watsonx.ai

Train, validate, tune and deploy AI models

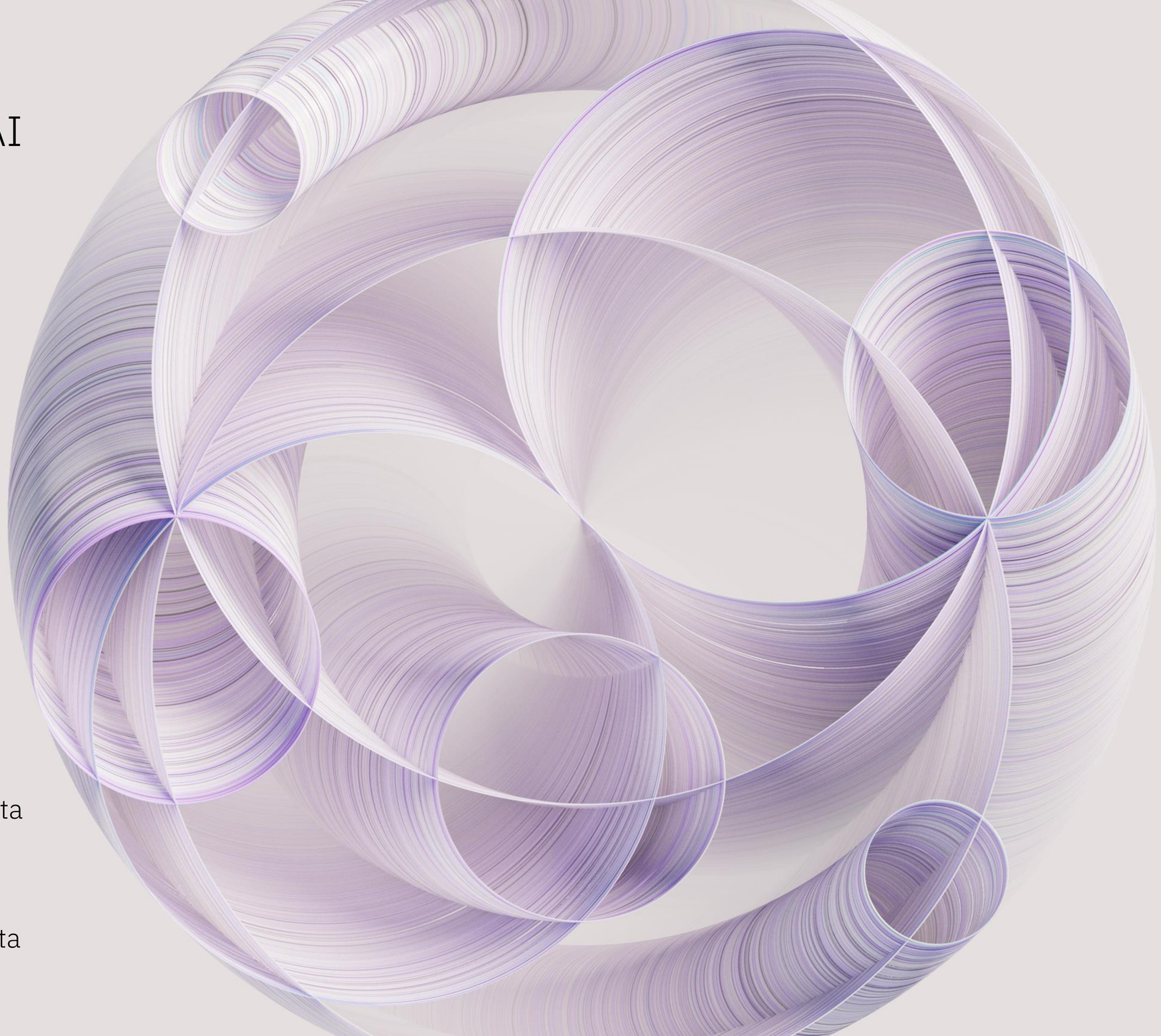
Client presentation

Linsay Wershaw
Lindsay.Beth.Wershaw@ibm.com
Senior Product Marketing Manager, IBM **watsonx.ai**

Angela Jamerson
Angela.Jamerson@ibm.com
Program Director, Product Management, **watsonx.ai**

Felix Lee
felix@ca.ibm.com
Principal, Learning Content Development, AI and Data

Anshupriya Srivastava
anshupriya.srivastava@ibm.com
Advisory, Learning Content Development, AI and Data



Contents

Foundation Models and Generative AI
are bringing an inflection point in AI...

...but how enterprises adopt and execute will define whether they unlock, create value, unleash innovation at scale and with speed

Generative AI and traditional AI

Both traditional AI and generative AI are useful for enterprises.

Neither replaces the other, generative AI opens new possibilities

Generative AI

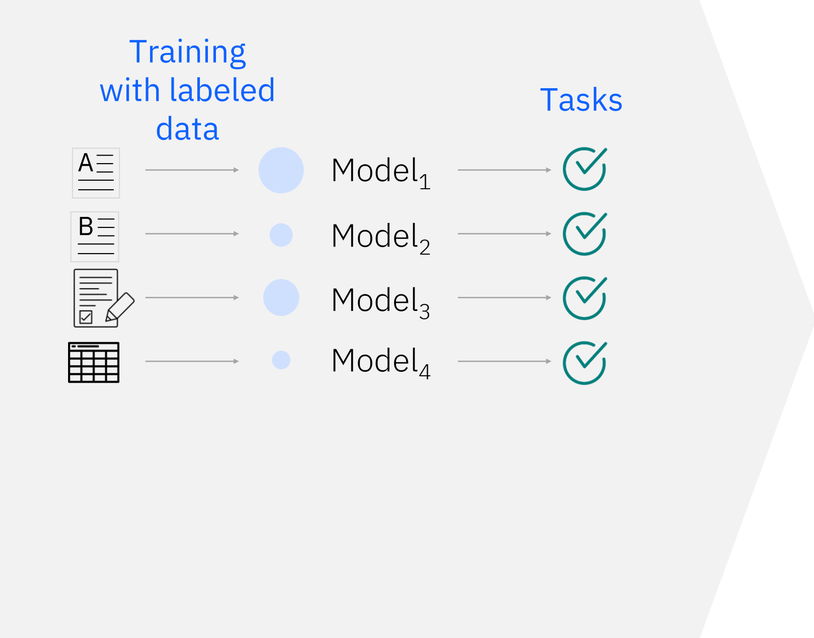
- Foundation models trained with unlabeled data
- Unsupervised
- Trained on very big data sets
- No specific task
- Transferable
- Works well for general tasks and can improve for specific tasks with less training
- Need to monitor bias and drift

Traditional AI

- Traditional Machine learning (ML/AI) model trained with “labeled” data
- Training is supervised
- Trained on proper, large data sets
- Trained for a specific task
- Does not transfer well to other tasks
- A tuned model can be very efficient for the specific task it was designed for
- Need to monitor bias and drift

