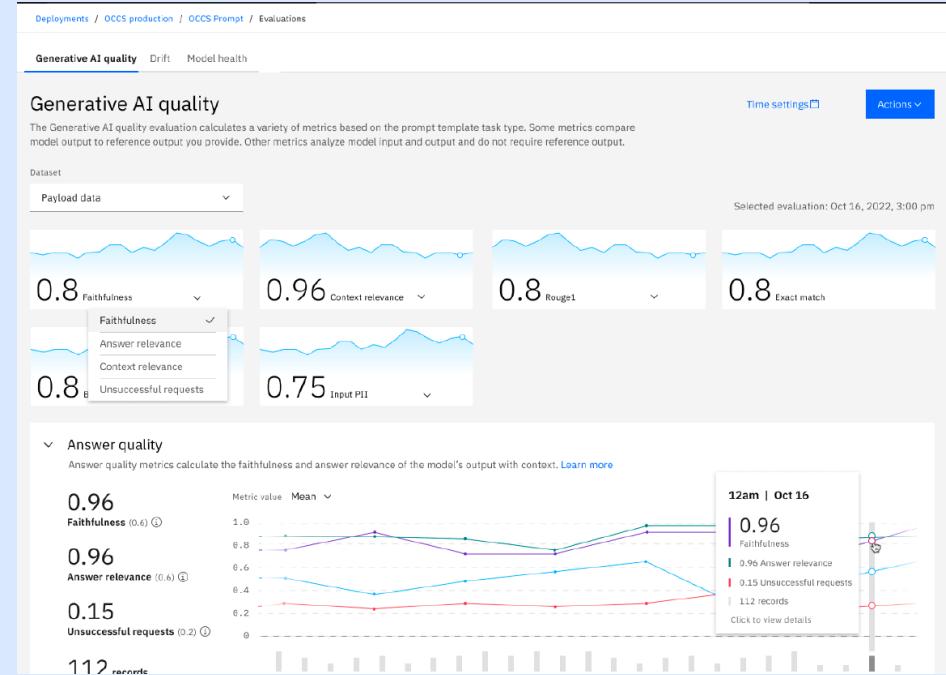
AI Engineers and users that are developing, validating, and deploying RAG to understand the quality of responses

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**

Entity Extraction Metrics

- [Seq eval](#)
- [Language Translation Evaluation Metrics](#)
- [Character](#)
- [charcut_mt](#)
- [Chrf](#)
- [google_bleu](#)
- [super_glue](#)
- [TER](#)
- [nist_mt](#)
- [Poseval](#)
- [sacrebleu](#)
- [XTREME-S](#)

20

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

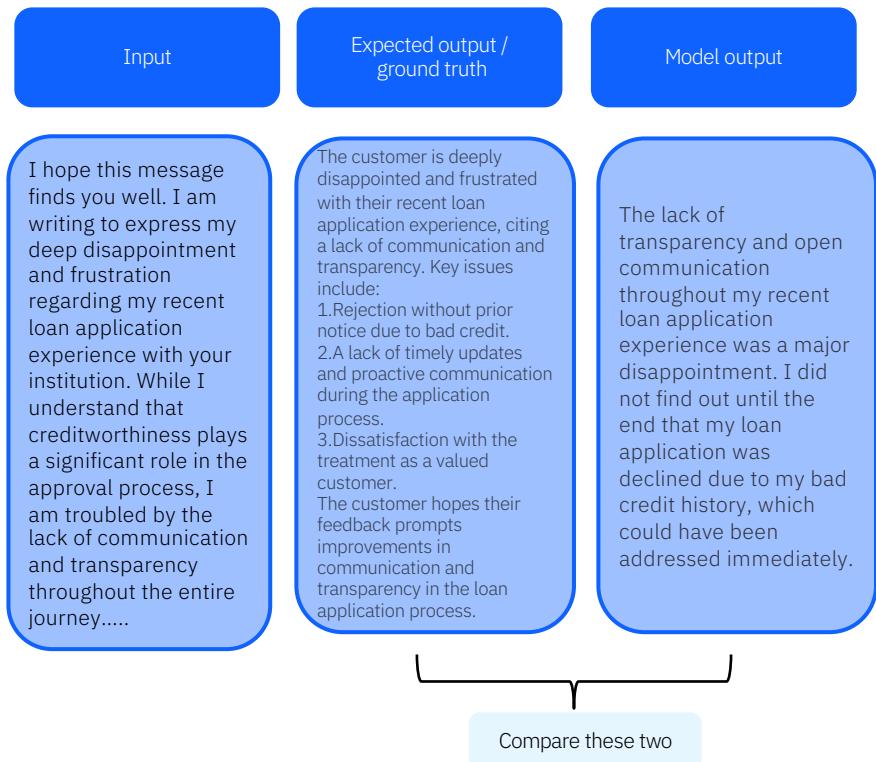
Under Exploration from SVL Research Labs

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

** Uses Metrics Computation Model

Summarization Example

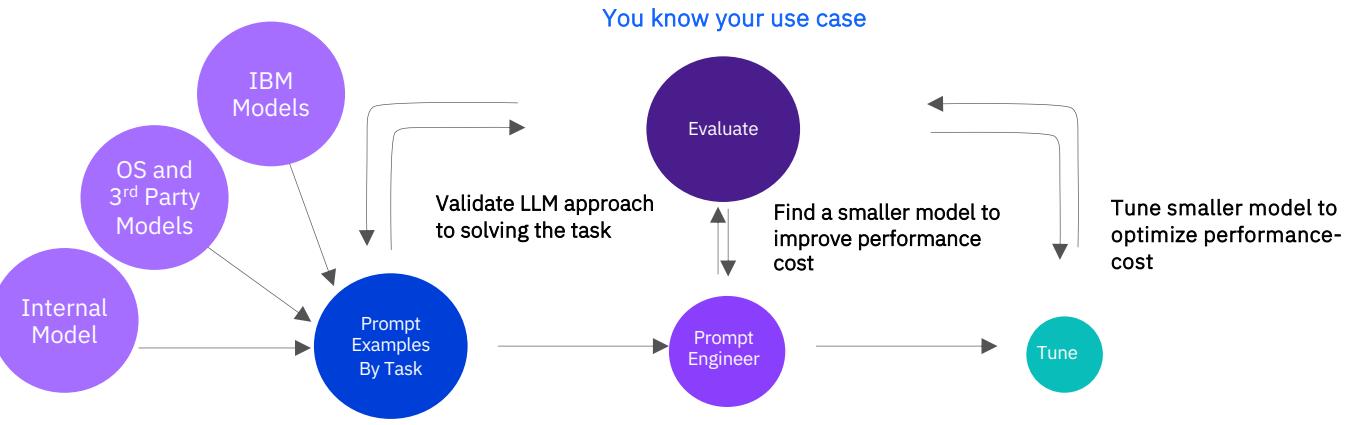
Read the customer review and write a short summary.



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

© 2024 International Business Machines Corporation
IBM and the IBM logo are trademarks of IBM
Corporation, registered in many jurisdictions
worldwide. Other product and service names might be
trademarks of IBM or other companies. A current list
of IBM trademarks is available on ibm.com/trademark.

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.

Red Hat and OpenShift are registered trademarks of
Red Hat, Inc. or its subsidiaries in the United States
and other countries.

