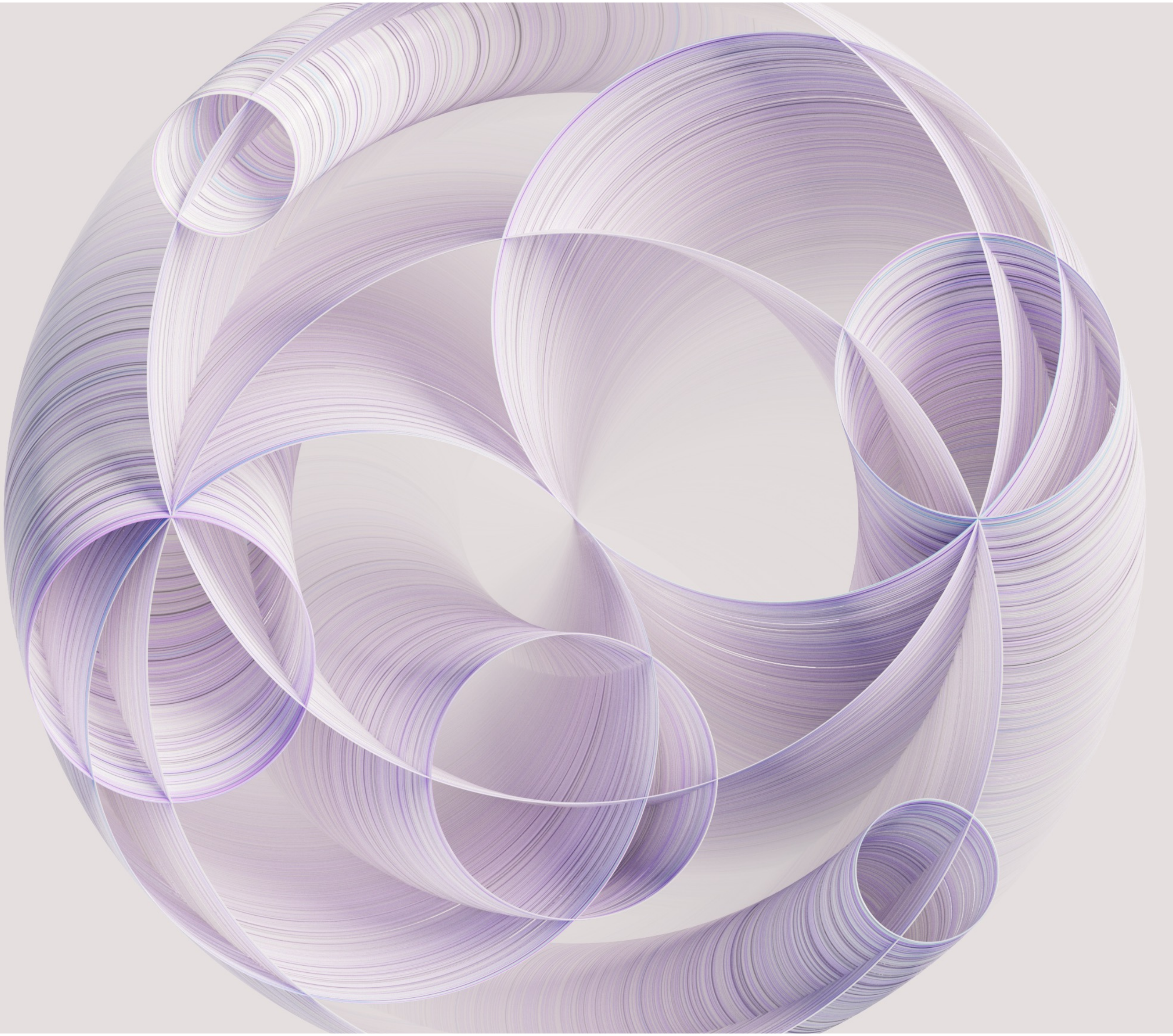
watsonx.ai

Train, validate, tune and deploy AI models



Contents

- **Introduction**
  - Generative AI and traditional AI
  - Foundation models and generative AI
  - Common generative AI tasks
  - Risks and requirements for a generative AI platform
- **Watsonx** and **watsonx.ai**
  - IBM **watsonx** and its components
  - IBM **watsonx.ai**
    - Train, validate, tune, and deploy AI models
  - IBM **watsonx.ai** components
    - Foundation models library
    - Prompt lab
    - Tuning studio *
- **Watsonx.ai value propositions**
- **Getting started with watsonx.ai**

AI adoption more than doubled since 2017

2017

20% → 2.5X → 50%

2022

Source: McKinsey – The State of AI in 2022 – and a half decade in review

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

Enterprise leaders are faced with unprecedented challenges to scale AI

1 in 5

Leaders cite difficulties integrating data across any cloud

25%

of organizations lack the tools or platforms to develop models

34%

of businesses lack the necessary AI skills, expertise or knowledge to keep up with AI innovation

74%

of leaders haven’t taken the necessary steps to reduce bias in the organization’s AI

Source: 2022 AI Adoption Index https://www.ibm.com/downloads/cas/GVAGA3JP

Impact of generative AI

The speed, scope, and scale of generative AI impact is unprecedented

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

Massive early adoption

80%

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

Broad-reaching and deep impact

Generative AI could raise global GDP by

7%

within 10 years

Critical focus of AI activity and investment

Generative AI expected to represent

30%

of overall market by 2025

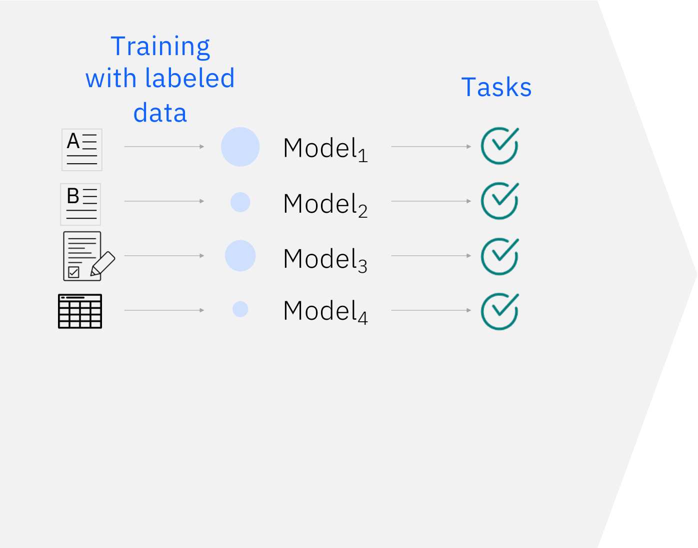
Generative AI and traditional AI

Both traditional AI and generative AI are useful for enterprises.
Neither replaces the other, generative AI opens new possibilities

Generative AI	Traditional AI
Foundation models trained with unlabeled data	Traditional Machine learning (ML/AI) model trained with “labeled” data
Unsupervised	Training is supervised
Trained on very big data sets	Trained on proper, large data sets
No specific task	Trained for a specific task
Transferable	Does not transfer well to other tasks
Works well for general tasks and can improve for specific tasks with less training	A tuned model can be very efficient for the specific task it was designed for
Need to monitor bias and drift	Need to monitor bias and drift

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

Traditional AI models



Training with labeled data

Tasks

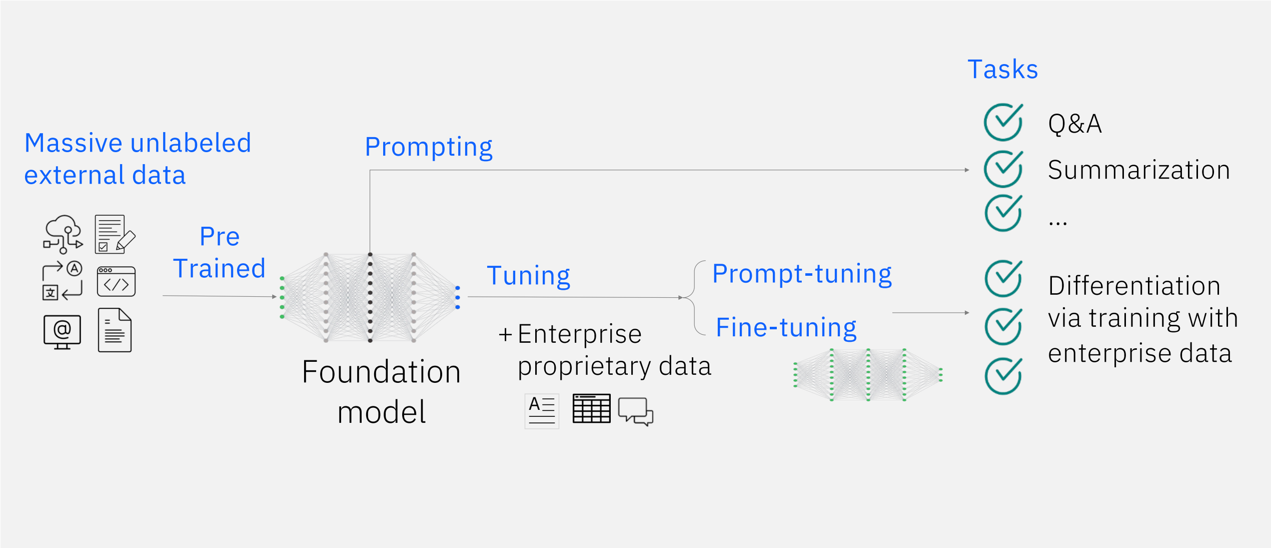
Model$_1$

Model$_2$

Model$_3$

Model$_4$

Foundation Models



Massive unlabeled external data

Pre Trained

Foundation model

Prompting

Tuning

+ Enterprise proprietary data

Prompt-tuning

Fine-tuning

Tasks

Q&A

Summarization

...