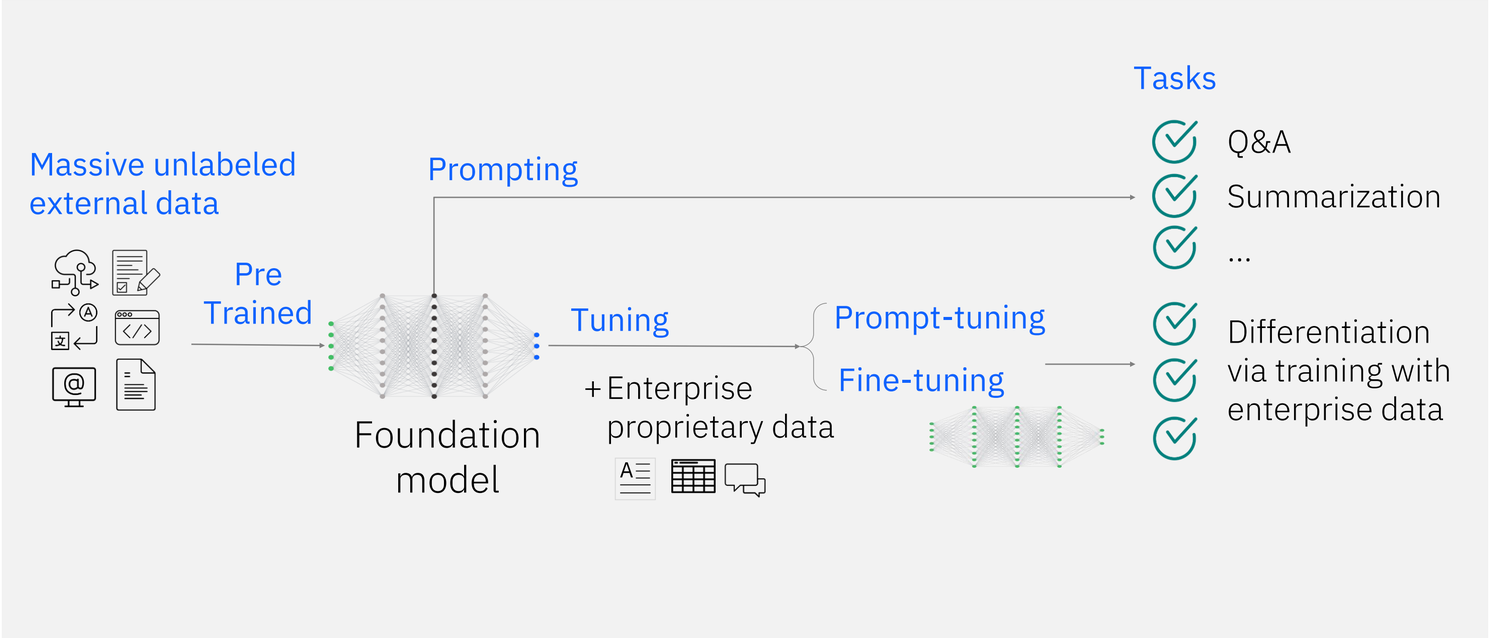
Foundational models enable a new paradigm of data-efficient AI development – generative AI

Traditional AI models



- Individual siloed models
- Require task specific training
- Lots of human supervised training

Foundation Models



- Rapid adaptation to multiple tasks with small amounts of task-specific data
- Pre-trained unsupervised learning

Impact of generative AI

Scale of impact points to swift adoption over next 3 years

Global economic boost

$3-4T

forecasted economic benefits to the global economy across industries

Massive early adoption

81%

of enterprises are working with or planning to leverage foundation models and adopt generative AI

Broad-reaching and deep impact

7%

Potential rise in global GDP within 10 years

Critical focus on AI investment

70%

of software vendors will integrate Gen AI in their enterprise applications by 2026

Sources: Mckinsey; Goldman Sachs; IBM Institute for Business Value; Gartner. Scale Zeitgeist: AI Readiness Report, a survey of more than 1,600 executives and ML practitioners

Five truths of generative AI

Truth 1

Multi-model

Two thirds of 150+ enterprises surveyed report pursuing a multi-model strategy

- 60% + of enterprises pursuing multi-model are experimental with commercial & open-source models
- Commercial & open-source innovation
- Quickly prioritize use cases that will outlive the model
- Multi-modal (text, image, audio, etc.)
- One model will not rule them all

Truth 2

Multi | hybrid cloud

Gartner reports that most enterprises will deploy generative AI across hybrid / multicloud environments

- Run where the workflows, apps and data live
- Infer where business runs to drive performance, cost, and simplicity
- Data location to drive security benefits
- Regulatory compliance to influence location selection

Truth 3

Governance

Surveyed companies report governance as a top requirement, impact of generative AI makes governance more difficult

- Businesses must control bias and monitor drift
- Organizations must actively monitor hallucinations and ensure model explainability
- Leaders must seek practices and tools to ensure model and data provenance

Truth 4

Scale for value

Critical to pick the right use cases and deployment for generative AI ROI

- Different work tasks have strongly positive or negative ROI impact
- Time savings for a meaningful product innovation +40%; business problem solving -23% time needed
- 60+ points difference in value for work tasks
- 25x difference in cost per inference, depending on model and deployment

Truth 5

Data matters

Generative AI pilots have not made it to production due to challenges with data quality, access, and security

- Short run: model innovation creates value
- Long run: data quality will decide which enterprises win with generative AI

Most common generative AI tasks implemented today

Summarization

Transform text with domain-specific content into personalized overviews that capture key points.

Conversation summaries, insurance coverage, meeting transcripts, contract information

Classification

Read and classify written input with as few as zero examples.

Sorting of customer complaints, threat and vulnerability classification, sentiment analysis, customer segmentation

Generation

Generate text content for a specific purpose.

Marketing campaigns, job descriptions, blog posts and articles, email drafting support

Extraction

Analyze and extract essential information from unstructured text.

Medical diagnosis support, user research findings

Question-answering

Create a question-answering feature grounded on specific content.

Build a product specific Q&A resource for customer service agents.

Generative AI adoption considerations, inhibitors and fears

80% of business leaders see at least one of these ethical issues as a major concern

Explainability

Believe decisions made by generative AI are not sufficiently explainable.

Bias

Believe that generative AI will propagate established biases.

Ethics

Concerned about the safety and ethical aspects of generative AI.

Trust

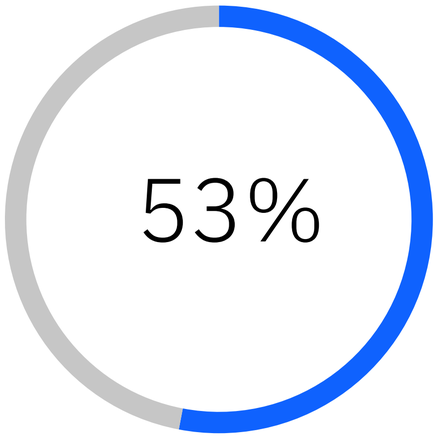
Believe generative AI cannot be trusted.

