

# watsonx.ai

Seller presentation

Jessica Ridella

[jnridell@us.ibm.com](mailto:jnridell@us.ibm.com)

World Wide Sales Lead for **watsonx.ai**

Felix Lee

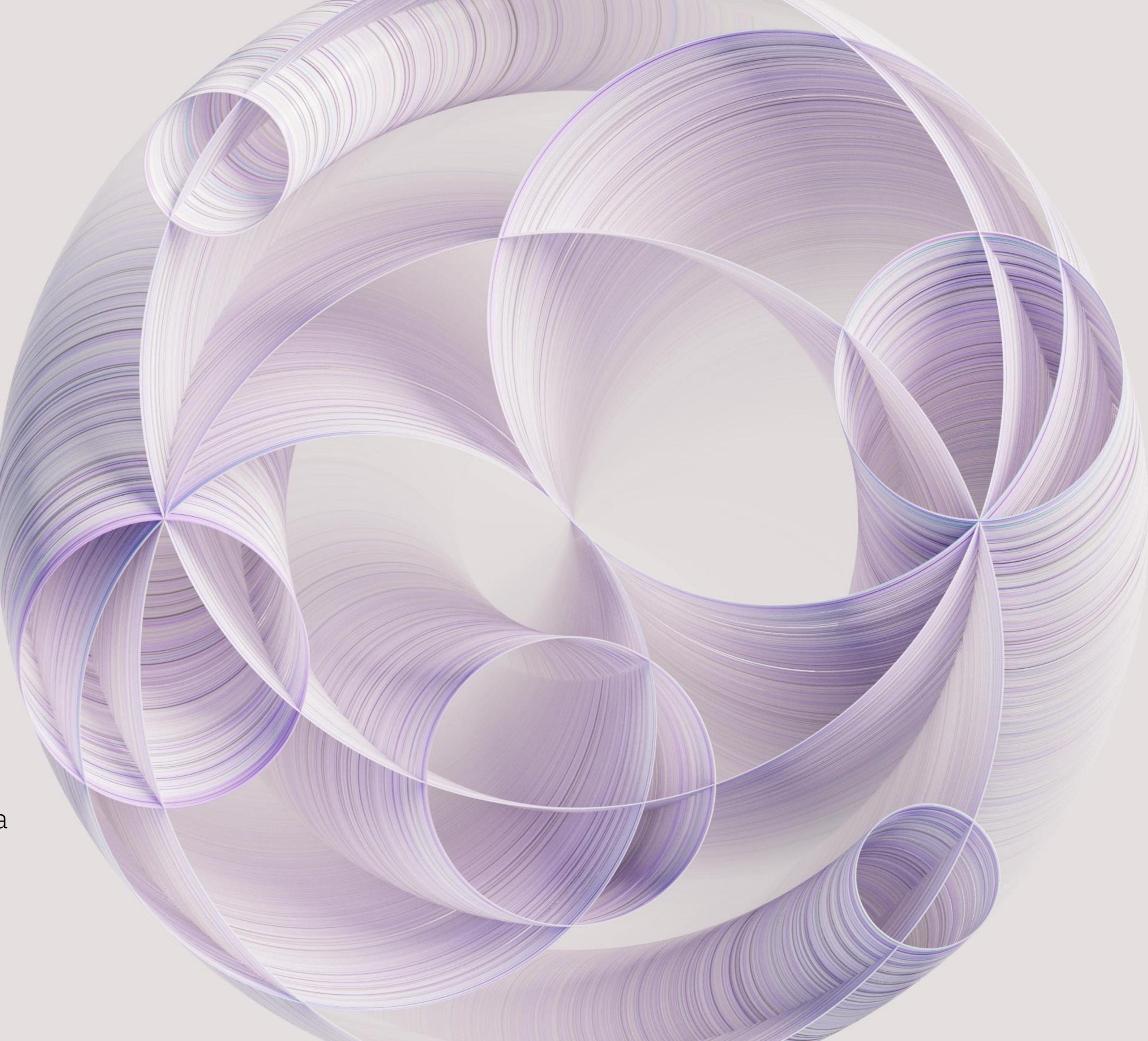
[felix@ca.ibm.com](mailto:felix@ca.ibm.com)

Principal, Learning Content Development  
AI and Data

Anshupriya Srivastava

[anshupriya.srivastava@ibm.com](mailto:anshupriya.srivastava@ibm.com)

Advisory, Learning Content Development, AI and Data



# Seller guidance and legal disclaimer

IBM and Business Partner  
**Internal Use Only**

Slides in this presentation marked as "IBM and Business Partner Internal Use Only" are for IBM and Business Partner use and should not be shared with clients or anyone else outside of IBM or the Business Partners' company.

© IBM Corporation 2024.  
**All Rights Reserved.**

The information contained in this publication is provided for informational purposes only. While efforts were made to verify the completeness and accuracy of the information contained in this publication, it is provided AS IS without warranty of any kind, express or implied. In addition, this information is based on IBM's current product plans and strategy, which are subject to change by IBM without notice. IBM shall not be responsible for any damages arising out of the use of, or otherwise related to, this publication or any other materials. Nothing contained in this publication is intended to, nor shall have the effect of, creating any warranties or representations from IBM or its suppliers or licensors, or altering the terms and conditions of the applicable license agreement governing the use of IBM software.

References in this presentation to IBM products, programs, or services do not imply that they will be available in all countries in which IBM operates. Product release dates and/or capabilities referenced in this presentation may change at any time at IBM's sole discretion based on market opportunities or other factors and are not intended to be a commitment to future product or feature availability in any way. Nothing contained in these materials is intended to, nor shall have the effect of, stating or implying that any activities undertaken by you will result in any specific sales, revenue growth, or other results.

All client examples described are presented as illustrations of how those clients have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by client.

# Content

## The AI landscape

- IBM Watson and AI
- Traditional and generative AI
- Foundation models
- Generative AI use cases
- Benefits and risks
- Enterprise requirements

## Watsonx and **watsonx.ai**

- **Watsonx** family
- **Watsonx.ai** components
  - Foundation model library
  - Prompt Lab
  - Tuning Studio
  - Hugging face partnership
- **Watsonx.ai** and other IBM technology pillars

## Watsonx.ai competitors

## Sellers – call to action

# Gartner reports:

Global spending on artificial intelligence (AI) software will grow to \$298 billion by 2027

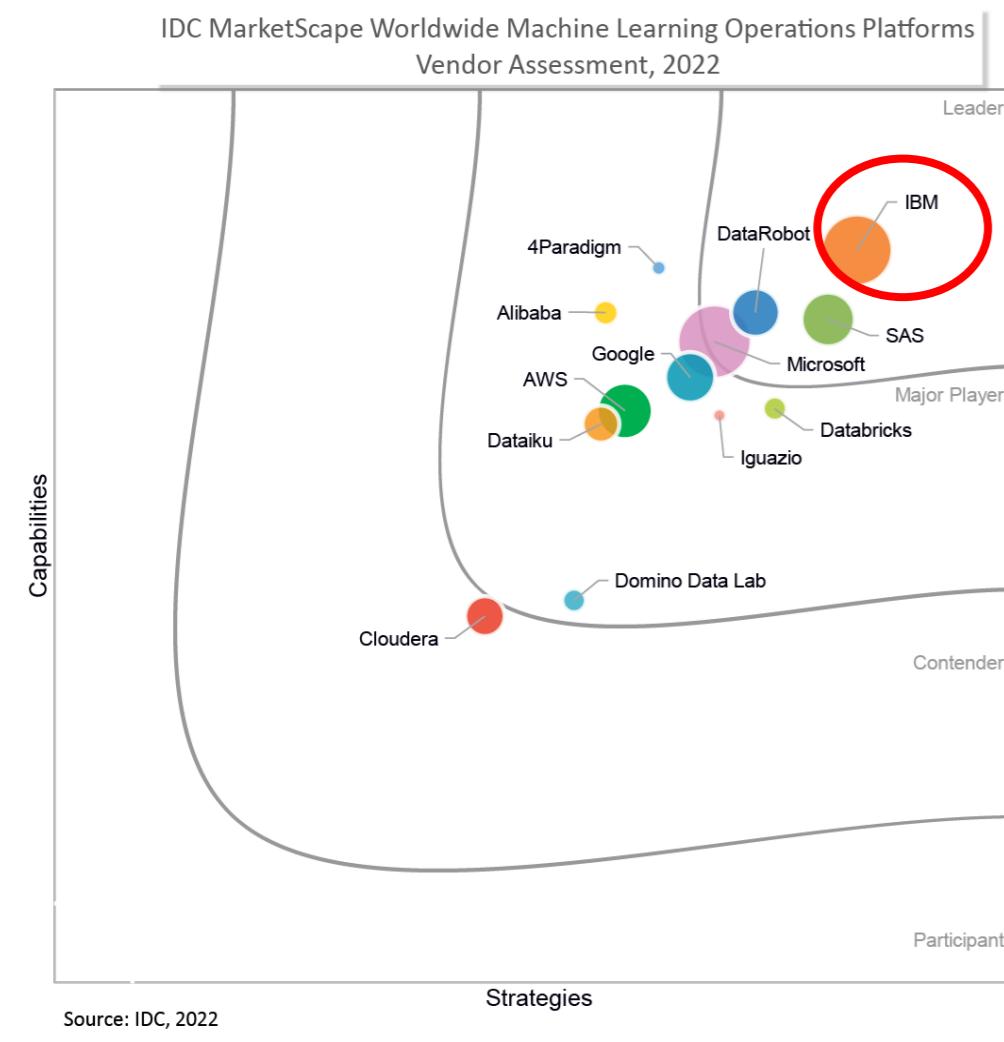
\$298B<sup>1</sup>

Gartner Forecasts Companies to Increase Spend on AI Solutions from 8% in 2023 to 35% by 2027.