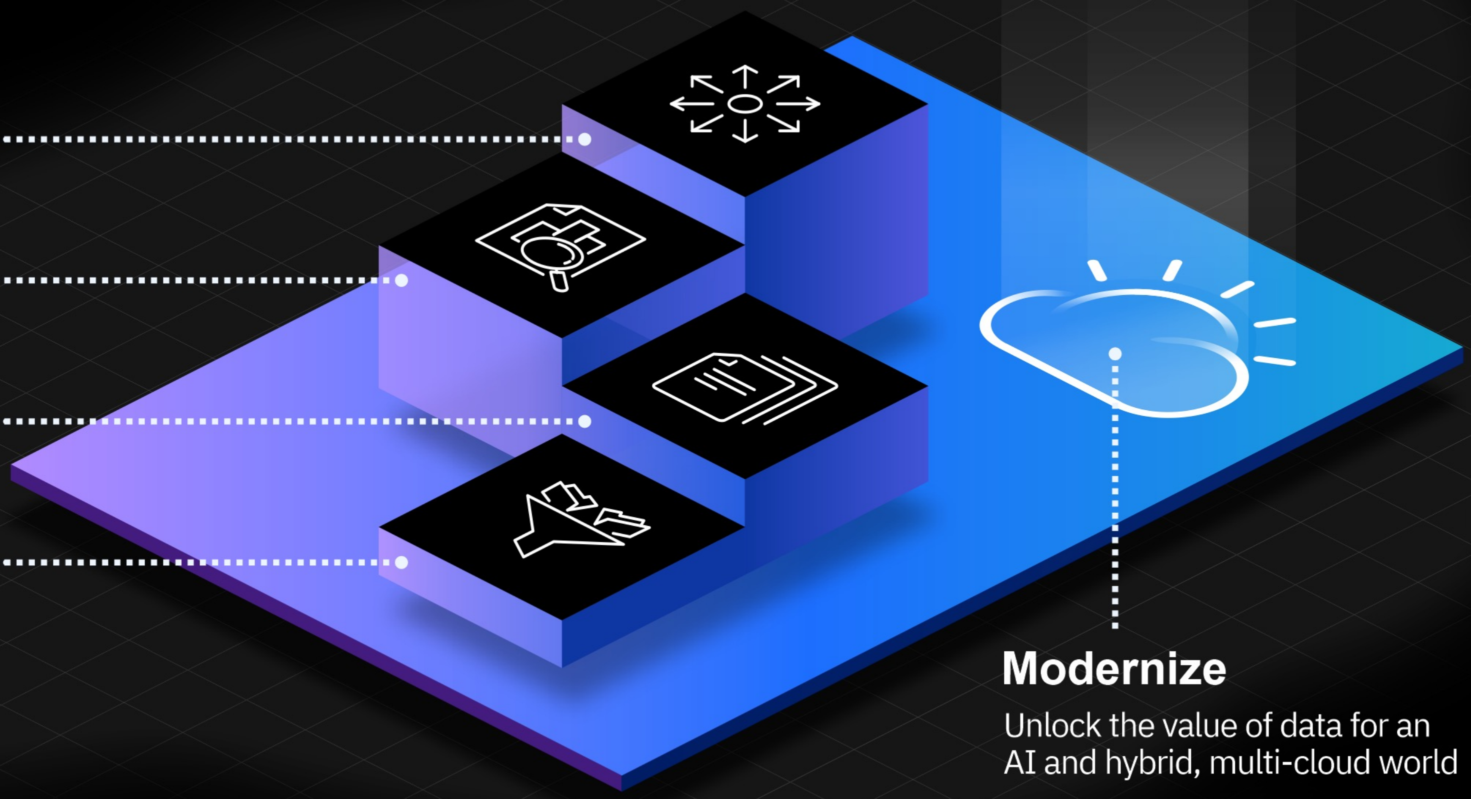
Differentiation via training with enterprise data

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

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

The AI Ladder

IBM’s prescriptive approach to the journey to AI



Infuse
Operationalize AI
throughout the enterprise

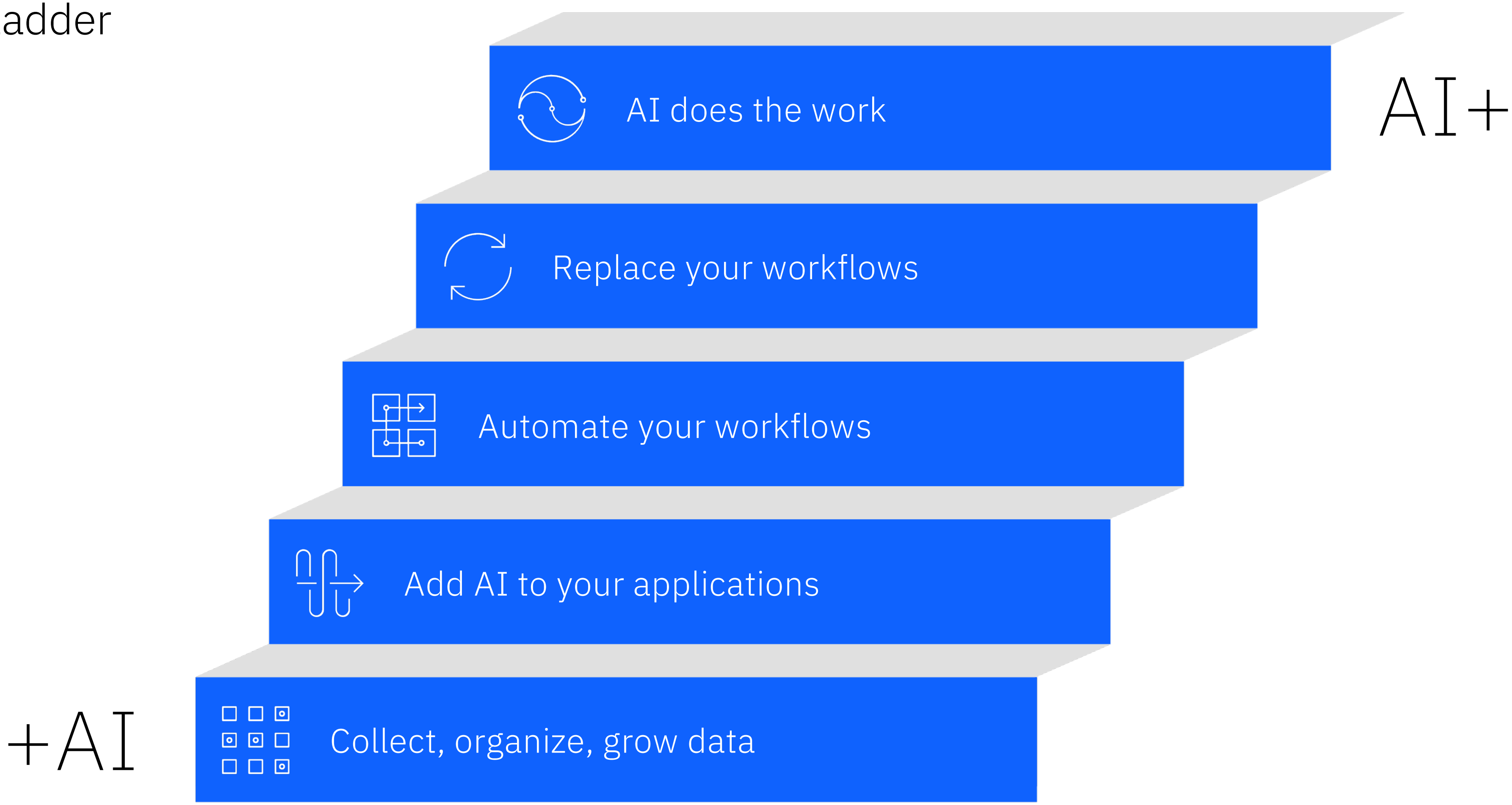
Analyze
Build and scale AI with
trust and transparency

Organize
Create a business-ready
analytics foundation

Collect
Make all data simple
and accessible

Modernize
Unlock the value of data for an
AI and hybrid, multi-cloud world

The modern-day AI ladder



AI does the work

AI+

Replace your workflows

Automate your workflows

Add AI to your applications

Collect, organize, grow data

+AI

Introducing…

watsonx

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.