Source: IBM IBV “generative AI: The state of the market”, June 2023

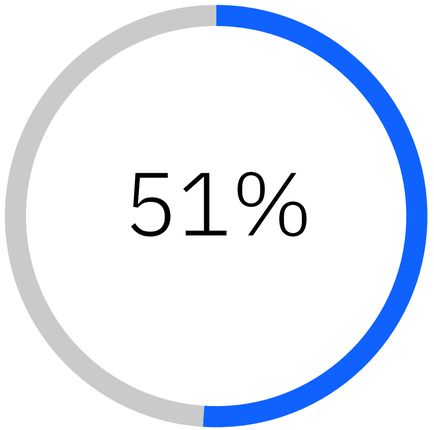
Generative AI adoption barriers

Executives highlight three top barriers to implementing generative AI

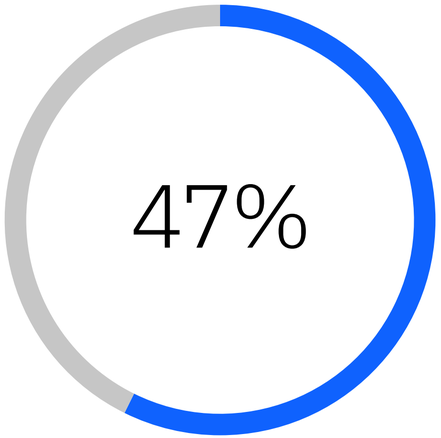
Cybersecurity



Privacy



Accuracy



Enterprises need more than an AI solution - they need a comprehensive and sound strategy for generative AI.

Source: IBM IBV “generative AI: The state of the market”, June 2023

Generative AI must be tailored to the enterprise

Open

→ Based on the best AI and cloud technologies available.

→ Giving access to the innovation of the open community and multiple models.

Trusted

→ Offering security and data protection.

→ Built with Governance, transparency, and ethics that support increasing regulatory compliance demands.

Targeted

→ Designed for targeted for business use cases, that unlock new value.

→ Models that can be tuned to your proprietary data and company guidelines.

Empowering

→ A platform to bring your own data and AI models that you tune, train, deploy, and govern.

→ Running anywhere, designed for scale and widespread adoption.

watsonx.ai

Put AI to work with **watsonx**

Scale and accelerate the impact of AI with trusted data on hybrid cloud

watsonx.ai	watsonx.data	watsonx.governance
Train, validate, tune and deploy AI models	Scale AI workloads, for all your data, anywhere	Enable responsible, transparent and explainable data and AI workflows

Red Hat OpenShift provides scalability, hybrid capability

watsonx

and its 3 components

The platform for AI and data

Scale and accelerate the impact of AI with trusted data.

watsonx.ai

Train, validate, tune and deploy AI models

A next generation enterprise studio for AI builders to 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

Enable responsible, transparent and explainable AI workflows

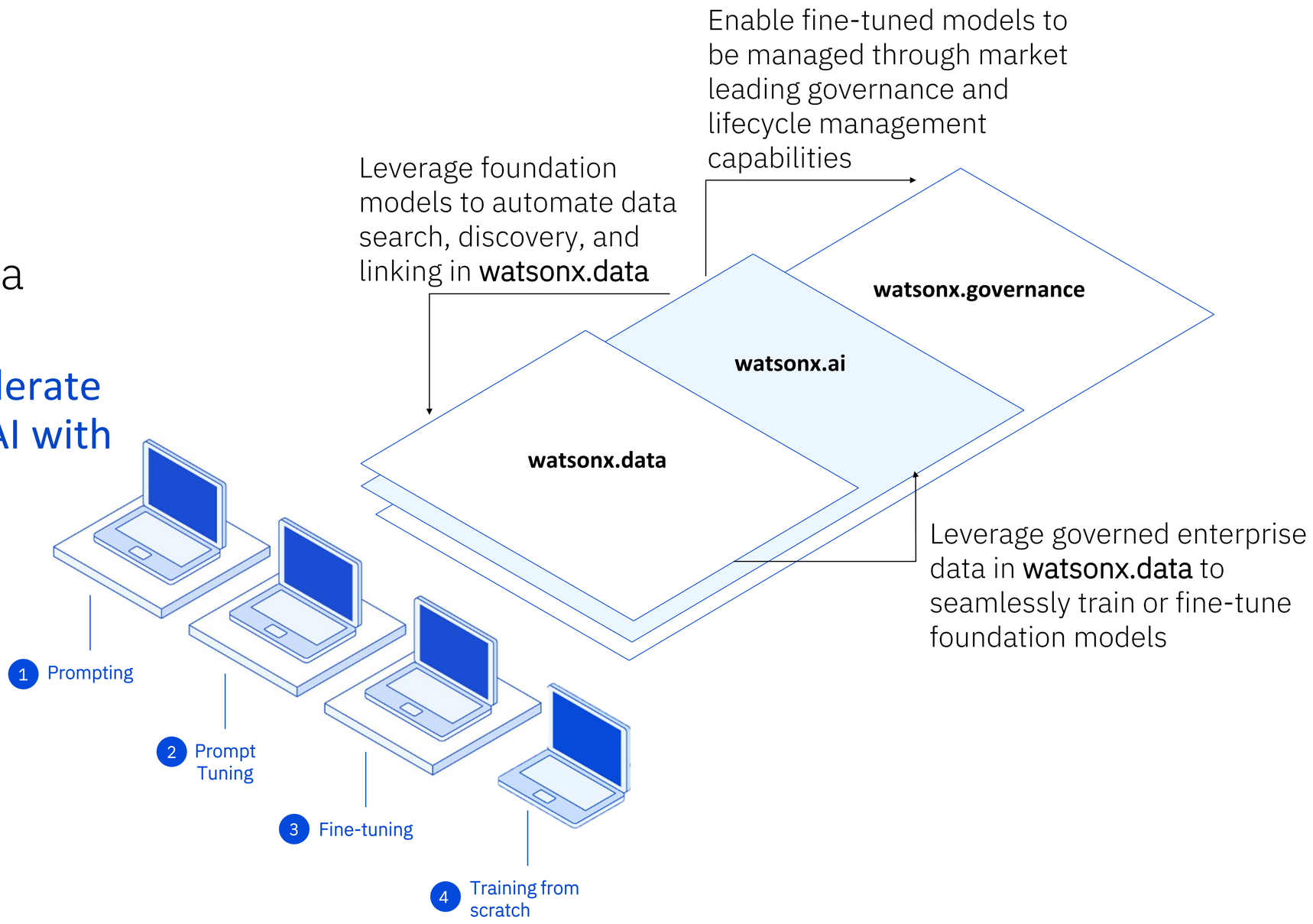
End-to-end toolkit encompassing both data and AI governance to enable responsible, transparent, and explainable AI workflows.

watsonx

and its 3 components

The platform for AI and data

Scale and accelerate the impact of AI with trusted data.



watsonx.ai

Clients can train, validate, tune, and deploy their AI models

Bring together AI builders

- Open-source frameworks
- Tools for code-based, automated, and visual data science capabilities
- All in a secure, trusted studio environment

Accelerate the full AI model lifecycle

- All the tools and runtimes are in one place to train, validate, tune, and deploy AI models.
- Hybrid and multicloud enabled

Leverage foundation models & generative AI

- Train with a fraction of the data, in less time, and with fewer resource
- Leveraged advanced prompt-tuning capabilities
- Full SDK and API libraries.