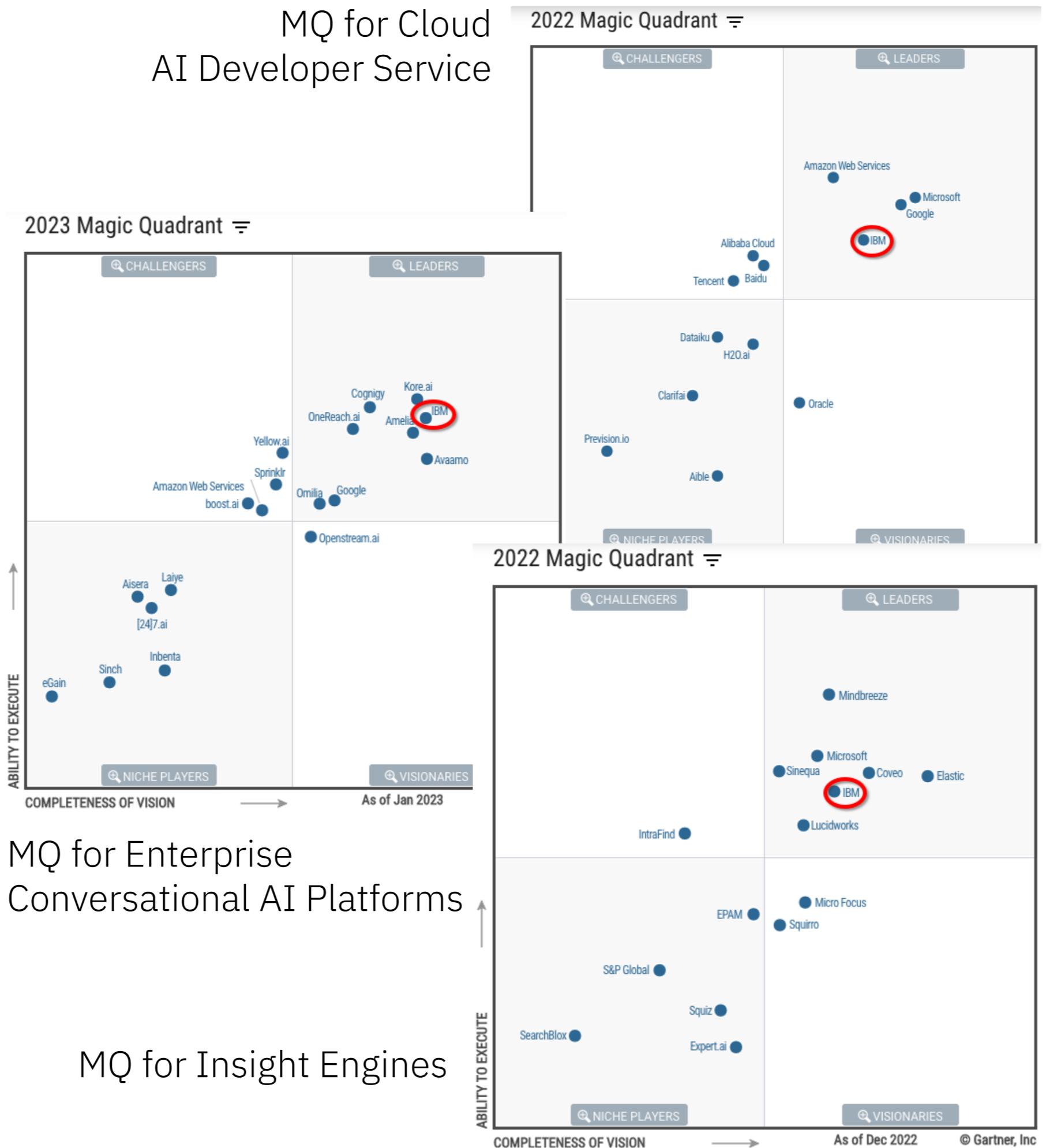
...spend is up but it does not mean that AI always comes easy

↑35%<sup>1</sup>

# IBM is a leader in AI



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

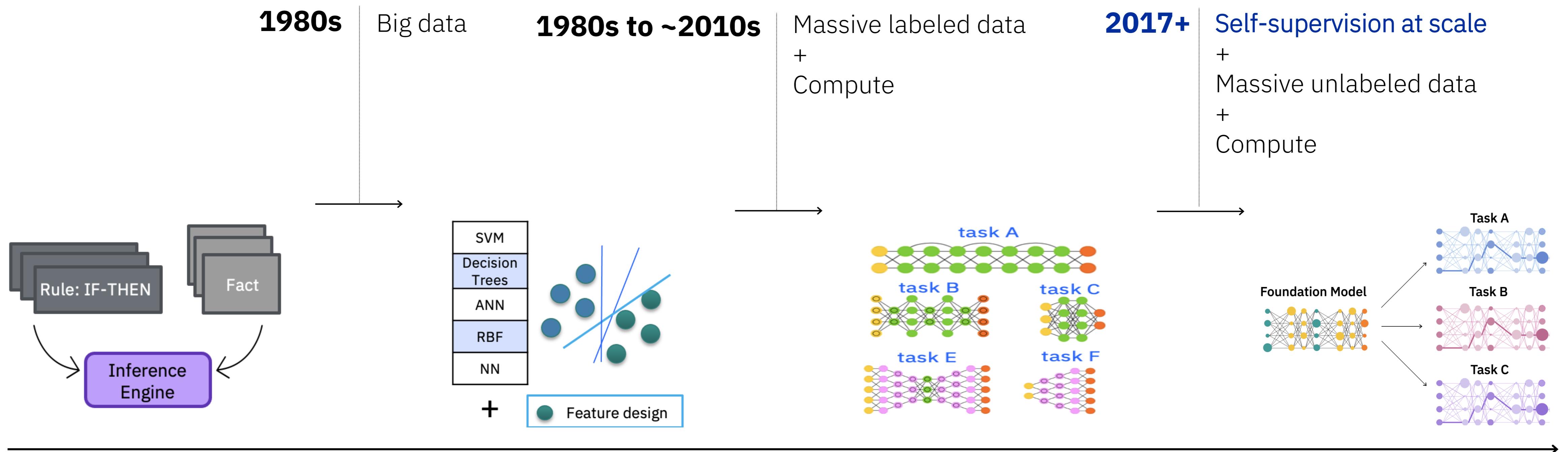


Multiple Gartner Magic Quadrants  
for AI-related capabilities



Forrester Wave:  
Multimodal Predictive  
Analytics and  
Machine Learning

# The story of AI is a story of data representations



## Expert systems

- Rule engines

## Machine learning

- ML algorithms
- Task-specific decision boundaries

## Deep learning

- Network architectures
- Task-specific decision boundaries + feature representations

## Foundation models

- Base model
- Adaptable decision boundaries and feature representations

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

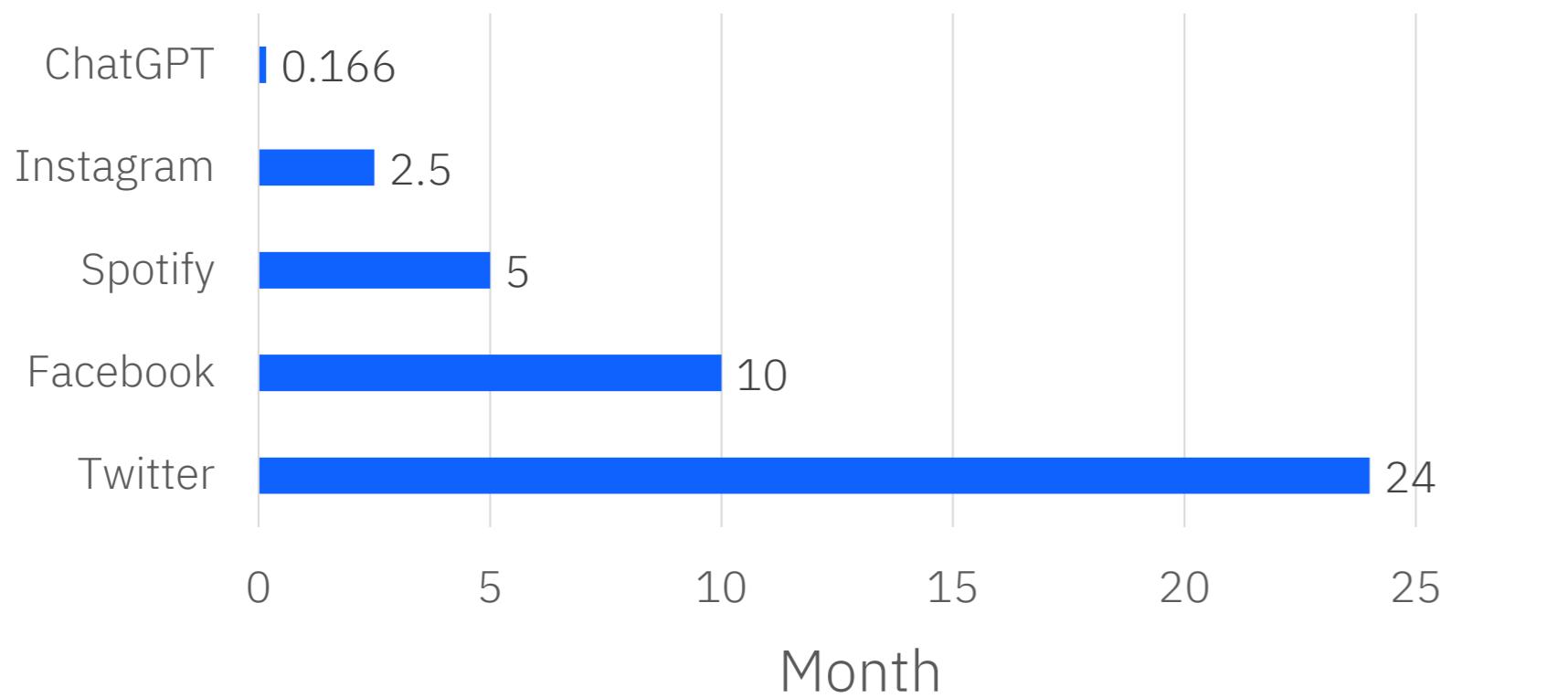
## Massive grassroots adoption



**1 M** users within 1 week

**100 M** users within 2 months

Months to reach 1 M



## Broad-reaching and deep impact

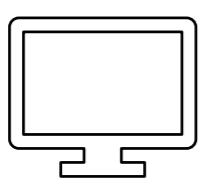
*Generative AI could raise global GDP by 7% within 10 years*

- Goldman Sachs

Generative AI has democratized AI and accelerated the largest commercial opportunity in today's economy, sized at USD 15.7 trillion of GDP by 2030.