Unleash the intelligence in your business with IBM Software

AI products

Digital Labor
- Watson Orchestrate
- Watson Assistant
- Watson Code Assistant
- Watson Discovery
- Planning Analytics

IT Automation
- Turbonomic
- Instana
- AIOps Insights
- Hybrid Cloud Mesh
- SevOne

Security
- QRadar
- Randori Recon
- Guardium
- MaaS360 | Verify
- Trusteer

Sustainability
- Envizi
- EIS
- Maximo
- Sterling

Application Modernization
- API Connect
- App Connect
- Event Automation

Software and SaaS partners

AI and data platform

watsonx
- watsonx.ai
- watsonx.data
- watsonx.governance

Cloud Paks

Hybrid cloud platform

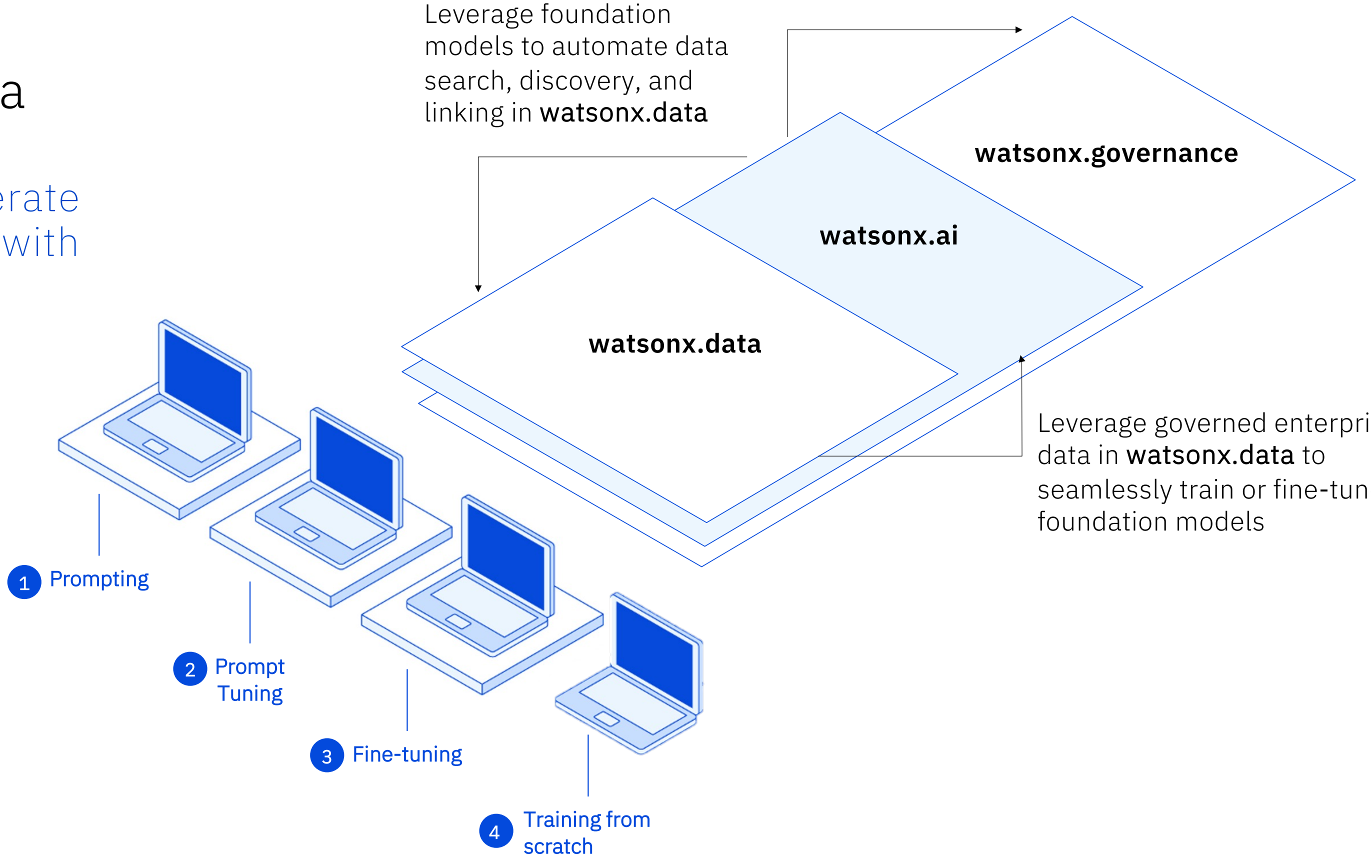
Red Hat
- OpenShift
- Enterprise Linux
- Ansible Automation Platform

watsonx

and its 3 components

The platform for AI and data

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**

watsonx.governance

watsonx.ai

watsonx.data

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

1. Prompting
2. Prompt Tuning
3. Fine-tuning
4. Training from scratch

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*

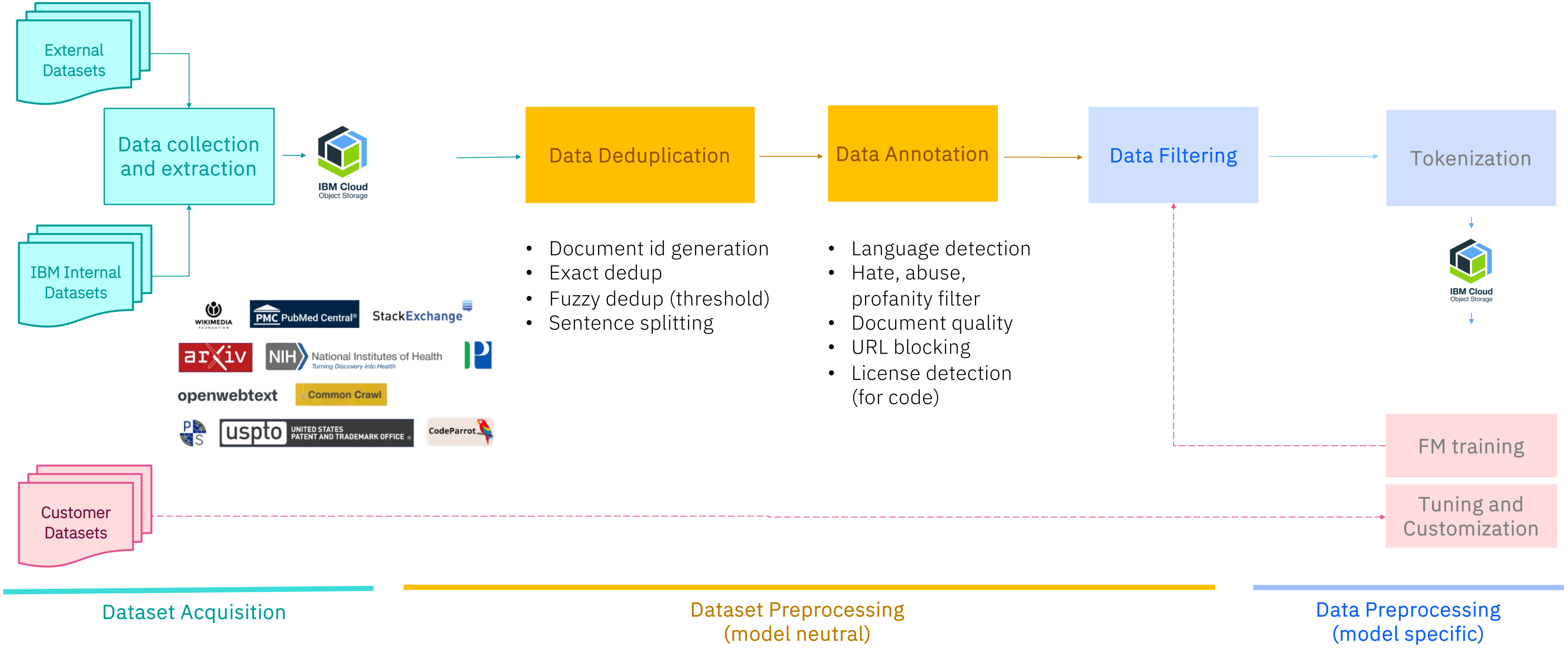
Plus, a proven studio for machine learning