watsonx.ai – generative AI with traditional AI features

Train, validate, tune, and deploy AI models with confidence

Generative AI capabilities

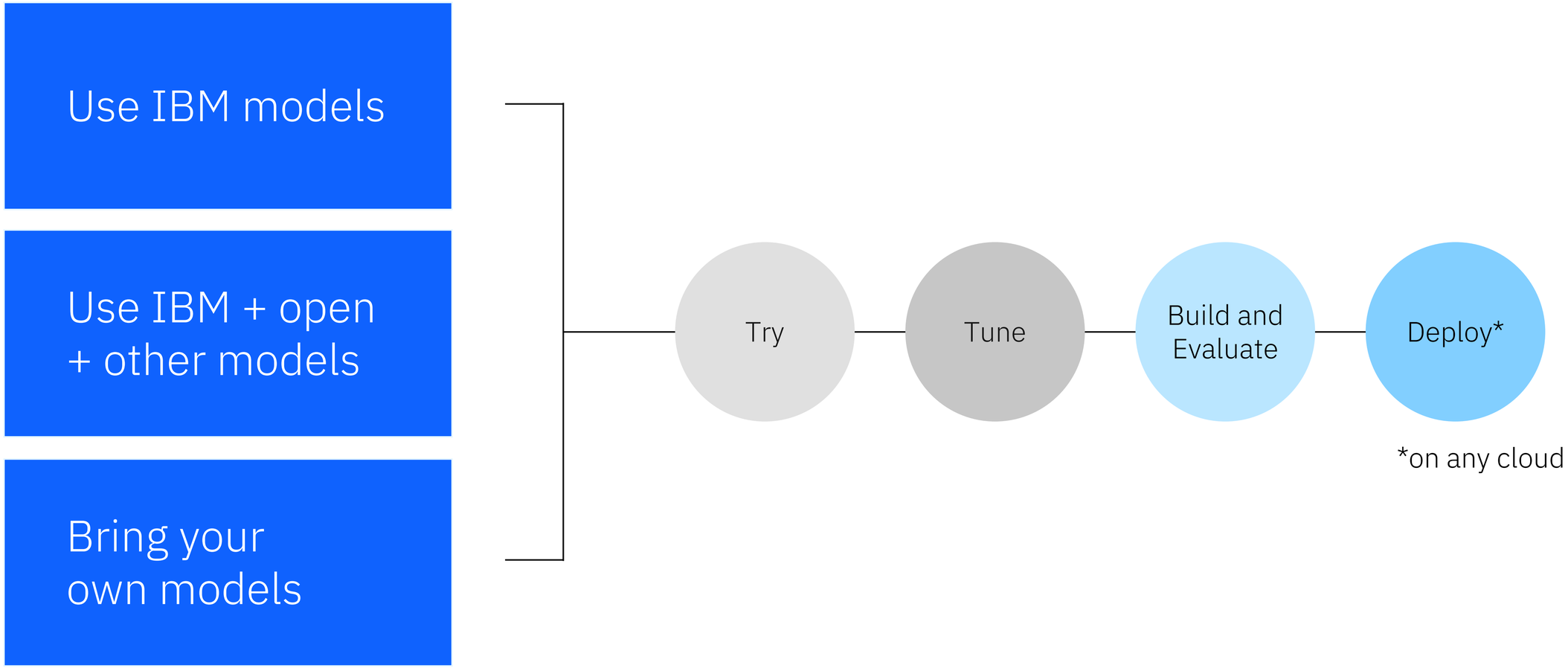
- Foundation model library
- Prompt lab
- Tuning studio

Plus, a proven studio for machine learning

- ModelOps
- Automated development
- Decision optimization

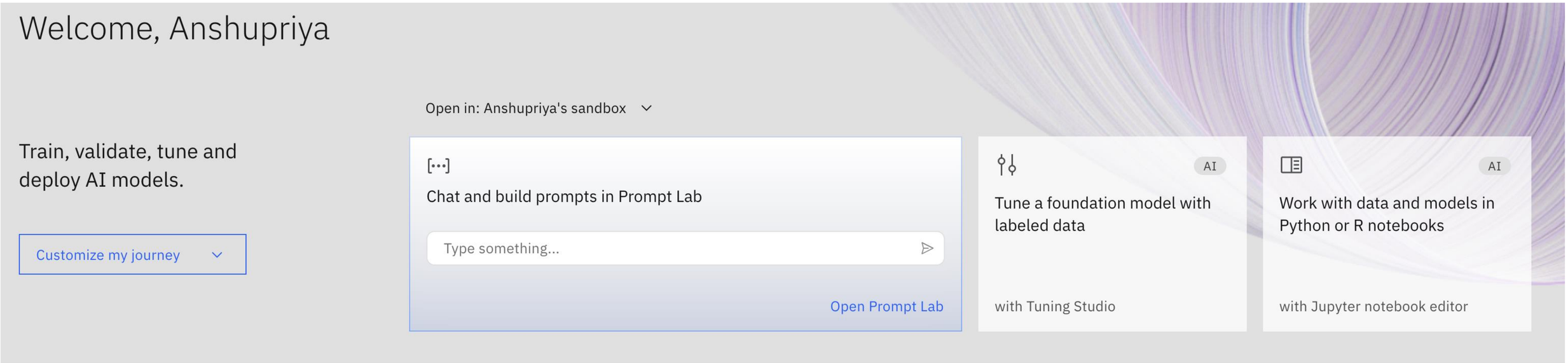
Team collaboration and data preparation

watsonx.ai is based on foundation models that are multi-model on multi-cloud with no lock-in



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

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 Foundation Model Library

Model variety to cover enterprise use cases and compliance requirements

IBM models

IBM's suite of foundation models is designed to ensure model trust and efficiency in business applications. Our suite of models features:

Transparent Pre-Training on IBM's trusted Data Lake

- One of the largest repositories of enterprise-relevant training data
- Verified legal and safety reviews by IBM
- Full, auditable data lineage available for any IBM Model

Compute-Optimal Model Training and Architectures

- Granite
  Decoder only transformers
- Slate
  Encoder only transformers
- Sandstone
  Encoder-decoder transformers

Efficient Domain and Task Specialization

Models Coming Soon:

- Finance
- Cybersecurity
- Legal, etc.

Opensource models

Experiment with open source models

IBM and Hugging Face partnership demonstrates our shared *commitment to delivering to clients an open ecosystem approach* that allows them to define the best models for their business needs.

Bring-your-own-model

Optional add-on for more flexibility
Partner with IBM Research to pre-train your own foundation models.

watsonx.ai – Foundation Models available

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 tasks.

Provider: IBM
Type: InstructLab

mt0-xxl-13b

An instruction-tuned iteration on mT5.

Provider: BigScience
Type: Provided model

flan-t5-xl-3b

A pretrained T5 - an encoder-decoder model pre-trained on a mixture of supervised / unsupervised tasks converted into a text-to-tex...

Provider: Google
Type: Provided model

flan-t5-xxl-11b

flan-t5-xxl is an 11 billion parameter model based on the Flan-T5 family.

Provider: Google
Type: Provided model

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
Type: Provided model

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 tasks.