- IBM Tech-led disruptions are accelerating, driven by Generative AI

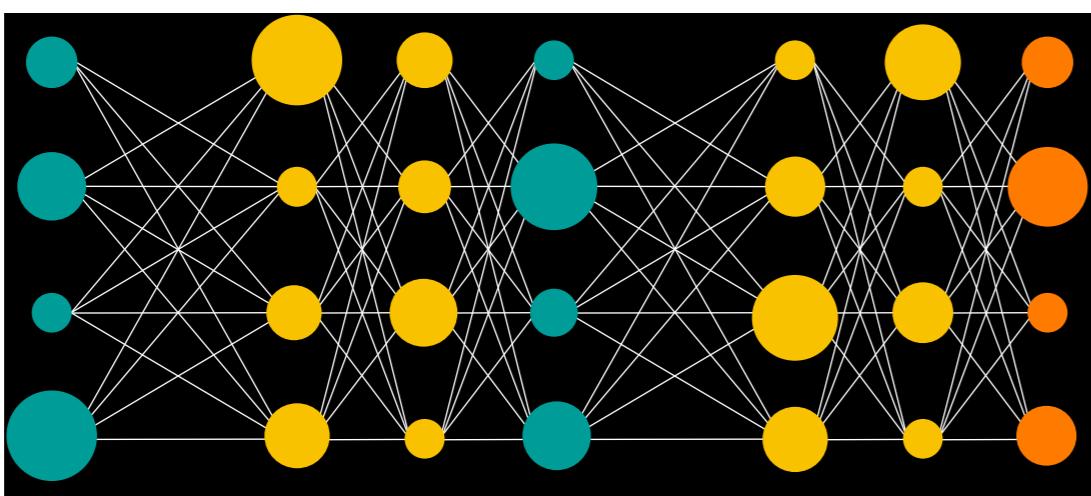
# Foundation models are ...



Self-supervised  
training



↓



Foundation model

Pre-trained

On unlabeled datasets  
of different modalities  
(e.g., language, time-  
series, tabular)

Multiple applications

Able to learn  
generalizable and  
adaptable data  
representations that can  
be effectively used in a  
variety of domains and  
tasks (code generation,  
question answering,  
sentiment analysis)

Self-learning

Systems that leverage  
self-supervised learning

Large language models

A type of foundation  
model trained withy  
language-related data

ChatGPT is based on a  
large language model

# Foundation models: generalizable and adaptable

Translation  
prompt

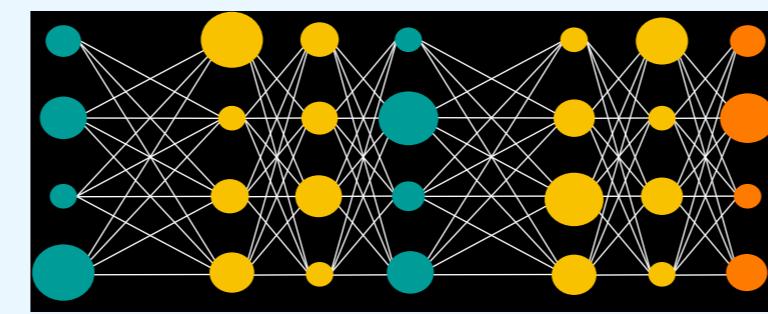
Summarization  
prompt

Answer finding  
prompt

+

Input  
text

+



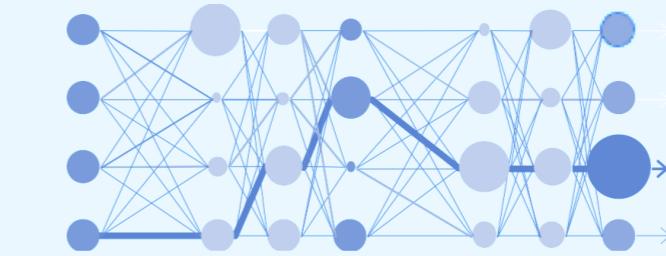
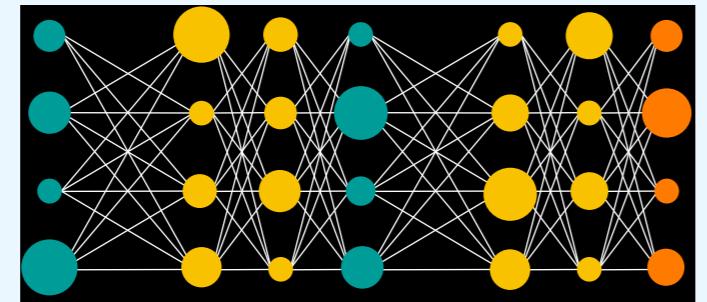
→ Translated input

→ Summarized input

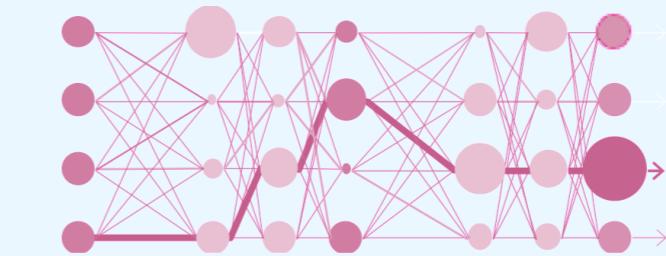
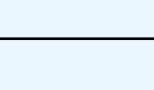
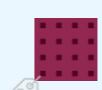
→ Answer to the  
input question



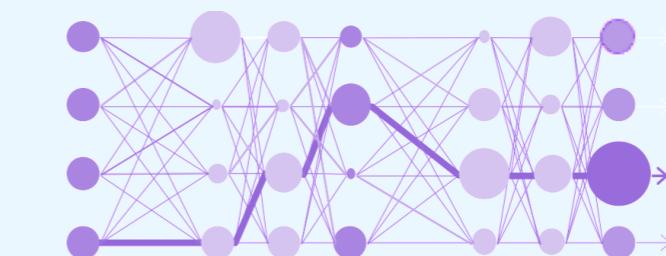
Task-specific  
fine-tuning



Translation  
model



Summarization  
model



Answer finding  
model

# Leveraging foundation model capabilities across various domains

	<b>Customer Care</b> Watson Assistant, Cloud Pak for Data	<b>Digital Labor</b> Watson Orchestrate, Cloud Pak for Integration/Automation, Wisdom in Ansible	<b>IT Operations</b> Turbonomic, Instana, Cloud Pak for Watson AIOps	<b>Cybersecurity</b> QRadar, Cloud Pak for Security
<b>Summarization</b>  Summarizing large documents, conversations, and recordings to key takeaways	<ul style="list-style-type: none"><li>• Call center transcripts</li><li>• Omnichannel journey summary</li><li>• Summarizing search snippets to augment chatbots</li><li>• Summarize events, analyst reports, financial info etc. for advisor</li><li>• Sentiment analysis</li></ul>	<ul style="list-style-type: none"><li>• Summarize documents, contracts, technical manuals, reports, etc.</li><li>• Transcribe videos to text and summarize</li><li>• Summarizing reports on Form 10K</li></ul>	<ul style="list-style-type: none"><li>• Summarize alerts, technical logs, tickets, incident reports, etc.</li><li>• Summarize policy, procedure, meeting notes, etc.</li><li>• Vendor report QBR summarization</li></ul>	<ul style="list-style-type: none"><li>• Summarize security event logs</li><li>• Summarize steps to recap security incident</li><li>• Summarize security specs</li></ul>
<b>Extraction</b>  Extract structured insights from unstructured data	<ul style="list-style-type: none"><li>• Extracting interaction history with clients</li><li>• Extract information from specific types/categories of incidents</li></ul>	<ul style="list-style-type: none"><li>• Extract answers and data from complex unstructured documents</li><li>• Extract information from media files such as meeting records, audio, and video</li></ul>	<ul style="list-style-type: none"><li>• Extract key information from various sources for report automation</li><li>• Extract relevant system/network information for administration, maintenance, and support purpose</li></ul>	<ul style="list-style-type: none"><li>• Extract information from incidents, content for security awareness</li><li>• Extract key security markers and attributes from new threat reports.</li></ul>
<b>Generation</b>  Generate AI to create text	<ul style="list-style-type: none"><li>• User stories, personas</li><li>• Create personalized UX code from experience design</li><li>• Training, and testing data for chatbots</li><li>• Automate responses to emails and reviews</li></ul>	<ul style="list-style-type: none"><li>• Automate the creation of marketing material and language translation</li><li>• Automate image, text, and video creation for articles, blogs, etc.</li><li>• Create automation scripts for various workflows across applications</li></ul>	<ul style="list-style-type: none"><li>• Create technical document from code</li><li>• Automate scripts to configure, deploy, and manage hybrid cloud</li><li>• Co-pilot to create code across multiple programming languages</li></ul>	<ul style="list-style-type: none"><li>• Automate report generation</li><li>• Social engineering simulation</li><li>• Security documentation creation</li><li>• Automate threat detection by looking for anomaly patterns</li></ul>
<b>Classification</b>  For sentiment or topics	<ul style="list-style-type: none"><li>• Classify customer sentiments from feedback or chatbot interaction</li><li>• Classify typical issues raised by clients for focused improvements</li></ul>	<ul style="list-style-type: none"><li>• Classify documents by different criteria – types, contents, keywords</li><li>• Sort digital contents in storage into pre-defined categories</li></ul>	<ul style="list-style-type: none"><li>• Classify incident reports</li><li>• Automate workflow based on analysis of items/status/reports</li></ul>	<ul style="list-style-type: none"><li>• Classify flagged items properly as threats or other categories</li><li>• Classify the type of security risks and find the best response</li><li>• Classify log and other monitoring output to determine the next action</li></ul>
<b>Question answering</b>  Knowledge base search across the company's proprietary data.	<ul style="list-style-type: none"><li>• Knowledgebase articles</li><li>• Augment chatbot w/search</li><li>• Agent assist</li><li>• Contract intelligence</li><li>• Search in technical manuals, HR documents, ethics codes, product documentation, etc.</li></ul>	<ul style="list-style-type: none"><li>• Analyze emails, attachments, documents, invoices, reports, etc.</li><li>• Knowledge search for company information to provide in-house day-to-day assistance and automation</li></ul>	<ul style="list-style-type: none"><li>• Knowledge search for IT helpdesk</li><li>• Ticket resolution by suggesting solutions from resolved tickets</li><li>• Error log and root cause analysis</li><li>• Compliance monitoring</li></ul>	<ul style="list-style-type: none"><li>• Knowledge search across security spec documents</li><li>• External threat intelligence</li><li>• Error log and root cause analysis</li><li>• Security incident search @ forensics</li></ul>

# Generative AI has benefits, but should be used with caution

## Generative AI

- AI techniques that learn a representation of artifacts from data, and use it to generate brand-new, completely original artifacts that preserve a likeness to original data. - Gartner
- Generative AI is trained on foundation models with massive amounts of data.
- The most well-known type of foundation models are large language models (LLMs) used with natural language processing (NLP)

## Great at NLP tasks

Generative AI from LLMs like ChatGPT are good at:

- Creating personal performance, summary reports and resumes
- Passing bar exams
- Generating original write up on various topics with little input
- Providing great chatbot interactions

[But is this what businesses need from generative AI?](#)

## There are risks

- Training requires good AI skills and talent
- Privileged few – to build requires huge R&D and hardware resources available to few vendors
- Easy to be misused
- Black-box nature – insufficient transparency and difficult to understand

[Non-technology concerns, but what should businesses look for in a generative AI platform?](#)

# Generative AI platform questions

Clients cannot just use *any* generative AI platform

## How was it trained?

- Garbage in → garbage out
- An enterprise cannot use a foundation model trained with Web crawl
- The training material needs to be huge and comprehensive but must also be curated

## Can it detect & minimize bias & hallucinations?

- How does the platform detect and correct bias?
- How can it prevent hallucination (providing random and untrue answers with absolute aplomb and convictions)?