- ModelOps
- Automated development
- Decision optimization

Team collaboration and data preparation

watson*x*.data: curated IBM Data Pile

Enterprise-ready data acquisition, curation, provenance, and governance



External Datasets

IBM Internal Datasets

Data collection and extraction

Customer Datasets

Data Deduplication

- Document id generation
- Exact dedup
- Fuzzy dedup (threshold)
- Sentence splitting

Data Annotation

- Language detection
- Hate, abuse, profanity filter
- Document quality
- URL blocking
- License detection (for code)

Data Filtering

Tokenization

FM training

Tuning and Customization

Dataset Acquisition

Dataset Preprocessing (model neutral)

Data Preprocessing (model specific)

Foundation model libraries

IBM watsonx.ai provides open foundation models as well as IBM's securely trained models including encoder, decoder, & encoder-decoder foundation models.

Encoder

Powers search engines and customer-service chatbots, including IBM Watson Assistant.

Encoder-only models are widely used for non-generative tasks like classifying customer feedback and extracting information from long documents.

Encoders are not trained to predict, but rather to find & extract useful information from models trained with relatively smaller sets of data.

Google's Bidirectional Encoder Representation from Transformer (BERT) is an example

Decoder

Trained to predict the next word without an encoded representation.

Decoder-only models are used with chat platforms like ChatGPT to generate new responses, answers, or text based on simple input from customers.

Decoders are trained for predictive, generative tasks based on their massive pre-trained data sets.

The Generative Pre-Trained Transformer (GPT) is an example

Encoder-decoder

Combines features of both encoder and decoder models.

They can do many of the generative tasks that decoder-only models can, but their compact size makes them faster and cheaper to tune and serve.

Google's Text-to-Text Transfer Transformer (T5) is an example.

watsonx.ai Foundation Model Library

Model variety to cover enterprise use cases and compliance requirements

IBM Foundation Models

Slate (encoder-only) NLP models

Slate
multilingual
distilled
153 million

Fine Tuning ***Required*** to support:

- Extract
- Classify

Open-Source Large Language Models

Encoder/decoder & decoder-only Large Language Models available in *Prompt Lab*
(Fine tuning NOT required for most tasks)

flan-ul2 *20 billion* encoder/decoder	gpt-neox *20 billion* decoder only	mt0-xxl *13 billion* encoder/decoder	flan-t5-xxl *11 billion* encoder/decoder	mpt-instruct2 *7 billion* decoder only
Q&A	Q&A	Q&A	Q&A	Q&A
Generate	Generate	Generate	Generate	Generate
Extract		Extract	Summarize	
Summarize		Summarize	Classify	
Classify		Classify		

Note: Slate models are fine-tuned
via notebooks + API

Open-source models are sourced from Hugging Face

Q&A	Model responds to a question in natural language
Generate	Model generates content in natural language
Extract	Model extracts entities, facts, and info. from text
Summarize	Model creates summaries of natural language
Classify	Model classifies text (e.g. sentiment, group, etc..)

watsonx.ai and its own foundation models

Model architectures	Architecture name	Use case
Encoder-only	Slate (use through a Watson Studio notebook)	Best cost performance trade-off for non-generative use cases but require task-specific labeled data for fine tuning.
Decoder-only	Granite Post July 2023	Designed explicitly for generative AI use cases; represents the architectures used in GPT-3 and other popular LLMs.
Encoder-decoder	Sandstone Post July 2023	Support both generative and non-generative use cases. Best cost performance trade-off for generative use cases when input is large but generated output is small.

The initial focus is on two categories for each model: language and code. A model's name will identify its characteristics

Model naming convention: **<category>/<architecture><opt sub arc>.<size>.<opt info>** where the 2 initial categories are fm.code and fm.language or fm.code/<architecture> where fm is the name of the foundation model

Examples:

- **fm.language/sandstone.3b** is the Encoder-decoder model on language-related tasks with 3b parameters.
- **fm.code/granite.350m.ansible** is the Decoder model on code-related tasks with 350m parameter for ansible

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
- Sandstone
  Encoder-decoder transformers
- Obsidian (in progress)
  Sparse universal transformers

Efficient Domain and Task Specialization

Models Coming Soon:

- Finance
- Cybersecurity
- Legal, etc.

Opensource models

Experiment with opensource 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: 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

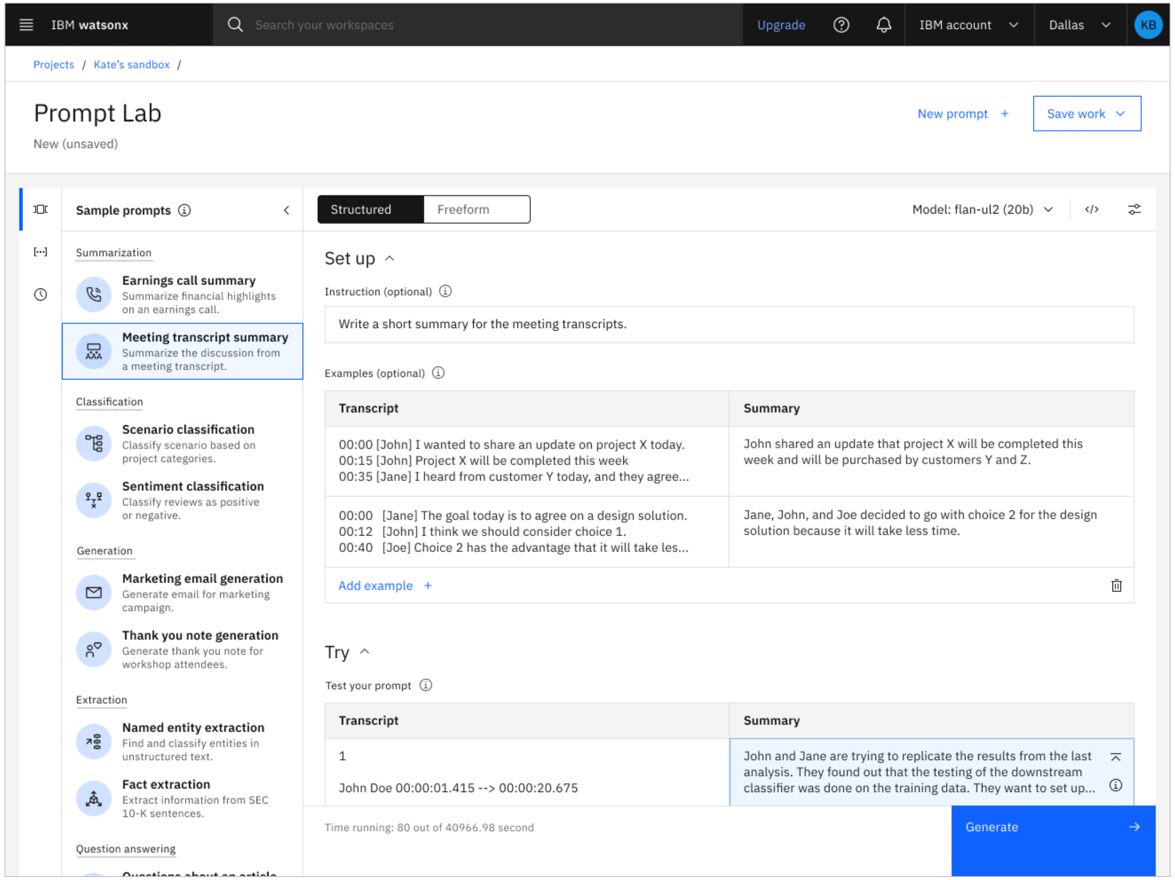
Experiment with prompt engineering

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
Search your workspaces
Upgrade
IBM account
Dallas
KB
Projects / Kate's sandbox /
Prompt Lab
New (unsaved)
New prompt +
Save work
Sample prompts
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
Structured
Freeform
Model: flan-ul2 (20b)
Set up
Add example +
Try
Generate

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.