Provider: IBM
Type: Provided model

granite-20b-multilingual

The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for generative tasks.

Provider: IBM
Type: InstructLab

granite-7b-lab

The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for generative tasks.

Provider: IBM
Type: InstructLab

llama-2-13b-chat

Llama-2-13b-chat is an auto-regressive language model that uses an optimized transformer architecture.

Provider: Meta
Type: Provided model

llama-2-70b-chat

Llama-2-70b-chat is an auto-regressive language model that uses an optimized transformer architecture.

Provider: Meta
Type: Provided model

llama-3-70b-instruct

Llama-3-70b-instruct is an auto-regressive language model that uses an optimized transformer architecture.

Provider: Meta
Type: Provided model

llama-3-8b-instruct

Llama-3-8b-instruct is an auto-regressive language model that uses an optimized transformer architecture.

Provider: Meta
Type: Provided model

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

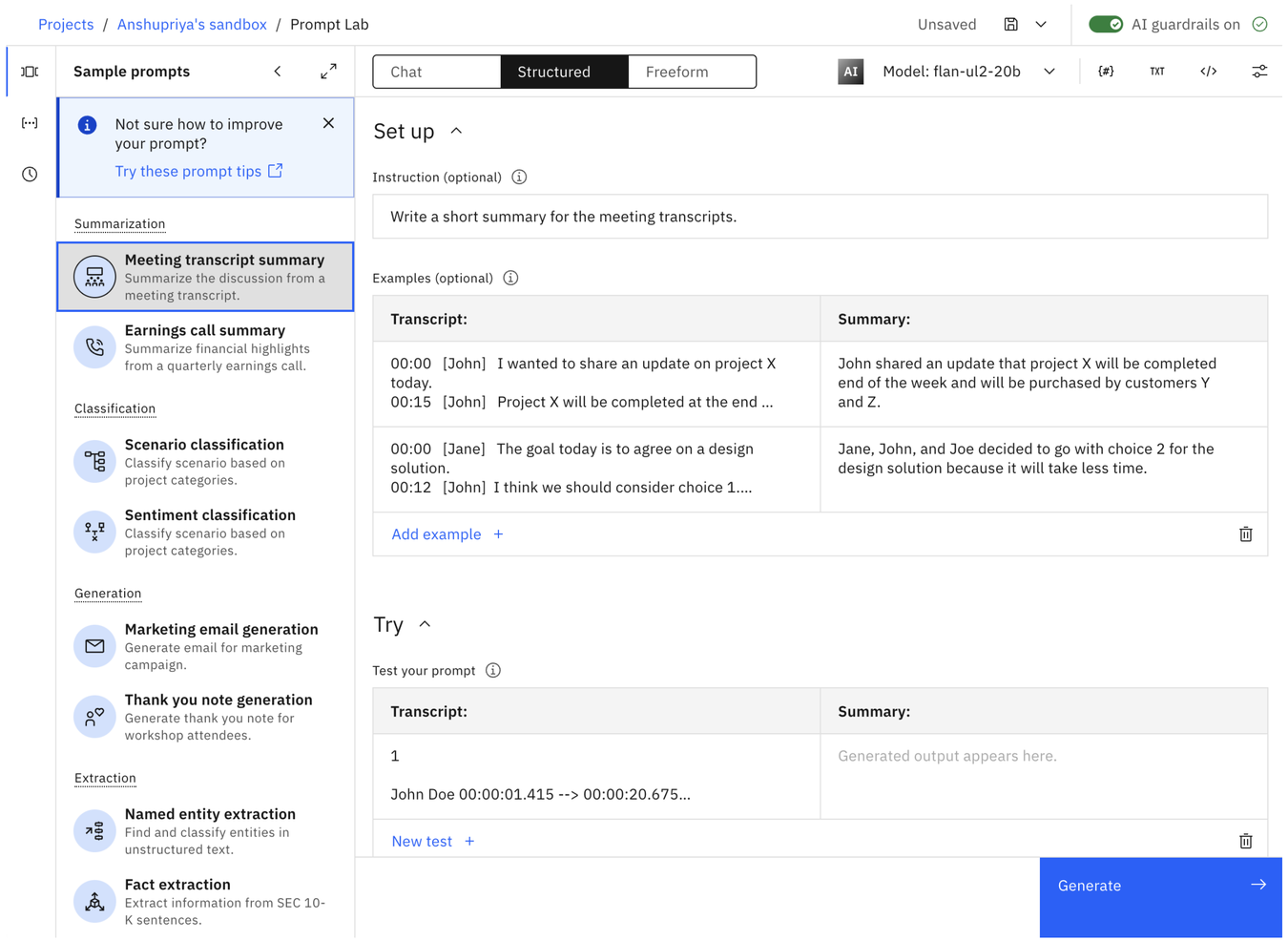
Experiment with prompt engineering

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



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

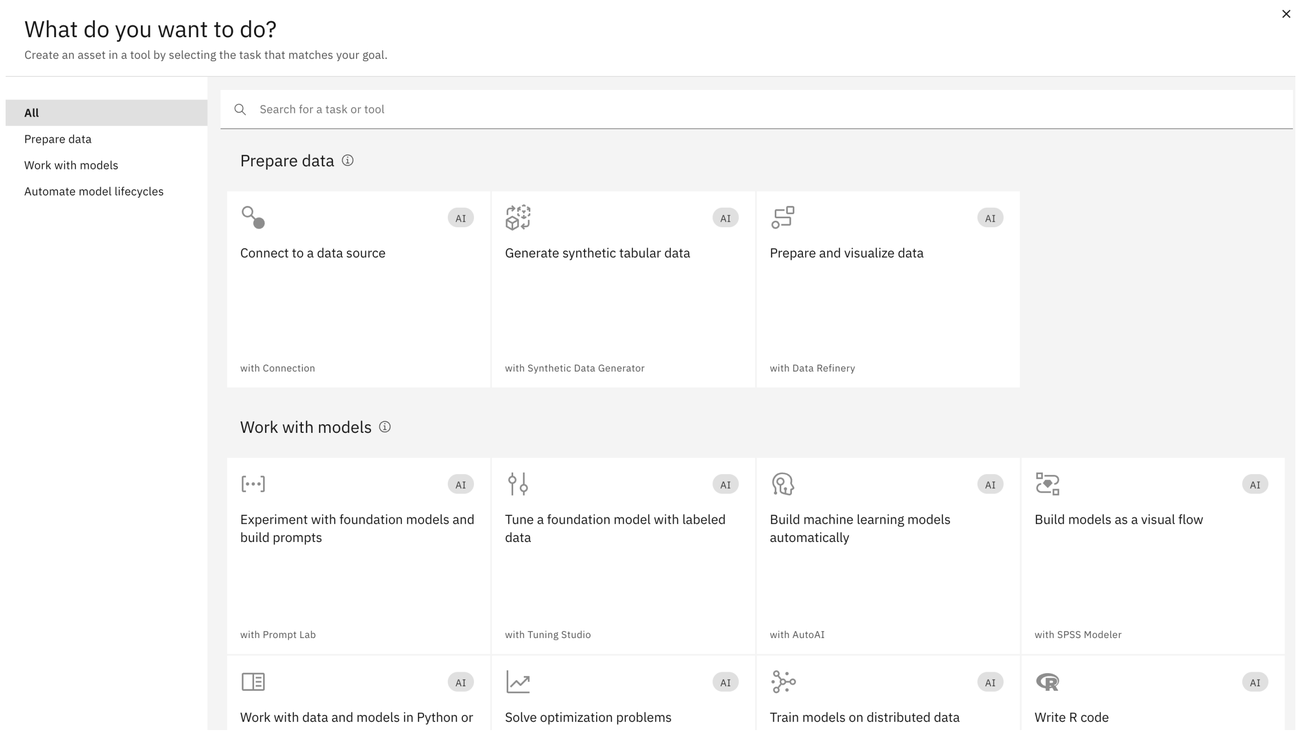
AutoAI, including preparing data for machine learning and generating and ranking candidate model pipelines