## Is it transparent?

- Open vs black-box
- How to audit and explain the model, and the answers it generates?
- Does the model track drift and bias? And how does it address them?

## Does it support regulatory compliance?

- How do foundation models and their usage comply with privacy and government regulations?
- What are the guardrails?
- Who is responsible for an inadvertently exposed PII or a “wrong answer”?

## Is it safe?

- Who has control over the model, input data, and output data?
- How to ensure that confidential information is not given out?
- How is it monitored?
- What safety features and guardrails are in place?

## Can it be customized?

- Hybrid and multicloud?
- Can the model be fine-tuned with clients’ data?
- How can clients update, and extend the model to make it more suitable for their use cases?
- How to integrate with applications? What APIs are in place?

# Put AI to work with **watsonx**

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

## watsonx.ai

Train, validate, tune  
and deploy AI models

## watsonx.data

Scale AI workloads, for  
all your data, anywhere

## watsonx.governance

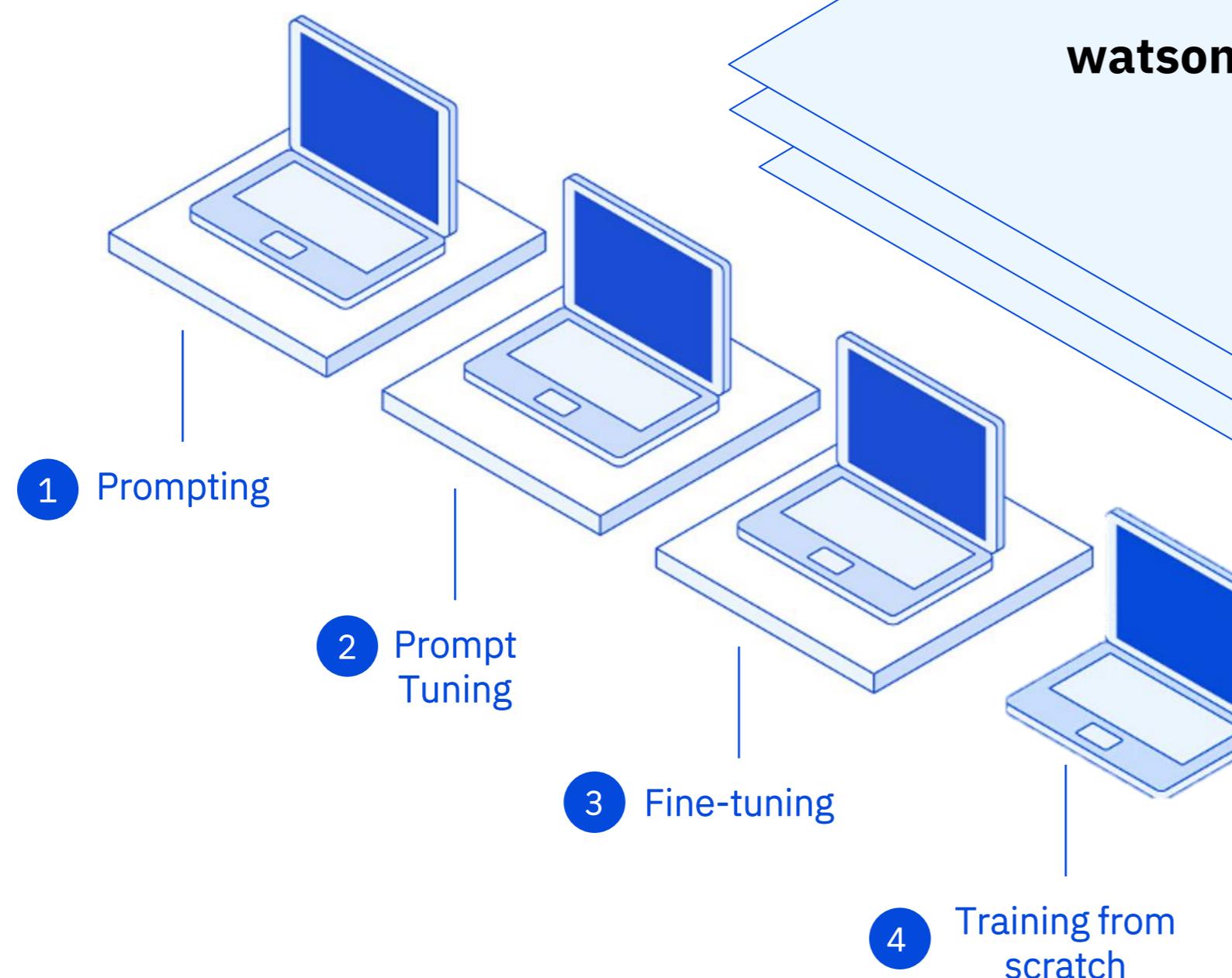
Enable responsible, transparent and  
explainable data and AI workflows

**Red Hat OpenShift** provides scalability, hybrid capability

# watsonx

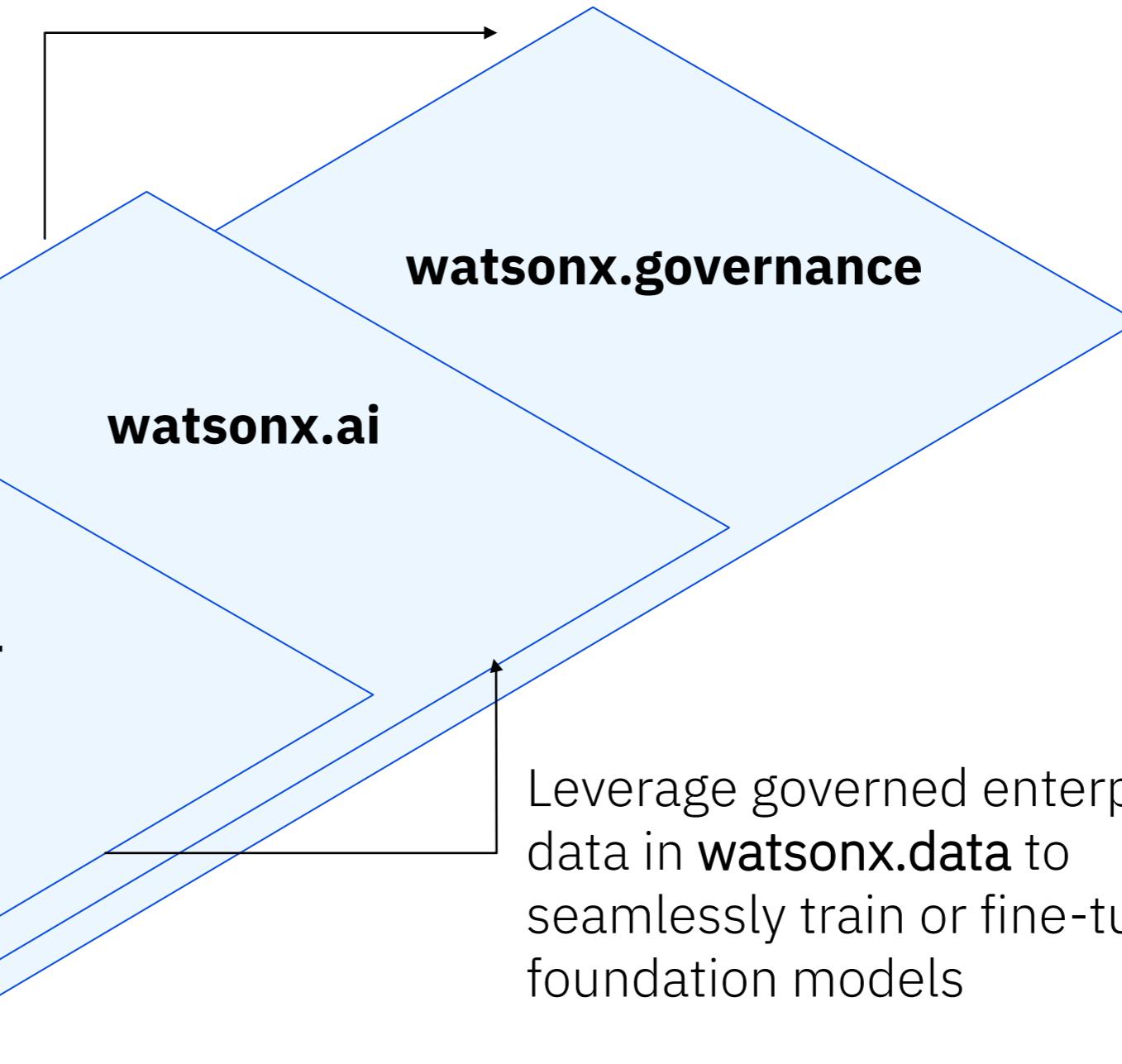
The platform  
for AI and data

Scale and  
accelerate the  
impact of AI  
with trusted data.