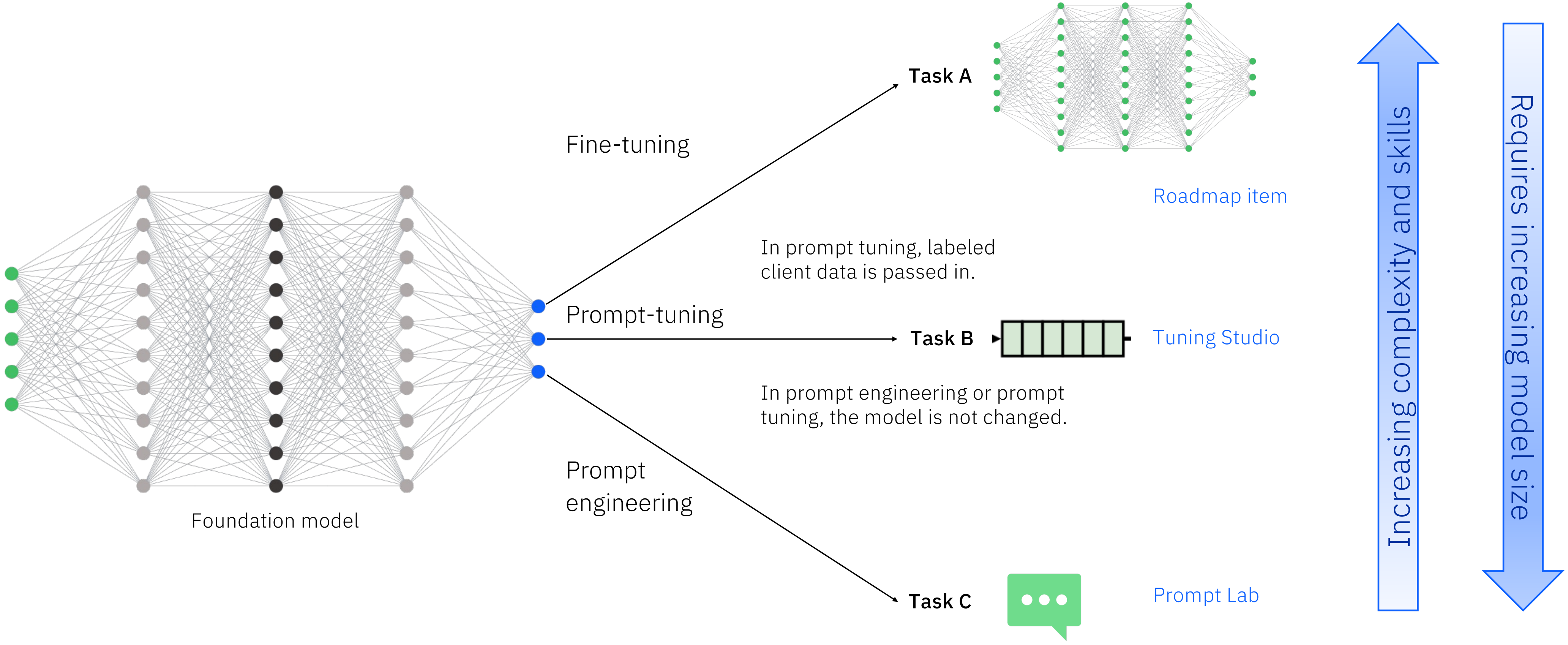
Test your prompt

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

Rapid adaptation to multiple tasks with small amounts of task-specific data

Fine-tuning requires labeled data and more resources to tune the model. When a model is fine-tuned, some of the weights are modified and clients get a private instance of the model.



In prompt tuning, labeled client data is passed in.

In prompt engineering or prompt tuning, the model is not changed.

watsonx.ai

Prompt engineering

Benefits to clients:

- There is no change to the model (no need for an expensive rebuild)
- No need to pass in labeled data
- Much faster way of training the prompt to properly respond to requests
- Can be passed in via APIs once a prompt has been properly engineered
- Can be used by anyone to guide the model to respond in a desirable way

Prompt tuning

Benefits to clients

- There is no change to the model (no need for an expensive rebuild)
- Requires labeled data but can achieve better performance even with using a smaller size model.
- Can achieve close to fine-tuning results without model modification
- Can be passed in via API
- Can be used by anyone to work with the model

watsonx.ai

Prompt Engineering

Prompt engineering is an art – and watsonx.ai Prompt Lab provides a rich and guided learning experience.

Foundation models are adaptable. One way to adapt the model is using prompt engineering.

What is prompt engineering?

- Prompt engineering is a new discipline for finding the optimal prompt to use with a foundation model for the best performance.

Why is it important?

- For most generative AI, it is not so much "answering" a question than simply appending the most likely text
- Simply asking a straightforward question may not yield the best result

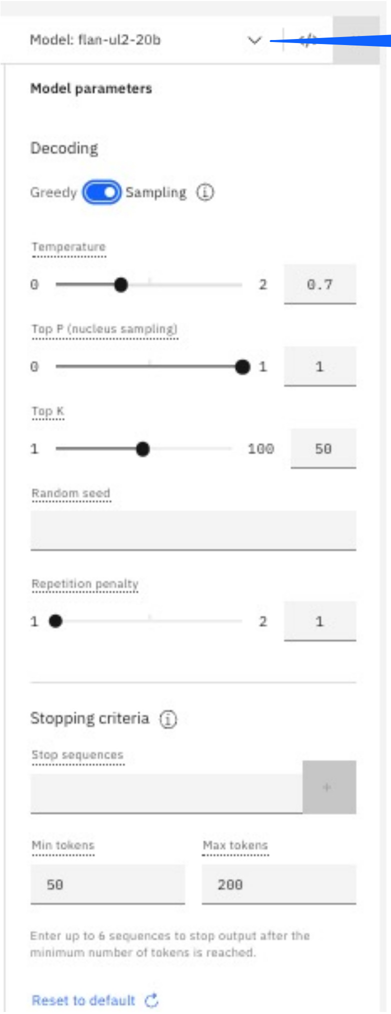
- A prompt is a way to communicate with the foundation model to:
  - Pass additional instructions on how the model should respond
  - Feed task-specific context to the model
- IBM **watsonx.ai** provides a Prompt Lab with an interactive prompt builder:
  - Includes prompt examples for various use cases
  - Clients can experiment with different prompts, save and reuse older prompts, use different models and vary different parameters

watsonx.ai

Prompt Lab

The Prompt Lab allows other modifications on various parameters such as:

- **Choice of foundation models to use** – some are better at certain tasks than others
- **Types of decoding used** – the process the model uses to generate the output
- **Repetition penalty** – prevent the model from generating repeating phrases
- **Stop sequences** – specifies sequences whose appearances should stop the model
- **Number of min and max new tokens** – specifies boundaries of the number of tokens in the response



Model: flan-ul2-20b

Model parameters

Decoding

Greedy ● Sampling

Temperature: 0 – 2: 0.7

Top P (nucleus sampling): 0 – 1: 1

Top K: 1 – 100: 50

Random seed:

Repetition penalty: 1 – 2: 1

Stopping criteria

Stop sequences:

Min tokens: 50

Max tokens: 200

Enter up to 6 sequences to stop output after the minimum number of tokens is reached.

Model selection

Parameters for prompt engineering

watsonx.ai

Prompt Lab: email generation

Prompt: Write a thank you note for attending a workshop.
Attendees: Enterprise clients
Topic: Introduction to IBM watsonx
Tone: Appreciative

Output:
We would like to thank all the attendees that came out and participated in today's introduction to watsonx workshop. It was a great turnout, and you are making an impact by engaging your enterprise with Watson. Thank you for taking the time, spending the day, and sharing your experience with it. We hope that you enjoyed the workshop and learned something new. If you have any questions about watsonx, just let me know!

Append prompt with additional context

Prompt: Write a thank you note for attending a workshop.
Attendees: Enterprise clients
Topic: Introduction to IBM watsonx
Tone: Appreciative
Expectation: discuss business opportunities

Output:
We would like to thank you for taking the time to attend our workshop today. It was a great day, and we are pleased that you were able to participate. This is an exciting time for Watson Health as we continue to expand our capabilities in clinical care, cognitive computing, analytics and mobile technology. As IBM watsonx continues to expand, we hope that you will reach out to us with questions or concerns regarding the product. In addition, we welcome feedback on how we can improve your experience with watsonx. Thank you again

watsonx.ai

Prompt Lab: client complaint

The following paragraph is a consumer complaint.
The complaint is about one of these options: credit cards, credit reporting, mortgages and loans, retail banking, or debt collection.
Read the following paragraph and list all the issues.