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: Tuning Studio

Tune your foundation models with labeled data

Prompt tuning

Efficient, low-cost way of adapting an AI foundation model to new downstream tasks

Tune the prompts with no changes to the underlying base model or weights

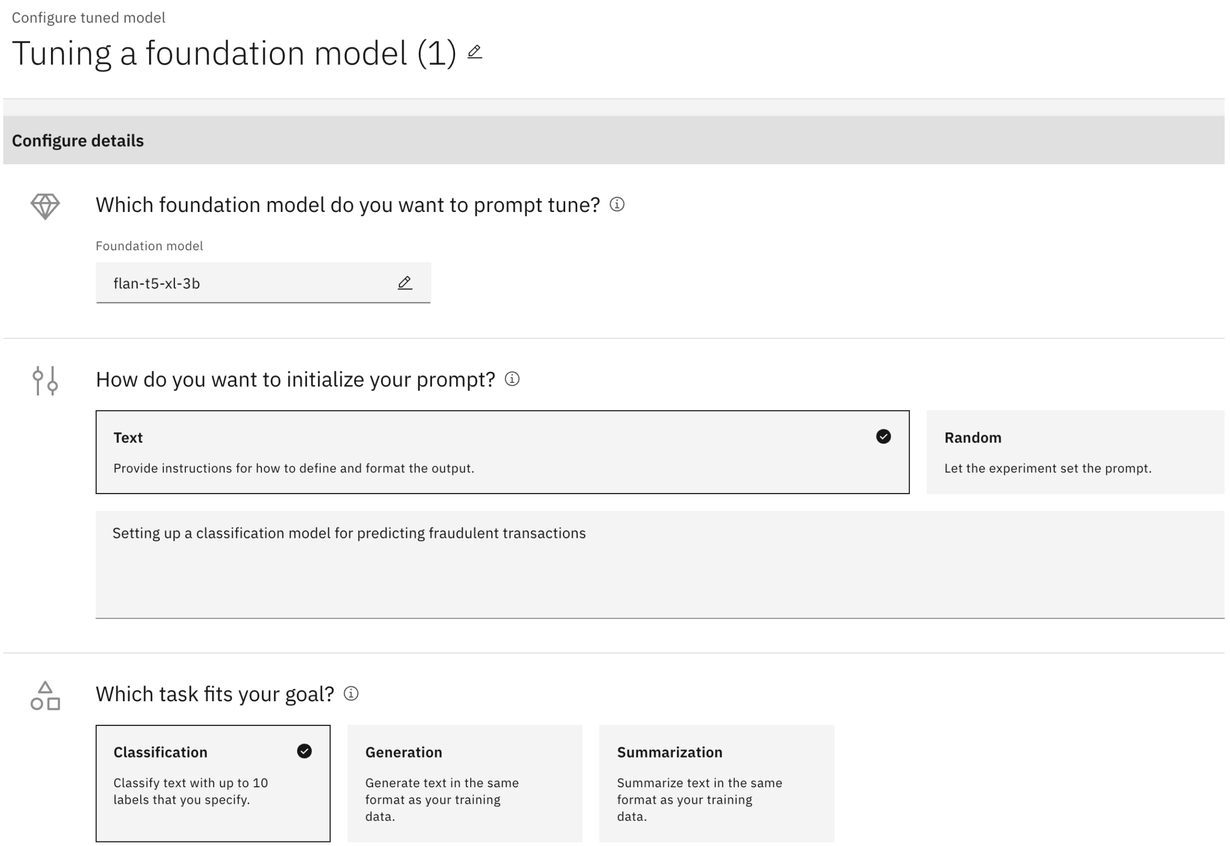
Unlike prompt engineering, prompt tuning allows clients to further enhance the model with focused, business data

Task support in the Tuning Studio

Models support a range of Language Tasks: Q&A, Generate, Extract, Summarize, Classify

Requires a small set of labelled data to perform specialized tasks

Can achieve close to fine-tuning results without model modification, at a lower cost to run



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.

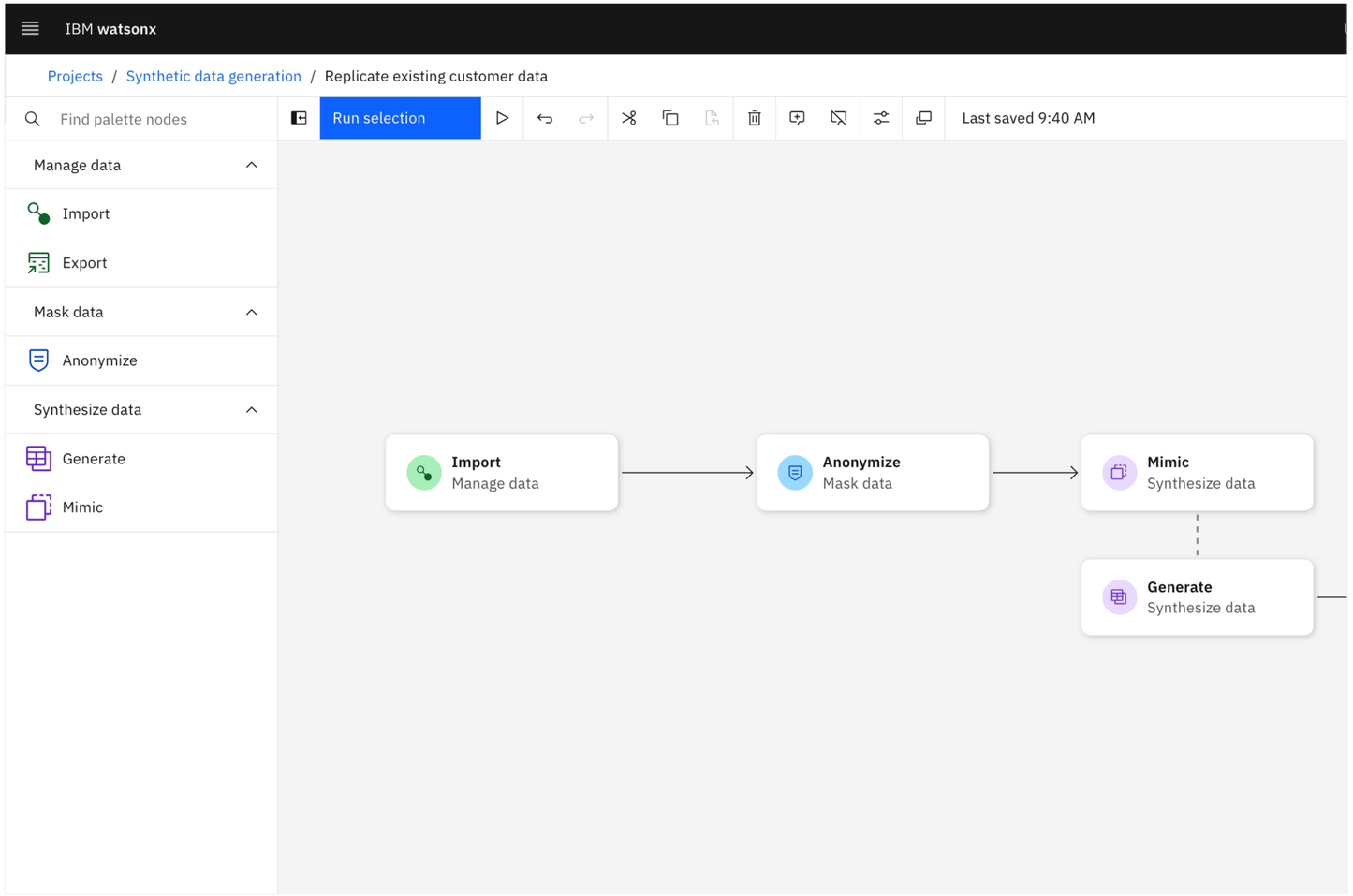
Create synthetic data using your existing data in a database or by uploading a file. If no data exists or can't be accessed, you can design your own data schema.