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

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



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

watsonx.data

Scale workloads,  
for all your data,  
anywhere

watsonx.ai

Train, validate, tune  
and deploy AI  
models

watsonx.governance

Enable responsible,  
transparent and  
explainable AI workflows

# watsonx.ai – generative AI with traditional AI features

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

## Generative AI capabilities



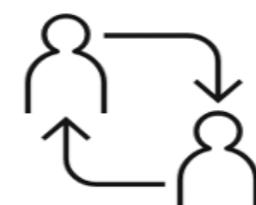
Foundation model library



Prompt lab

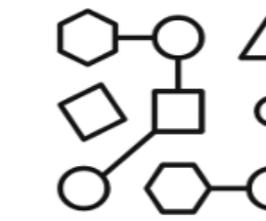


Tuning studio

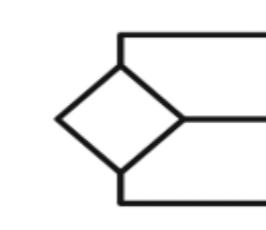


Team collaboration and data preparation

## Plus, a proven studio for machine learning



ModelOps



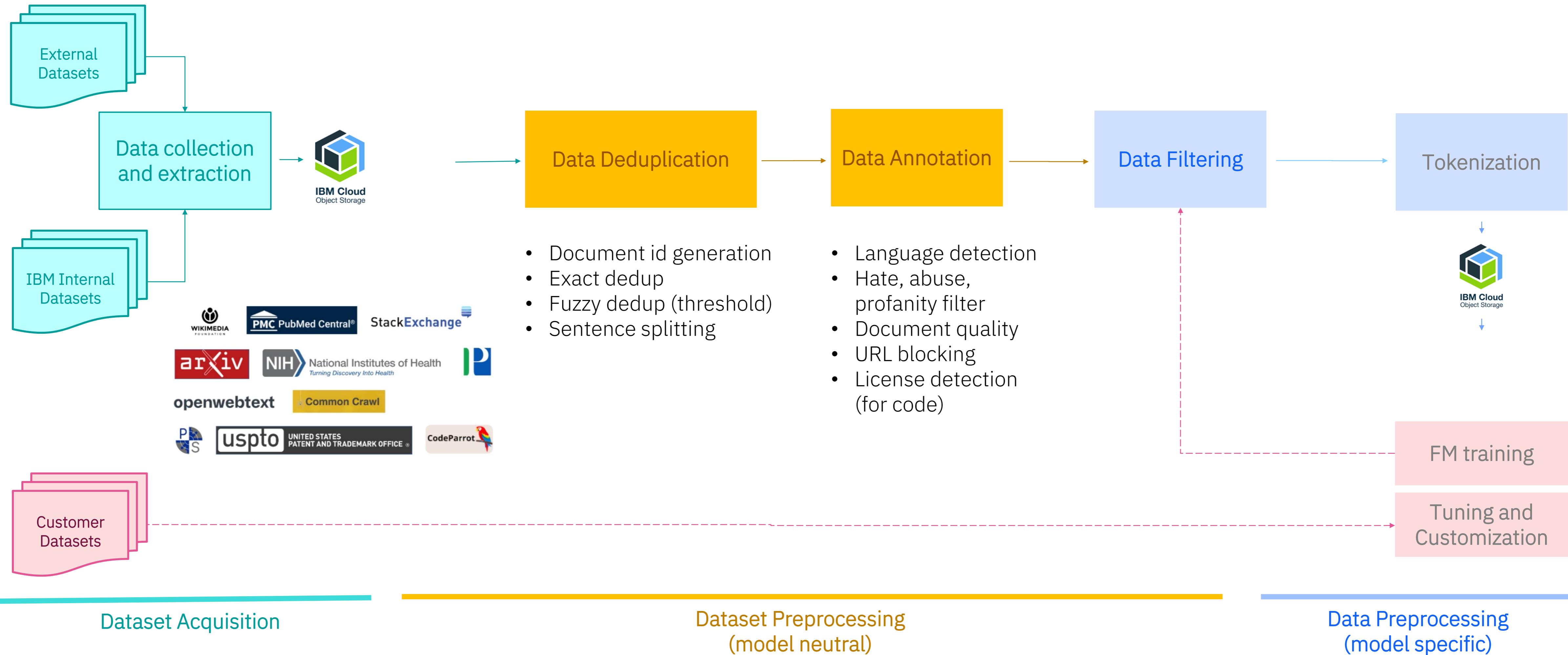
Automated development



Decision optimization

# watsonx.data: curated IBM Data Pile

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

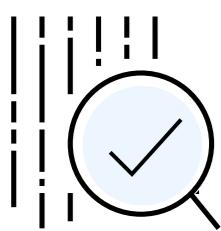


# watsonx.ai Foundation Model Library

Model variety to cover enterprise use cases  
and compliance requirements

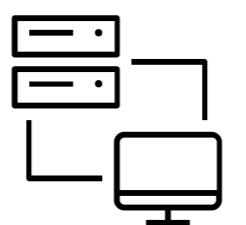
## IBM models

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



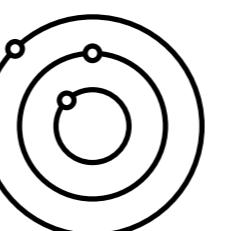
### Transparent Pre-Training on IBM's trusted Data Lake

- 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.

# 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.

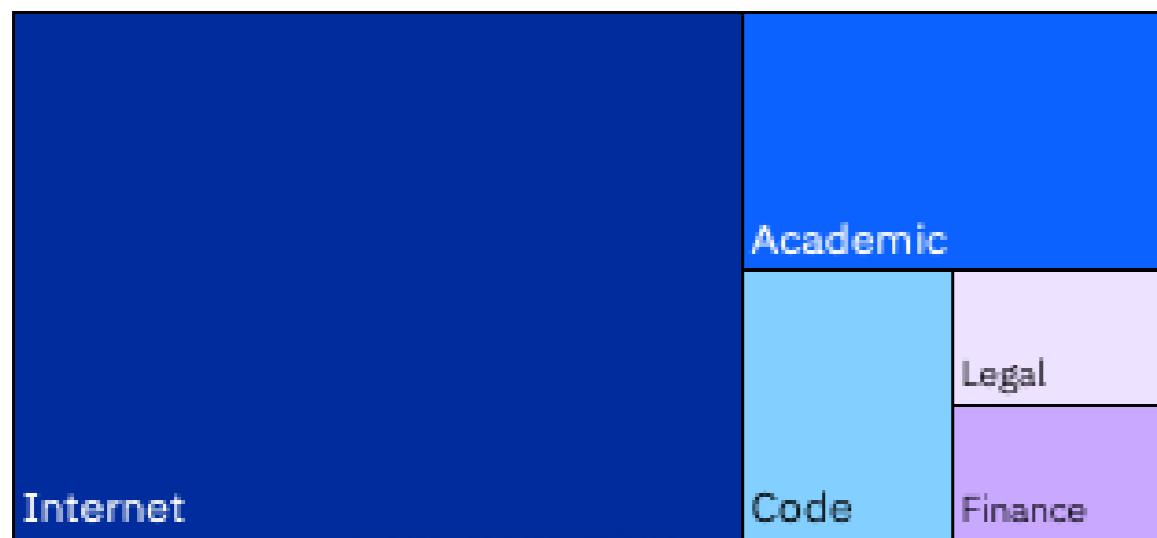
# AI for business - IBM Granite

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.

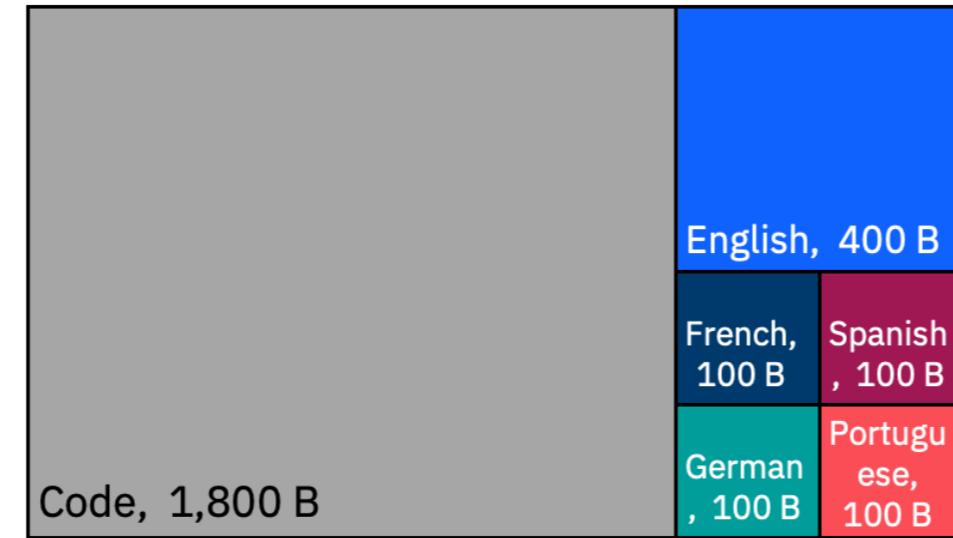
[granite-13b-v2](#) ( English LLM )  
*-chat-v2.1, -instruct-v2*

**13B** parameters in size  
**2.5T** tokens of data



[granite-20b-multilingual](#)

**20B** parameters in size  
**2.6 T** tokens of Data



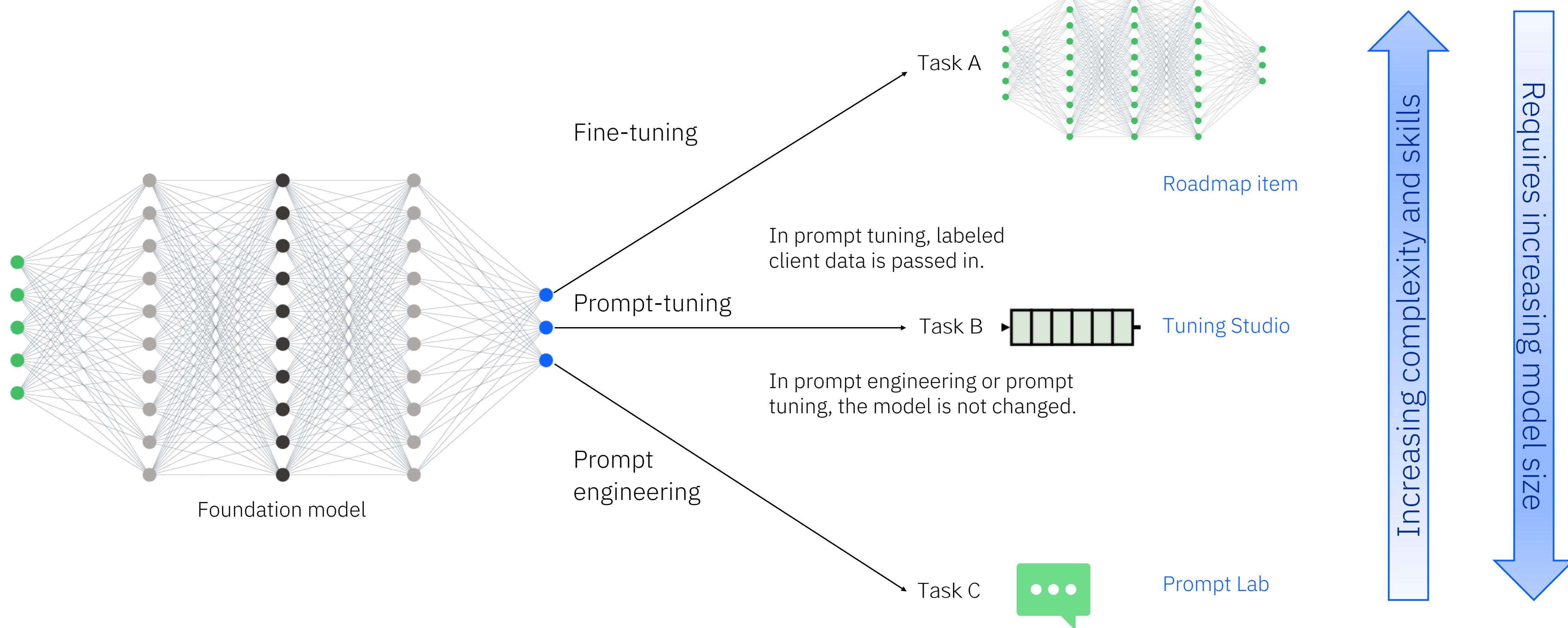
[granite-8b-japanese](#)