I called your helpdesk multiple times and every time I waited 10-15 minutes before I gave up. The first time I got through, the line got cut suddenly and I had to call back. This is just ridiculous. When I finally got through like after 3 days (yes, 3 days) your agent kept going over a long checklist of trivial things and asking me to verify, after I repeatedly told the agent that I am an experienced user and I know what I am doing. It was a complete waste of time. After like an eternity of this pointless conversation, I was told that an SME will contact me. That – was 2 days ago. What is the problem with your support system?

Retail banking

In this example, a modified prompt with an example is provided to the model to teach it what is expected.

I bought a GPS from your store and the instructions included are in Spanish, not English. I have to use Google Translate to figure out. The mounting bracket was broken, and so I need information on how to get a replacement. Moreover, the information seems to be outdated because I cannot see the new roads put in around my house within the last 12 months.

The list of issues is as follows:
1) The instructions are in Spanish, not English.
2) The mounting bracket is broken.
3) The information is outdated.

The next page shows the new output with this additional “one shot” input.

watsonx.ai

Prompt Lab:
client complaint continue

The following paragraph is a consumer complaint.
The complaint is about one of these options: credit cards, credit reporting, mortgages and loans, retail banking, or debt collection. Read the following paragraph and list all the issues.

I called your helpdesk multiple times and every time I waited 10-15 minutes before I gave up. The first time I got through, the line got cut suddenly and I had to call back. This is just ridiculous. When I finally got through like after 3 days (yes, 3 days) your agent kept going over a long checklist of trivial things and asking me to verify, after I repeatedly told the agent that I am an experienced user and I know what I am doing. It was a complete waste of time. After like an eternity of this pointless conversation, I was told that an SME will contact me. That – was 2 days ago. What is the problem with your support system?

With the added instructions from the previous slide, this is the new output:

```
The list of issues are as follows;
1)  Waited 10-15 minutes every time I called
2)  The line got cut suddenly
3)  The agent kept going over a long checklist of trivial things
4)  I was told that an SME will contact me
5)  No one called
```

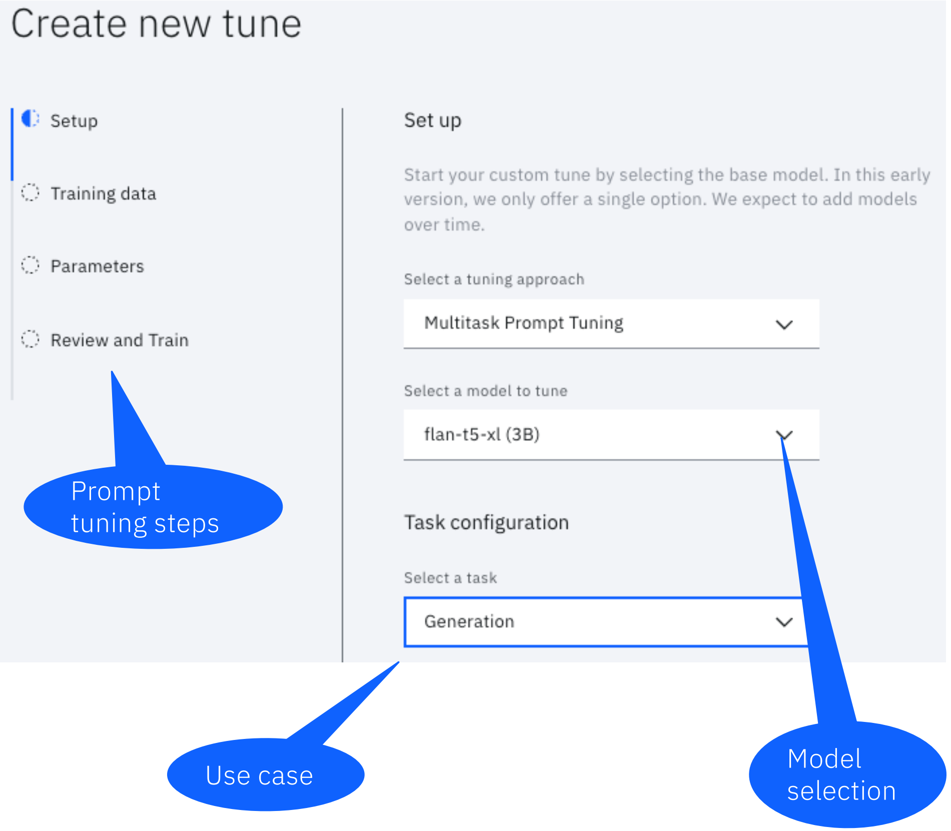
This is the expected results.
With the added instructions, the model further understands that the task is to extract a list of issues raised by the customer.

watsonx.ai

Prompt tuning

Another useful feature of **watsonx.ai** is the prompt tuning in the tuning studio.

- **Watsonx.ai** allows clients to further tune the prompts
- Unlike prompt engineering, prompt tuning allows clients to further train a model with focused, business data.
- In prompt tuning, the underlying model is not changed.
- The input prompt is changed via augmenting input with examples.



Create new tune

Set up

Start your custom tune by selecting the base model. In this early version, we only offer a single option. We expect to add models over time.

Select a tuning approach: Multitask Prompt Tuning

Select a model to tune: flan-t5-xl (3B)

Task configuration

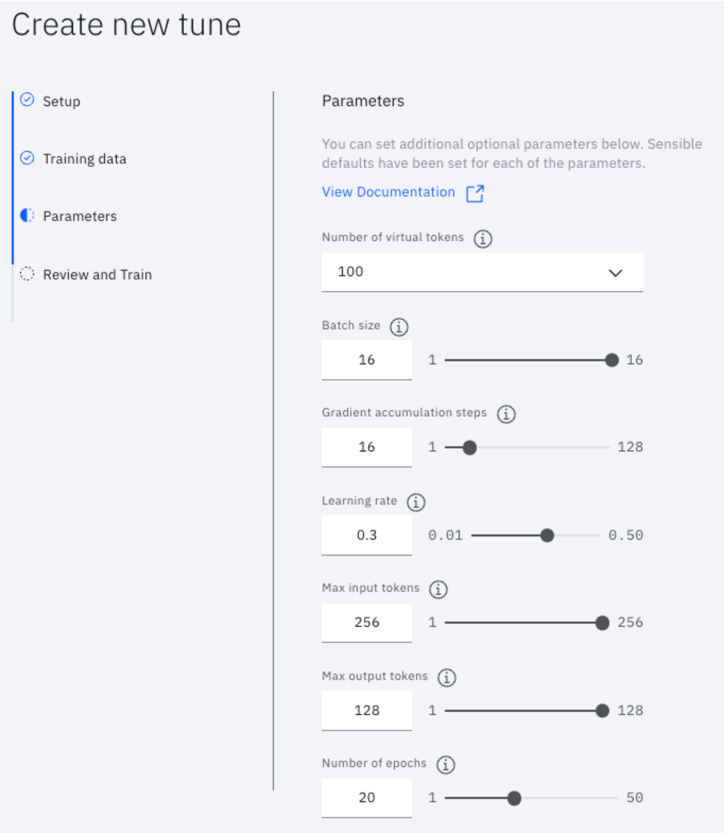
Select a task: Generation

watsonx.ai

Prompt tuning

Clients can:

- Name the finished model name
- Select a model to tune
- Provide their own training data
- Tune training parameters such as:
  - Batch size
  - Gradient accumulation steps
  - Learning rate
  - Max input and output tokens
  - Number of epochs
- Train and automatically deploy the model



Create new tune

Parameters

You can set additional optional parameters below. Sensible defaults have been set for each of the parameters.

Number of virtual tokens: 100

Batch size: 16 (1 – 16)

Gradient accumulation steps: 16 (1 – 128)

Learning rate: 0.3 (0.01 – 0.50)

Max input tokens: 256 (1 – 256)

Max output tokens: 128 (1 – 128)

Number of epochs: 20 (1 – 50)

Parameters for prompt tuning

IBM partnership with open-source models provider

HUGGING FACE

- IBM **watsonx.ai** clients have access to the latest and greatest open-source foundation models from Hugging Face.
- The IBM and Hugging Face partnership demonstrates a joint commitment to deliver an open ecosystem to clients, allowing them to find the best foundation models for their business needs.