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

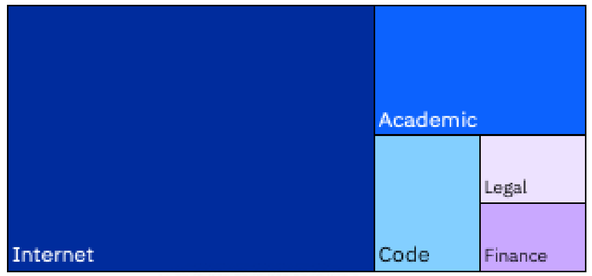
AI for business - IBM Granite

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.

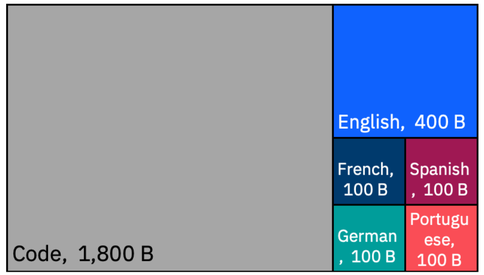
granite-13b-v2 (English LLM)
-chat-v2.1, -instruct-v2

13B parameters in size
2.5T tokens of data



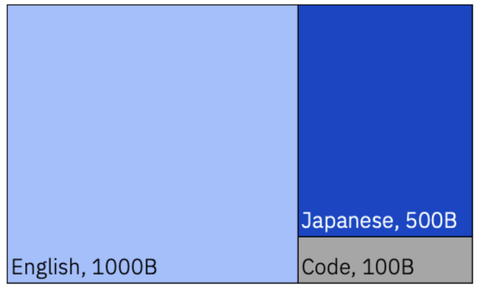
granite-20b-multilingual

20B parameters in size
2.6 T tokens of Data



granite-8b-japanese

8B parameters in size
1.6T tokens of Data



watsonx.ai value proposition

Improved performance

- Developing specialized models to produce better results for targeted tasks with lower infrastructure requirements to achieve improved price-performance, *(granite.13b for financial tasks).*

- Enhancement of models delivered through model refresh (granite.13b.V2), new models developed by IBM (e.g., granite.20b multilingual), or 3rd
- party models

3x Price-cuts

- granite.13b [3X less cost] available today at $0.0006 1,000 tokens (input/output)
- llama2.70b [2.7X less cost] available today at $0.0018 1,000 tokens (input/output)
- Llama2 13b [3X less cost] available today at $0.0006 1,000 tokens (input/output)

Multi-lingual support

Expanding language support beyond English through a combination of 3rd party model providers and IBM-developed multi-lingual models that support:

- English
- Japanese
- Spanish
- Portuguese
- French
- German

Differentiated Client Protection

IBM stands behind IBM-developed models and indemnifies the client against third-party IP claims.
IBM offers an additional peace of mind to clients by:

- not requiring them to indemnify IBM for their use of its models
- not capping its IP indemnification liability

watsonx.ai differentiators

Open

- **Built on open technologies**
  - IBM's hybrid cloud-native stack based on Red Hat OpenShift enables a flexible and secure deployment of **watsonx.ai.**
  - Hugging Face partnership provides access to the best open-source model collection.

Trusted

- IBM's suite of foundation models is designed to **ensure model trust** and efficiency in business applications.
- Models trained with scrutinized and copyright-free data
- Tight integration with **watsonx.governance** provides clients with a **trusted pathway** to operationalize AI confidently and at scale.

Targeted

- Designed for **targeted business use cases**, that unlock new value.
  - On-prem, hybrid cloud and IBM Cloud
  - Designed for scalability
  - Right model for the right task
- **Industry-leading support** for use case implementations.

Empowering

- For **value creators**, not just users
  - Tunable models at a fraction of the cost & time
  - Deploy anywhere
- An enterprise studio that allows clients build their own differentiated AI assets with their own proprietary data, creating a competitive edge.

watsonx.ai is transparent, responsible, and governed

Most AI models are trained on datasets of unknown quality, representing legal, regulatory, ethical, and inaccuracy nightmares. Data provenance and quality matters. **IBM ensures its AI can be trusted.**

watsonx.data

- Curates domain-specific and internet datasets, as well as ingesting your own
- Filters for hate, profanity, biased language, and licensing restrictions before training
- Tracks and manages every step of the process to meet legal and regulatory requirements

watsonx.governance

- Governs training data and the AI deployed
- Applies reinforcement learning with human feedback to align models with human values, reduce hallucinations, and build AI guardrails
- Finds and fixes AI biases before ML AI models are tuned and deployed

IBM's Center of Excellence for Generative AI

Over 1,000 IBM Consultants specialized in generative AI help you establish an organization to adopt and scale AI safely, detect and mitigate risks, and provide education and guidance

watsonx.ai is helping companies custom-build AI solutions to suit their specific needs.

Sicredi

Leveraged **watsonx.ai** foundation models to train their AI augment its support representatives' efforts. Built a system to improve customer service and employee satisfaction using the IBM watsonx.ai™, IBM Watson® Discovery and IBM watsonx Assistant solutions

SAMSUNG SDS

Exploring watsonx.ai generative AI capabilities for new solutions such as SDS's Zero Touch Mobility to deliver unprecedented product innovations to improve client experience.