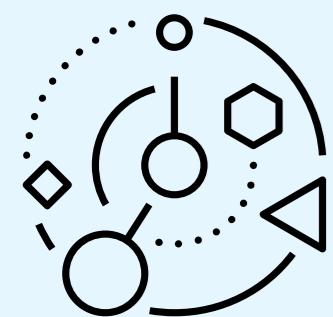
**8B** parameters in size  
**1.6T** tokens of Data



# 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.

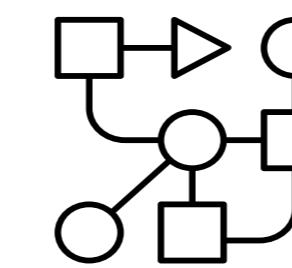




## Prompt engineering

### Benefits to clients:

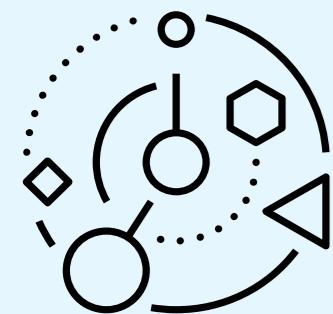
- There is no change to the model (no need for an expensive rebuild)
- No need to pass in labeled data
- Much faster way of training the prompt to properly respond to requests
- Can be passed in via APIs once a prompt has been properly engineered
- Can be used by anyone to guide the model to respond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



# 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

## 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 WatsonX.ai Prompt Lab interface. At the top, it displays 'Projects / Anshupriya's sandbox / Prompt Lab' and 'Sample prompts'. Below this, there are tabs for 'Chat', 'Structured', and 'Freeform', with 'Structured' selected. The 'Model' dropdown shows 'flan-ul2-20b'. On the right, there are buttons for '#', 'TXT', and '</>'. A checkbox for 'AI guardrails on' is checked.

The main area is divided into sections:

- Set up ^**: Contains an 'Instruction (optional)' field with the placeholder "Write a short summary for the meeting transcripts." and an 'Examples (optional)' section with a transcript table.
- Try ^**: Contains a 'Test your prompt' field with the number '1' and a 'Generate' button.

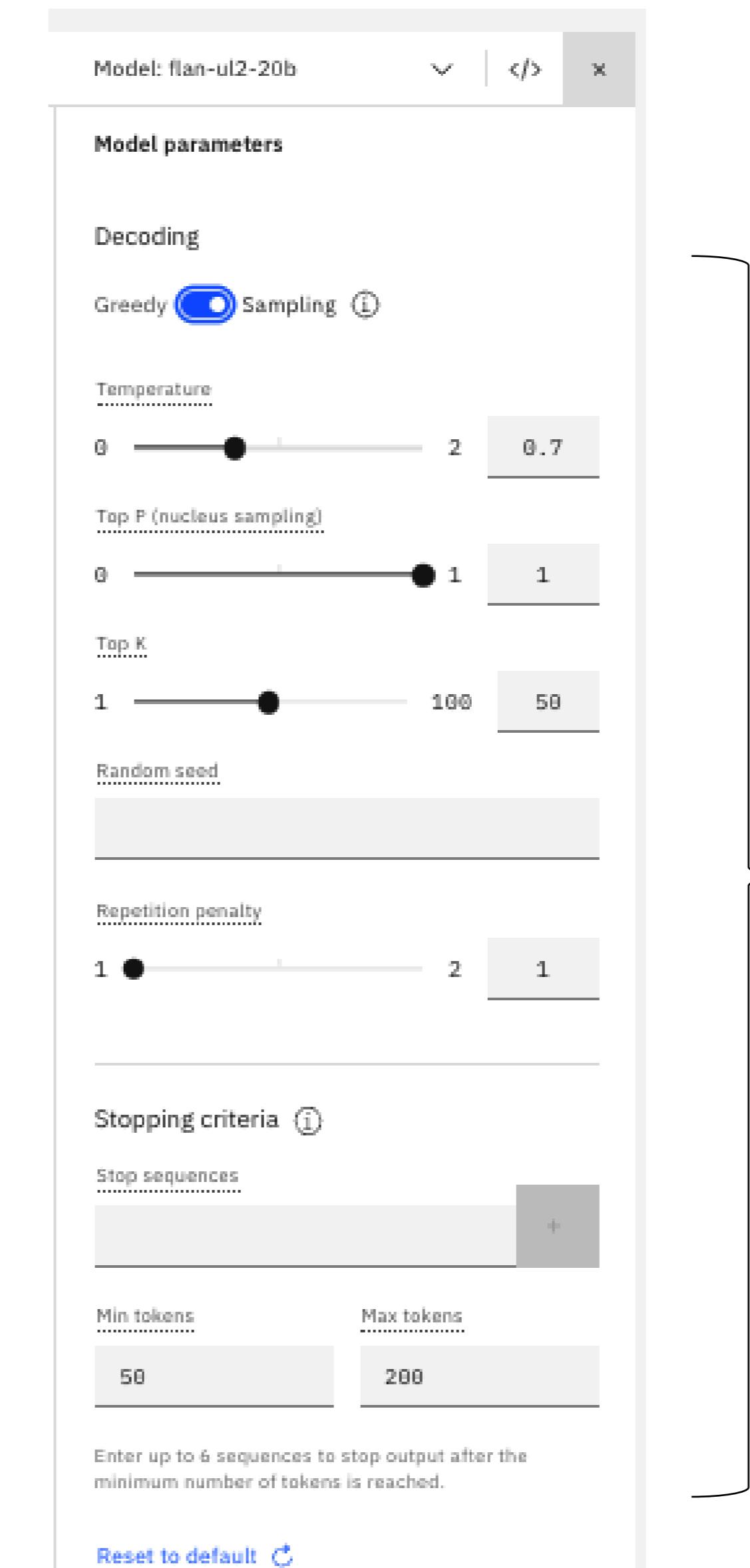
**Sample prompts** (left sidebar):

- Summarization**:
  - Meeting transcript summary**: Summarize the discussion from a meeting transcript.
  - Earnings call summary**: Summarize financial highlights from a quarterly earnings call.
  - Classification**:
    - Scenario classification**: Classify scenario based on project categories.
    - Sentiment classification**: Classify scenario based on project categories.
  - 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.

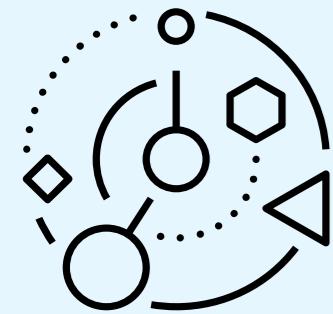
# watsonx.ai: Prompt Lab

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



Parameters for  
prompt engineering



## 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

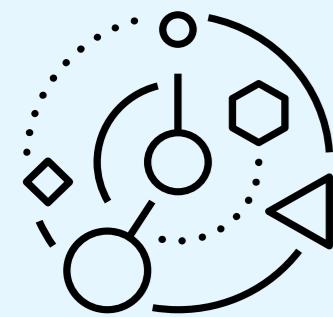
Topic: Introduction to IBM watsonx

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



## 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.

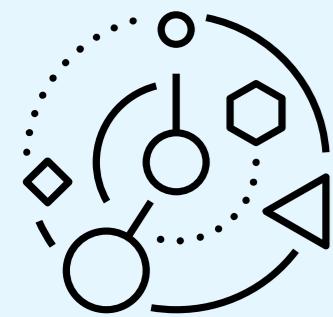
I called your helpdesk multiple times and every time I waited 10-15 minutes before I gave up. The first time I got through, the line got cut suddenly and I had to call back. This is just ridiculous. When I finally got through like after 3 days (yes, 3 days) your agent kept going over a long checklist 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?

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.



# 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: Tuning Studio

Tune your foundation models with labeled data

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.

The screenshot shows the 'Create new tune' interface in the WatsonX.ai Tuning Studio. On the left, a vertical navigation bar lists five steps: 'Setup' (selected), 'Training data', 'Parameters', 'Review and Train', and 'Set up'. To the right, three sections are displayed: 'Set up' (instructions to start by selecting a base model), 'Select a tuning approach' (set to 'Multitask Prompt Tuning'), and 'Select a model to tune' (set to 'flan-t5-xl (3B)'). Below these is a 'Task configuration' section with a dropdown menu set to 'Generation', which is highlighted with a blue border.

# watsonx.ai: Tuning Studio

## 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

Configure tuned model

### Tuning a foundation model (1) ↗

#### Configure details

Which foundation model do you want to prompt tune? ⓘ

Foundation model

flan-t5-xl-3b

How do you want to initialize your prompt? ⓘ

Text

Provide instructions for how to define and format the output.

Setting up a classification model for predicting fraudulent transactions

Random

Let the experiment set the prompt.

Which task fits your goal? ⓘ

Classification

Classify text with up to 10 labels that you specify.

Generation

Generate text in the same format as your training data.

Summarization

Summarize text in the same format as your training data.