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

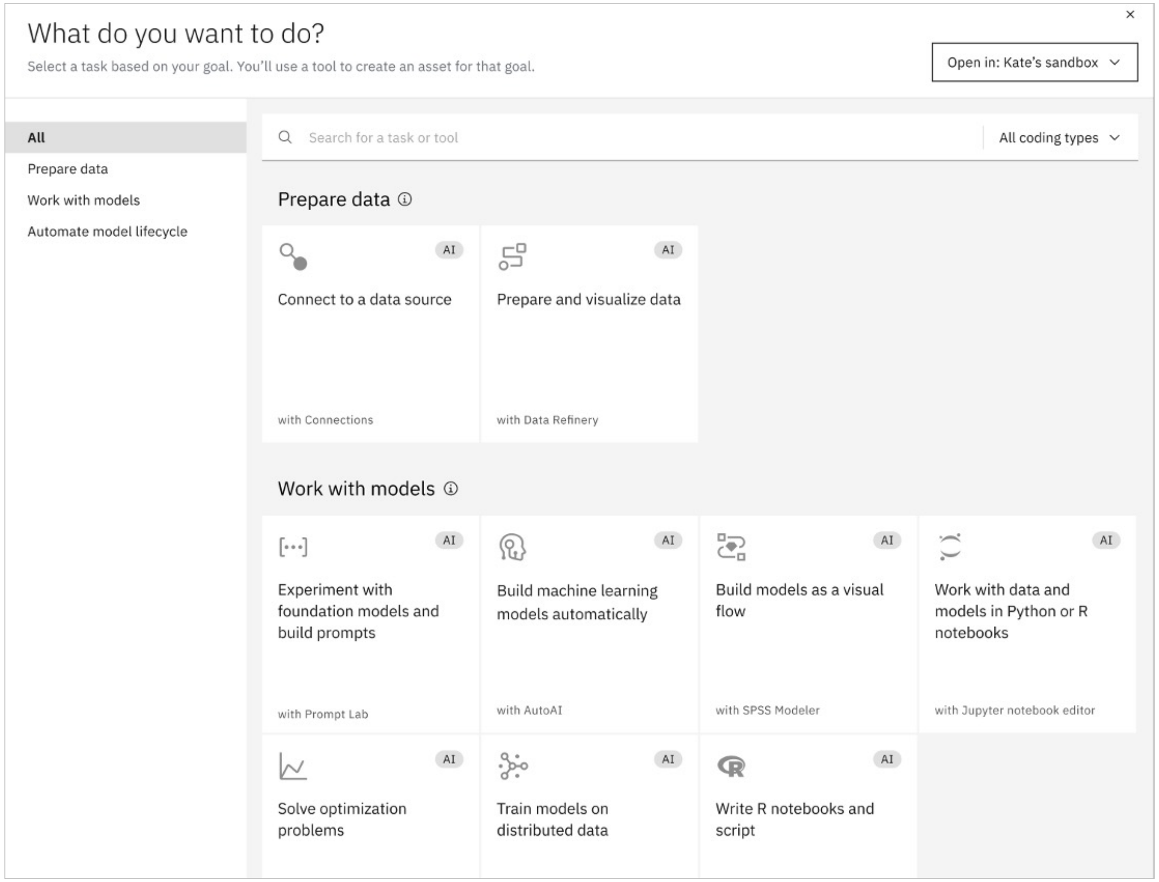
AutoAI, including preparing data for machine learning and generating and ranking candidate model pipelines

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



What do you want to do?

Select a task based on your goal. You'll use a tool to create an asset for that goal.

Open in: Kate's sandbox

Prepare data

Connect to a data source
with Connections

Prepare and visualize data
with Data Refinery

Work with models

Experiment with foundation models and build prompts
with Prompt Lab

Build machine learning models automatically
with AutoAI

Build models as a visual flow
with SPSS Modeler

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

Solve optimization problems

Train models on distributed data

Write R notebooks and script

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

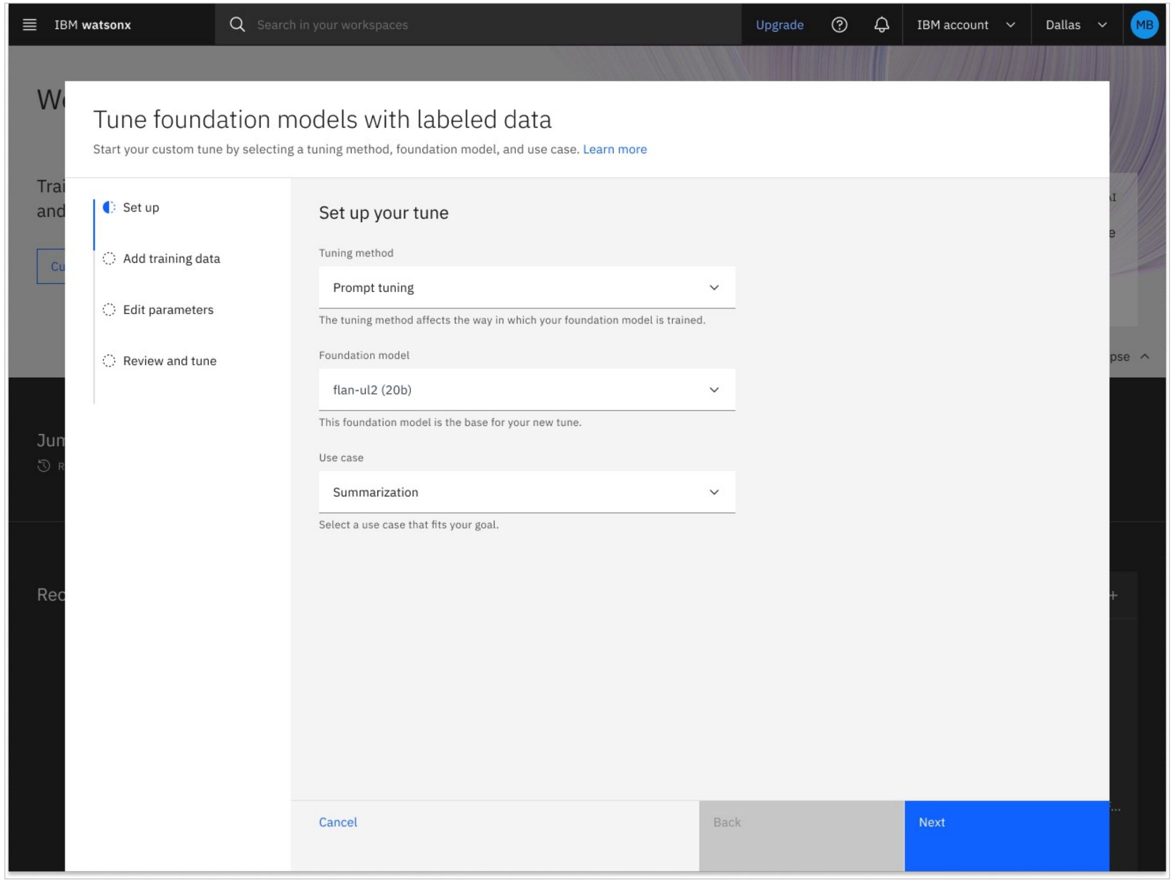
Unlike prompt engineering, prompt tuning allows clients to further train the model with focused, business data

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



Tune foundation models with labeled data

Start your custom tune by selecting a tuning method, foundation model, and use case. Learn more

Set up your tune

Tuning method: Prompt tuning

The tuning method affects the way in which your foundation model is trained.

Foundation model: flan-ul2 (20b)

This foundation model is the base for your new tune.

Use case: Summarization

Select a use case that fits your goal.

*Coming soon, available post-GA

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.