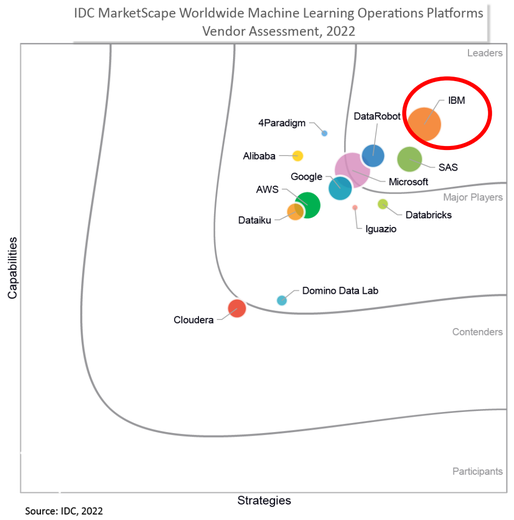
TechD

Using **watsonx.ai** to slash delivery time from 3-4 months down to 3-4 weeks for many customer care use cases.

Seismic

An early adopter of generative AI, has been exploring **watsonx.ai** to improve content discoverability, summarization and classification of data to enhance productivity.

IBM is a leader in AI



IDC Marketscape:
Leader in Worldwide
Machine Learning
Operations Platforms
2022 Vendor Assessment

MQ for Cloud
AI Developer Service

MQ for Enterprise
Conversational AI Platforms

MQ for Insight Engines

Multiple Gartner Magic Quadrants
for AI-related capabilities

Forrester Wave:
Multimodal Predictive
Analytics and
Machine Learning

How to get started with **watsonx.ai** today

IBM's investment in partnering with you

FREE TRIAL

Experience **watsonx.ai** yourself with a free trial through ibm.com/watsonx.

Try our free trial

CLIENT BRIEFING

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

PILOT PROGRAM

Watsonx.ai pilot develop with IBM Client Engineering and IBM Consulting to prove the solution's value for the selected use case(s) with a plan for adoption.

1-4 weeks

Backup

Supervised and Self Supervised Learning ↷

What's the difference?

Supervised learning	Self-supervised learning
Human powered	Computer powered
Requires intense labeling	Requires little labeling
Long, hard, expensive	Quick, automated, and efficient

Leveraging foundation model capabilities across various domains

	Customer Care Watson Assistant, Cloud Pak for Data	Digital Labor Watson Orchestrate, Cloud Pak for Integration/Automation, Wisdom in Ansible	IT Operations Turbonomic, Instana, Cloud Pak for Watson AIPOs	Cybersecurity QRadar, Cloud Pak for Security
Summarization Summarizing large documents, conversations, and recordings to key takeaways	• Call center transcripts • Omnichannel journey summary • Summarizing search snippets to augment chatbots • Summarize events, analyst reports, financial info etc. for advisor • Sentiment analysis	• Summarize documents, contracts, technical manuals, reports, etc. • Transcribe videos to text and summarize • Summarizing reports on Form 10K	• Summarize alerts, technical logs, tickets, incident reports, etc. • Summarize policy, procedure, meeting notes, etc. • Vendor report QBR summarization	• Summarize security event logs • Summarize steps to recap security incident • Summarize security specs
Extraction Extract structured insights from unstructured data	• Extracting interaction history with clients • Extract information from specific types/categories of incidents	• Extract answers and data from complex unstructured documents • Extract information from media files such as meeting records, audio, and video	• Extract key information from various sources for report automation • Extract relevant system/network information for administration, maintenance, and support purpose	• Extract information from incidents, content for security awareness • Extract key security markers and attributes from new threat reports.
Generation Generate AI to create text	• User stories, personas • Create personalized UX code from experience design • Training, and testing data for chatbots • Automate responses to emails and reviews	• Automate the creation of marketing material and language translation • Automate image, text, and video creation for articles, blogs, etc. • Create automation scripts for various workflows across applications	• Create technical document from code • Automate scripts to configure, deploy, and manage hybrid cloud • Co-pilot to create code across multiple programming languages	• Automate report generation • Social engineering simulation • Security documentation creation • Automate threat detection by looking for anomaly patterns
Classification For sentiment or topics	• Classify customer sentiments from feedback or chatbot interaction • Classify typical issues raised by clients for focused improvements	• Classify documents by different criteria – types, contents, keywords • Sort digital contents in storage into pre-defined categories	• Classify incident reports • Automate workflow based on analysis of items/status/reports	• Classify flagged items properly as threats or other categories • Classify the type of security risks and find the best response • Classify log and other monitoring output to determine the next action
Question answering Knowledge base search across the company's proprietary data.	• Knowledgebase articles • Augment chatbot w/search • Agent assist • Contract intelligence • mart search in technical manuals, HR documents, ethics codes, product documentation, etc.	• Analyze emails, attachments, documents, invoices, reports, etc. • Knowledge search for company information to provide in-house day-to-day assistance and automation	• Knowledge search for IT helpdesk • Ticket resolution by suggesting solutions from resolved tickets • Error log and root cause analysis • Compliance monitoring	• Knowledge search across security spec documents • External threat intelligence • Error log and root cause analysis • Security incident search @ forensics

Fusion HCI for watsonx

Fusion HCI

IBM Storage Ready Nodes for Ceph*

*target availability 4Q23

watsonx.data
DB2, Netezza, Presto, Spark

watsonx.ai
Inferencing/Fine tuning

watsonx.governance

1 Fusion HCI: watsonx compute and data acceleration appliance

Core Benefits

- watsonx.data fully qualified platform
- Appliance-like experience for watsonx on-premises
- High-speed cache for hybrid-cloud data sources (up to 7x faster queries)
- Built on proven, enterprise-hardened core data services, including backup

watsonx.data query engines ***scream*** with Fusion HCI's shared cache accelerator

2 Ceph: watsonx storage (data lake on-prem)

Storage Ceph Storage Ceph

Core Benefits

- **Included:** 'starter' capacity already included inside watsonx.data
- **Cloud-native:** self-healing, on-demand, secure, low-cost
- **Performance-at-Scale:** Smooth perf to 10B objects
- **Easy:** Ceph 'Ready Node', 1-command install, and Ceph Dashboard

watsonx.data lake house ***scales*** cost-effectively with IBM Storage Ceph

Fusion HCI for watsonx - Value proposition

watsonx in a box

Hardware architecture optimized for watsonx

No need to design/prove/maintain a custom architecture and deployment model

Time to value

Deploy the full solution in under a week

Automated OpenShift install

Storage pre-configured for running Cloud Pak for Data

Simplified Day 2 UX

Orchestrated firmware upgrades

Orchestrated scale out/up

Integrated appliance monitoring/management experience

High performance

watsonx.data storage acceleration improves query times by 7-90x

NVMe cache shared by all query engines and nodes

Dedicated 100 GbE storage network

Enterprise ready

Highly available storage and networking

Backup & restore – built in for recovering from data loss or corruption

Disaster recovery – built in for recovering from the loss of a data center

Service and support

Expert Labs installation

Technical account manager: concierge service for helping with the adoption of the solution and management of the appliance

Single point of entry into IBM for the entire solution

Model IP indemnification

→ Model IP (Intellectual Property) indemnification refers to a legal protection mechanism where the provider of a software or technology model assures the client against legal disputes arising from the use of their intellectual property (IP).

→ To learn more about how IBM's standard IP indemnification methods and where it stands against its competitors check this link.

IBM