# 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: 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.

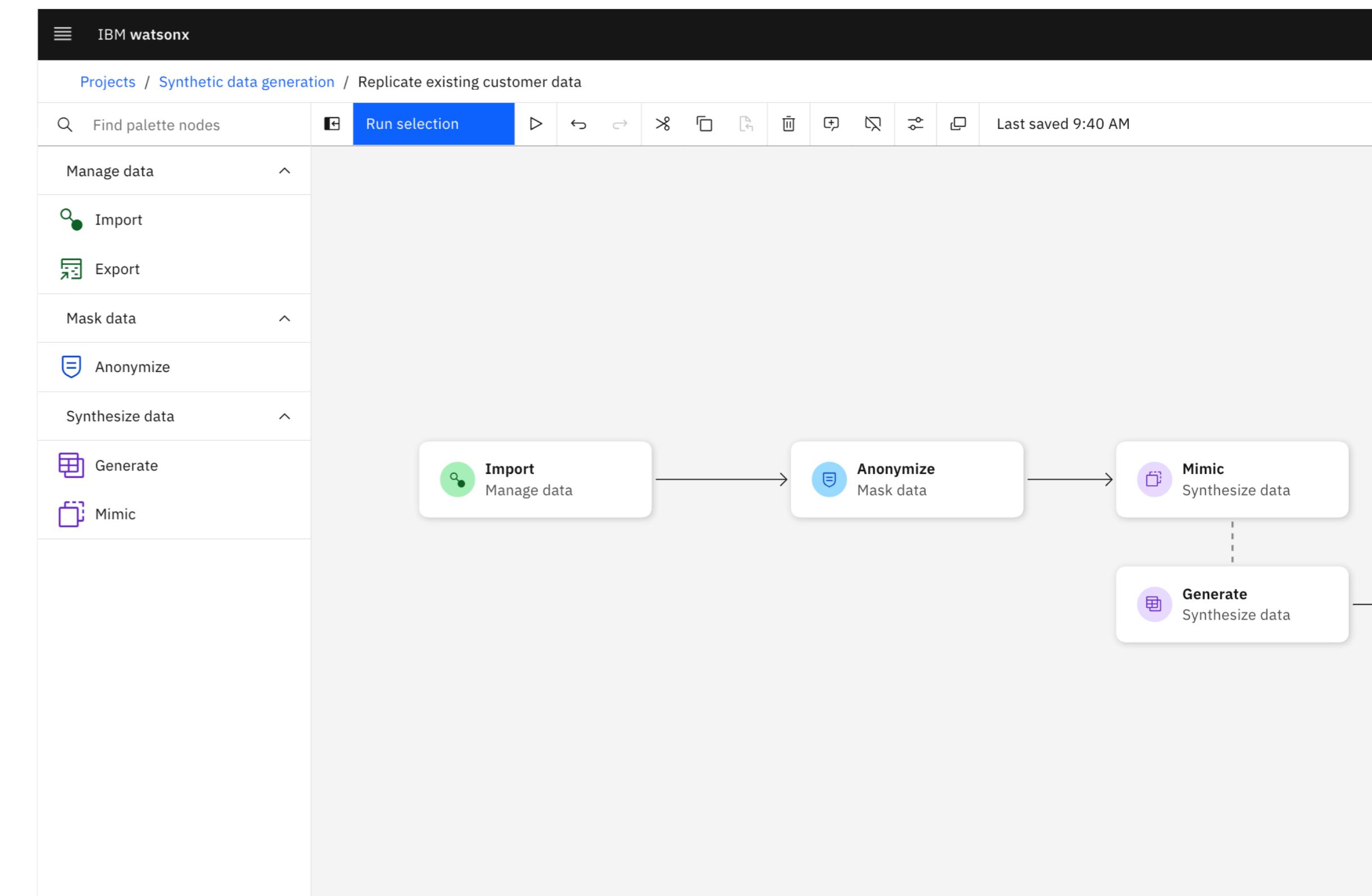
Create synthetic data using your existing data in a database or by uploading a file. If no data exists or can't be accessed, you can design your own data schema.

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.



# watsonx.ai

Value propositions -  
Addressing the needs  
of different persona

**watsonx.ai** caters  
to different types  
of AI builders



## Data Scientists

Data scientists leverage **watsonx.ai** to work with data, generative AI, and traditional machine learning (ML) to solve business problems.

### Key responsibilities:

- Data scientists help to figure out business problems and determine what data is applicable to solving those problems.
- They sometimes retrieve and clean data.
- They work with AI models to solve problems — creating traditional ML models, as well as working with foundation models to prompt engineers and tune models.
- They sometimes deploy AI models and monitor their lifecycle.



## Developers

Developers leverage **watsonx.ai** to work with foundation models and create prompts — enabling them to grab production-ready code they can put to use in an application.

### Key responsibilities:

- Developers develop and deploy software with high-quality and compliant code.
- They automate processes with the use of generative AI to avoid repetitive and menial work.
- Previously developers were not often able to work with ML, but generative AI enables developers to now work with foundation models.



## Analysts

Analysts leverage **watsonx.ai** to work with generative AI to automate their workflow. They also ensure that AI models accurately generate outputs that reflect their specific domain.

### Key responsibilities:

- Analysts review generative AI model outputs to ensure that the foundation model is generating accurate responses relevant to the business context.
- They use various tools to showcase and share reports, including AI tools that help automate their workflow.
- Their reports cover actionable business insights that are guiding others in making decisions for their company.
- They support others to assess the data available, understanding the existing processes, and resolving data issues.

Works closely with other IBM technology pillars

## Automation

IBM brings together AI, business process capability, and industry knowledge to support complex natural language interactions with users, process mining integrated with IT observability, and improved automation accuracy with governance, auditability, and trust.

## Hybrid Cloud

Provide generative AI capabilities everywhere, on-premises or in the cloud. Aiming to dynamically discover, deploy, and compose infrastructure and services via scalable advances in AI, automation, and programming models.

## Data

IBM **watsonx.ai** works closely with **watsonx.data** to provide a tightly integrated ecosystem:

- **watsonx.data** provides governed data for model development and training
- **watsonx.ai** provides generative AI models to analyze workloads

## Security

AI is used to create a unified security workflow across the multitude of security tools most companies need to manage and reduce incident response times. AI helps to protect data across the hybrid cloud and simplify compliance.

# watsonx.ai competitors

Multiple vendors are offering generative AI services. Here are some top competitors:

- Hyperscalers
  - AWS (Bedrock)
  - Azure (Azure OpenAI Service)
  - Google Vertex (Generative AI Studio)
  - Nvidia \*
- Others
  - AI21 Lab
  - Anthropic
  - Cohere
  - Databricks
- More will be investigated

ChatGPT is a chatbot – and this is not the top focus of IBM generative AI. As such, OpenAI is not a direct competitor of IBM.

Context is important in looking at **watsonx.ai**

- It is important to look at the ecosystem – in particular how vendors are addressing clients' main concerns in responsible AI
- Not everyone has the same terminology
  - Model tuning in Google is fine-tuning
- For competitor details, consult the **watsonx.ai** competitive Insight presentation [here](#).

# Price comparison points

The following table compares the cost of running the llama 2 Chat (70b) model from watsonx.ai in a summarization tasks (summarizing a 30-minute call into a one-page completion) against competitors:

- AWS Bedrock running the same model
- OpenAI GPT4
- OpenAI GPT4–turbo

Model	Price per 1,000 input tokens	Price per 1,000 output tokens	Inference Cost
Llama 2 Chat (70b) on watsonx.ai	0.0018	0.0018	0.0119988
Llama 2 Chat (70b) on AWS Bedrock	0.00195	0.00256	0.01340496
OpenAI GPT4	0.03	0.06	0.21996
OpenAI GPT4-turbo	0.01	0.03	0.07998

Inference cost: 6,000 input tokens and 666 completion tokens. Using llama 2 Chat (70B) on watsonx.ai as an example:

$$\text{Inference cost} = (6,000 / 1,000) \times 0.0018 + (666 / 1,000) \times 0.0018 = 0.0119988.$$

Watsonx.ai  
price  
advantage

**11%**

Less expensive than running  
Llama2 Chat (70b) on  
Amazon Bedrock

---

Summarize  
task using  
Llama 2 Chat (70b)

**18.3X**

Less expensive than GPT4

---

**6.6X**

Less expensive than GPT4-turbo

# Success story: Customer Success Management company

PoC with IBM watsonx.ai

Use case: RAG and Summarize

Competitor: GPT 3.5

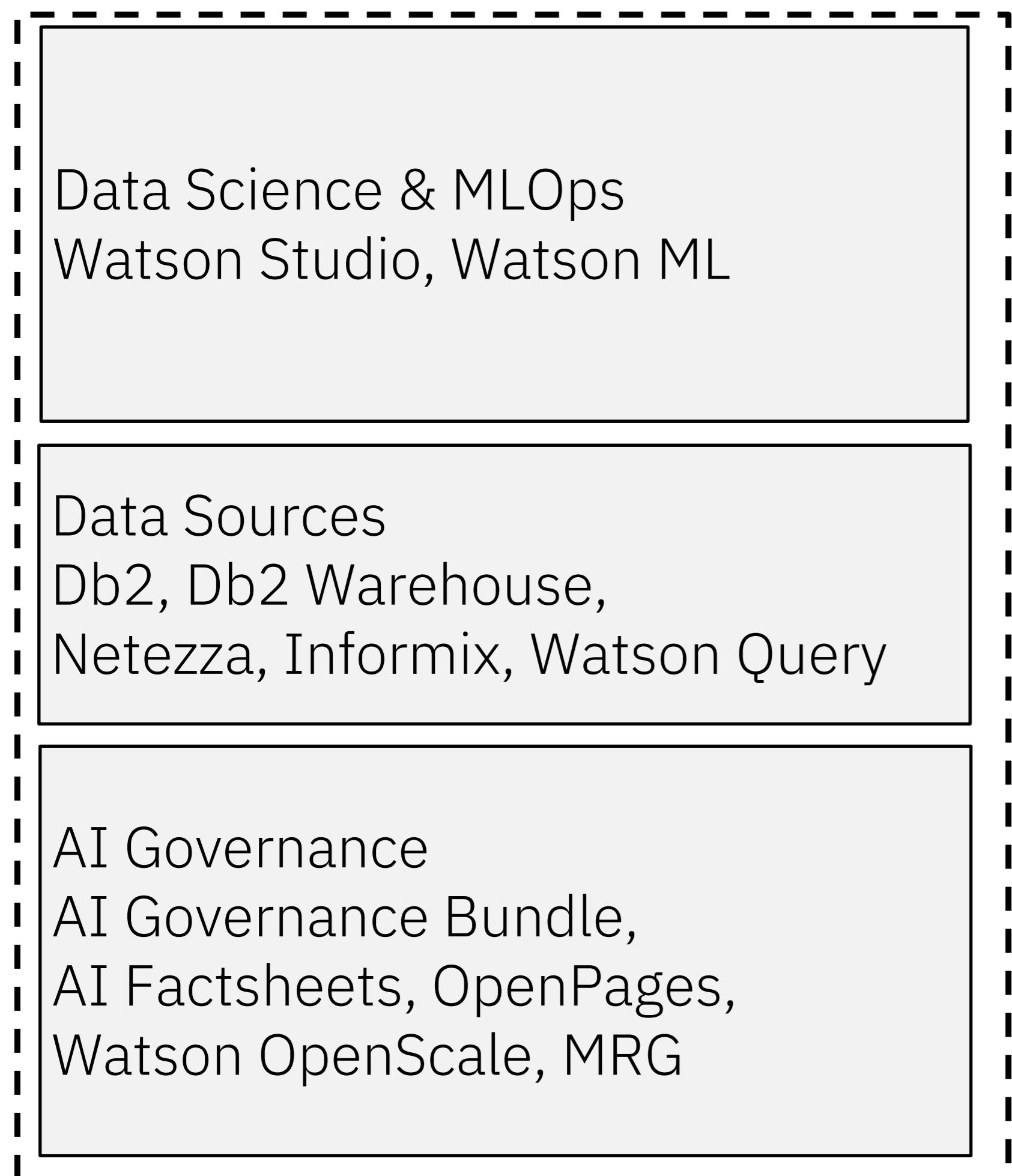
Successful PoC showing:

- 33% savings using llama 2 13B model
- Secure, dedicated, and robust LLM model workbench and tuning platform hostable within client's firewalls
- Simple, single, standard API/SDK entry point for different watsonx.ai models



# Connecting current software portfolio with **watsonX**

## Relevant CP4D and Cartridge Use Cases



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\*\*

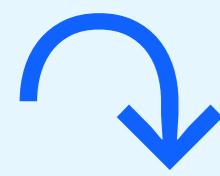
### watsonX.governance

- AI Factsheets\*
- OpenPages\*
- OpenScale\*

\* capabilities that will be enhanced in **watsonX**

\*\* brand new capabilities to **watsonX**

# Let's create a Pilot



## What does IBM offer?

A no-cost IBM multi-disciplinary team and expertise to jointly innovate and rapidly prove generative AI solutions to your business needs, leveraging watsonx.