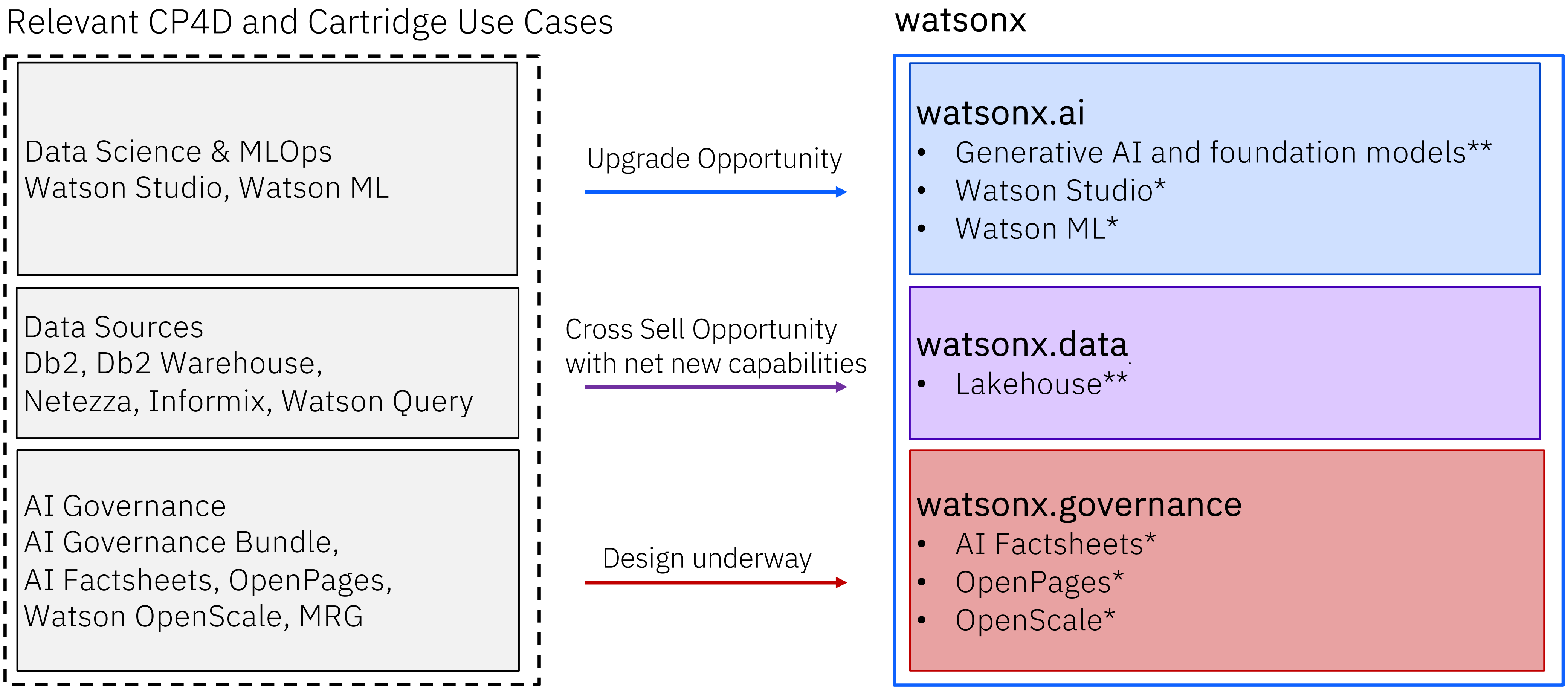
Empowering

- 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.

Why IBM?

- **Open - Hybrid and Multicloud capabilities -** Can work on cloud platform of choice for client as well as on-prem (in future). No cloud vendor lock-in
- **Trusted - Market leader Governance** - Provides tooling and capabilities for end-to-end Data and AI governance.
- **Trusted - IBM Foundation models** are trained on data which is checked and curated by legal and ethical teams and uses HAP filters, so that you can completely trust the data on which they are build.
- **Empowering - Integrated** capabilities for **generative AI** and **traditional AI**
- **Empowering** - Complete and Integrated capabilities for AI, Data and Governance.
- **Targeted - IBM’s Prompt Tuning** uses Multitask Prompt Tuning, developed by **MIT IBM Watson AI Labs** which is efficient, low-cost way, of adapting an AI foundation model to your custom tasks, using lesser training data and much lesser cost compared to fine tuning approach and still able to provide at par results with fine tuning models.
- **Targeted** - IBM is creating smaller foundation models which can be effective for specific tasks. Idea is to use diff models for diff tasks at much lower cost instead of using one large model for all the use cases.

Connecting current software portfolio with **watsonx**



Relevant CP4D and Cartridge Use Cases

Data Science & MLOps
Watson Studio, Watson ML

Data Sources
Db2, Db2 Warehouse,
Netezza, Informix, Watson Query

AI Governance
AI Governance Bundle,
AI Factsheets, OpenPages,
Watson OpenScale, MRG

Upgrade Opportunity

Cross Sell Opportunity
with net new capabilities

Design underway

watsonx

watsonx.ai

- Generative AI and foundation models**
- Watson Studio*
- Watson ML*

watsonx.data

- Lakehouse**

watsonx.governance

- AI Factsheets*
- OpenPages*
- OpenScale*

* capabilities that will be enhanced in **watsonx**
** brand new capabilities to **watsonx**

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

THE CHAMPIONSHIPS WIMBLEDON

Leveraged **watsonx.ai** foundation models to train their AI to create tennis commentary. Generated informative and engaging video clip narrations for fans with varied sentence structures and vocabulary.

SAMSUNG

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.

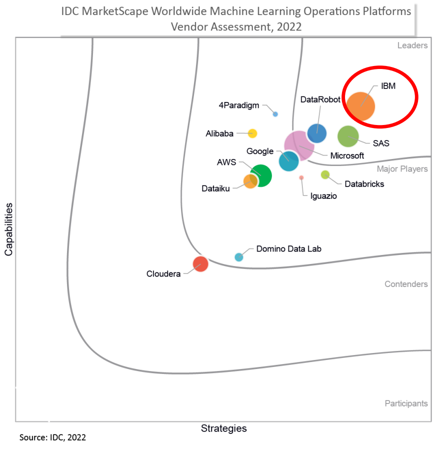
TechD

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

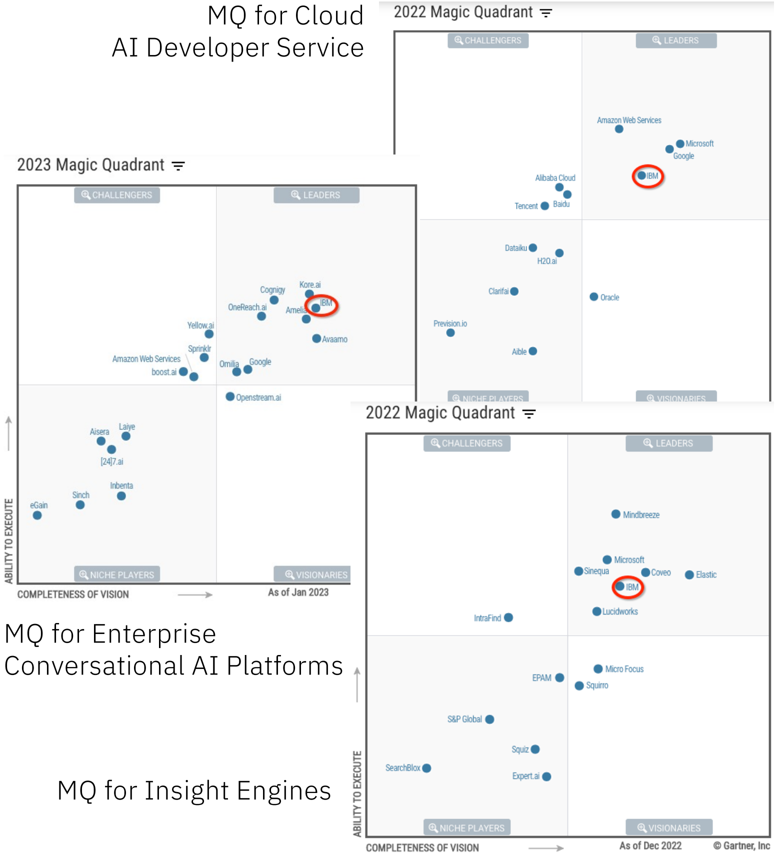


IDC Marketscape:
Leader in Worldwide
Machine Learning
Operations Platforms
2022 Vendor Assessment

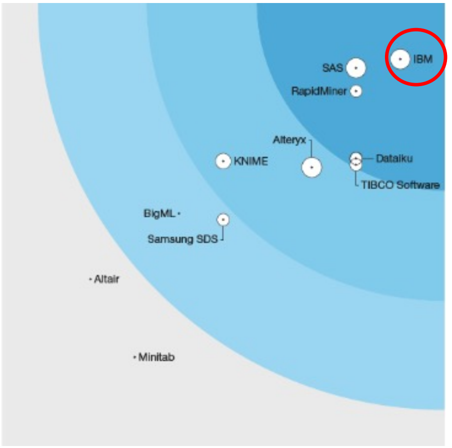
MQ for Cloud
AI Developer Service

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

Call to action

1 Challenge yourself to deliver a watsonx briefing for each of your clients

- Leverage the watsonx.ai Sales Kit, generative AI whitepaper and watsonx.ai video on seismic
- Engage clients in **watsonx.ai** pilot: https://ibm.biz/watsonx-pilot

2 Get Prepared - Complete required learning & earn your **watsonx.ai** sales foundation badge

- AI Fundamentals
- AI for Business
- **watsonx.ai** Level 2
- **watsonx.data** Level 2

3 Participate in the **watsonx** challenge
– Coming Soon…

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.
Generate 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
Classify 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