## What value do clients get?

A model that delivers the agreed upon pilot scope. Confidence that generative AI with watsonx can deliver a technical solution to your compelling business needs with trust and meeting your enterprise requirements.

## What is your commitment?

Your business and technology context, sponsorship, subject matter experts, and data.

## IBM watsonx Pilot Team

600+

AI Engineers

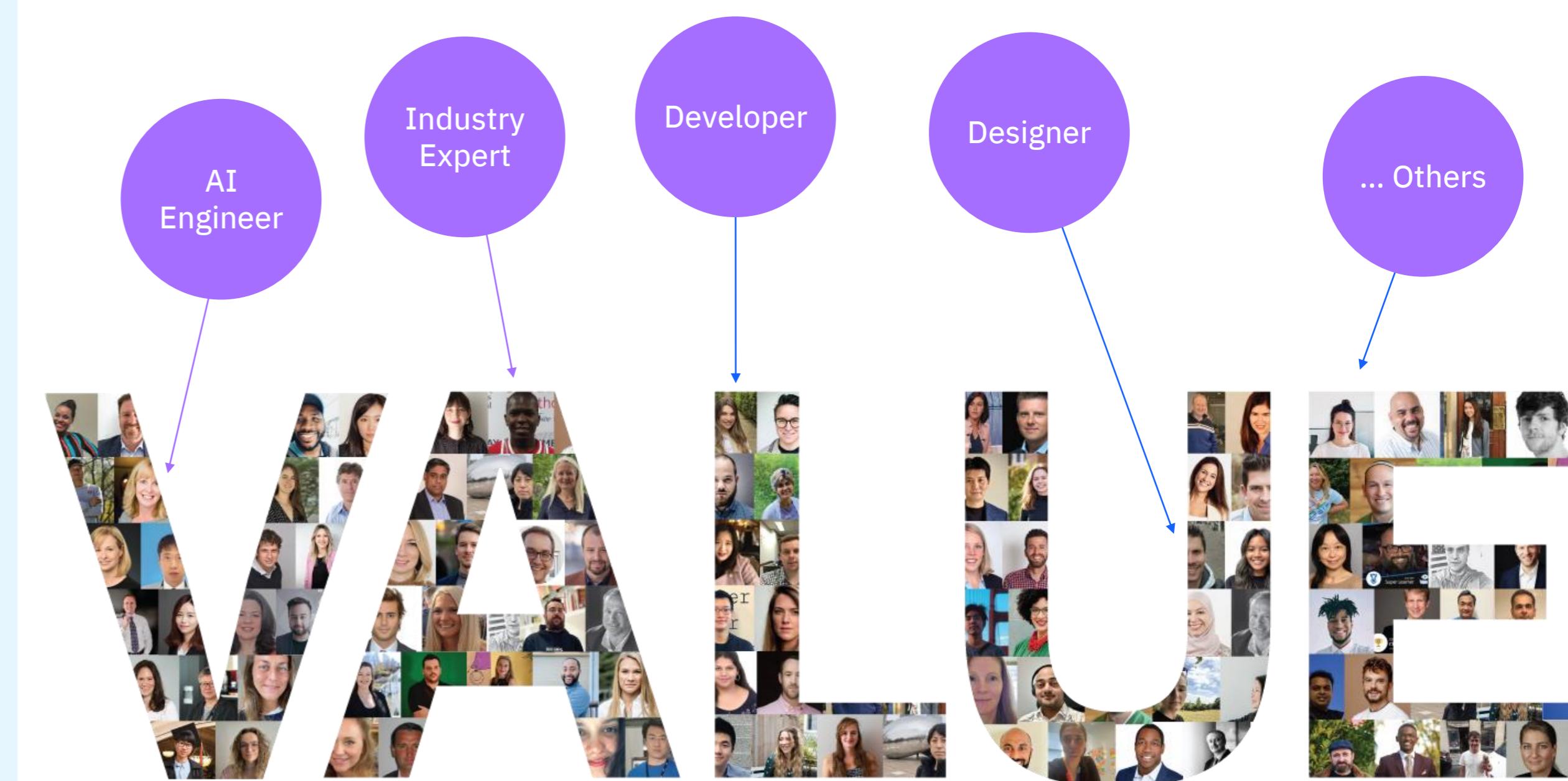
100s

Domain & Industry Experts

50+

Countries

Supported by  
IBM product and  
research teams



# Client path to watsonx.ai adoption through Technology Expert Labs →

1

Attach pre-scoped offerings  
to all new license sales

2

Expand with Expertise  
Connect

3

Engage Expert Labs  
Services Sellers early  
in the Software Sale

## Expert Labs Contacts

[Sarah Memon](#)

Services Product Manager, Data and AI

[Ted Trask](#)

Program Director, WW Data & AI Services  
Sales Leader

[Charley Beller](#) WW Solution Engineering Lead  
Data Insights & Applications

[Tal Shahar](#) WW Principal Delivery Practice  
Leader

Seller Locator [Tool](#)

Slack: [#ask-expert-labs](#)

Offering	Retail Price	Duration	Transact by SaaS	Transact by Non-SaaS
<a href="#">Build watsonx.ai</a> Conversational Search	\$90K USD	5W	Qty 5 x #D0G4ZX	NA
<a href="#">Build watsonx.ai</a> Summarization	\$130K USD	5W	Qty 7 x #D0GC4ZX	NA
<a href="#">Expertise Connect</a>	\$90K USD	6M	Qty 1 x D0676ZX	

# Technology Expert Labs **watsonx.ai** Companion Services

<b><u>Build Offering</u></b>	<b><u>Expertise Connect</u></b>	<b><u>Custom Services</u></b>
<b>5 weeks</b>  <ul style="list-style-type: none"><li>✓ Services offering to assist with technical implementation of watsonx.ai technology and integrate with the client IT environment.<ul style="list-style-type: none"><li>1. Infuse and optimize generative AI into customer care business processes &amp; applications</li><li>2. Additional integrations such as: Digital Avatars, Neural Seek</li><li>3. Use case implementation &amp; validation</li><li>4. User acceptance testing &amp; go live support</li></ul></li><li>❖ <b>Outcome:</b> Production level deployment of a business application and AI model</li></ul> <p>Supported Use Cases: Conversational Search, Summarization</p>	<b>1 Year</b>  <ul style="list-style-type: none"><li>✓ Subscription-based service that provides access to deep technical expertise in design, deployment, operations, and transformation while building your client's capacity and skills</li><li>❖ <b>Expertise Connect</b> is the insurance policy that helps clients succeed with IBM technologies and allows them to <b>move forward</b> in their IBM journey.</li></ul>	
		<ul style="list-style-type: none"><li>✓ Tailored services to help customers with additional LLM &amp; generative AI use cases, prompt-engineering and fine tuning with watsonx.ai<ul style="list-style-type: none"><li>▪ Planning</li><li>▪ Solution Design</li><li>▪ Studio Configuration</li><li>▪ Model Development &amp; Optimization</li><li>▪ User Acceptance Testing</li><li>▪ Go-Live Support</li><li>▪ Enablement</li></ul></li></ul> <p><a href="#"><u>Engage Solution Engineering</u></a></p>

\*On-prem install launch aligned with product availability (2024)

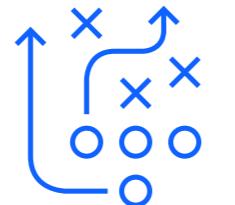
Have Questions?



#ask-expert-labs

Get your clients  
started with  
**watsonX** today

Three ways you  
can team up with  
**IBM Consulting** →



### **½ DAY STRATEGY BRIEFING**

- **Objective:** Initial discussion to learn more about IBM's **watsonX.ai** POV
- **Attendees:** Clients: LOB, IT, and IBM: generative AI expert, Designer
- **IBM Tech:** Client Engineering
- **Outcome:** Determine if generative AI can be leveraged now within the organization
- **Duration:** 2-4 hours
- **Fee:** Free



### **GARAGE DESIGN SESSION**

- **Objective:** Collaborative design session to determine the **watsonX.ai** roadmap
- **Attendees:** Clients: LOB, IT, CDO office and IBM: BTS, HCS based on client's focus, 30-50% Client Engineering
- **IBM Tech:** Client Engineering
- **Outcome:** generative AI **watsonX** roadmap of prioritized use cases for transformation
- **Duration:** 2-4 weeks
- **Fee:** \$150K



### **PILOT PROGRAM**

- **Objective:** Deliver results based on selected use case(s) leveraging **watsonX.ai** capabilities
- **Attendees:** Clients: LOB SMEs, IT, CDO office and IBM: BTS, HCS based on client's focus
- **IBM Tech:** Pre-GA software & Client Engineering
- **Outcome:** generative AI **watsonX** model to prove results for the selected use case(s)
- **Duration:** 4-6 weeks
- **Fee:** \$500K

# 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 L1](#)
- [Watsonx L1](#)



# Resources

## Seismic



- [watsonx](#)
- [watsonx.ai](#)
- [watsonx.data](#)
- [watsonx.governance](#)

## Internal



- [watsonx blog](#)
- [watsonx sales tactics](#)
- [SalesZone AI for Business](#)
- [Behind the news](#)

## External



- [watson](#)
- [watsonx](#)
- [watsonx.ai](#)
- [watsonx.data](#)
- [watsonx.governance](#)

## Enablement



- [AI for Business learnings](#)
- [Generative AI enablement series](#)
- [watsonx rally hosted by Rob Thomas & Dario Gil](#)
- More to come...

## Events



- [watsonx.ai tech preview](#)  
(clients)
- [watsonx.ai alpha environment](#)  
(IBMer)

## FAQ



<https://ibm.seismic.com/LInk/Content/DCGq8T9hmFHbF82P2PPmQH8jpmBB>

