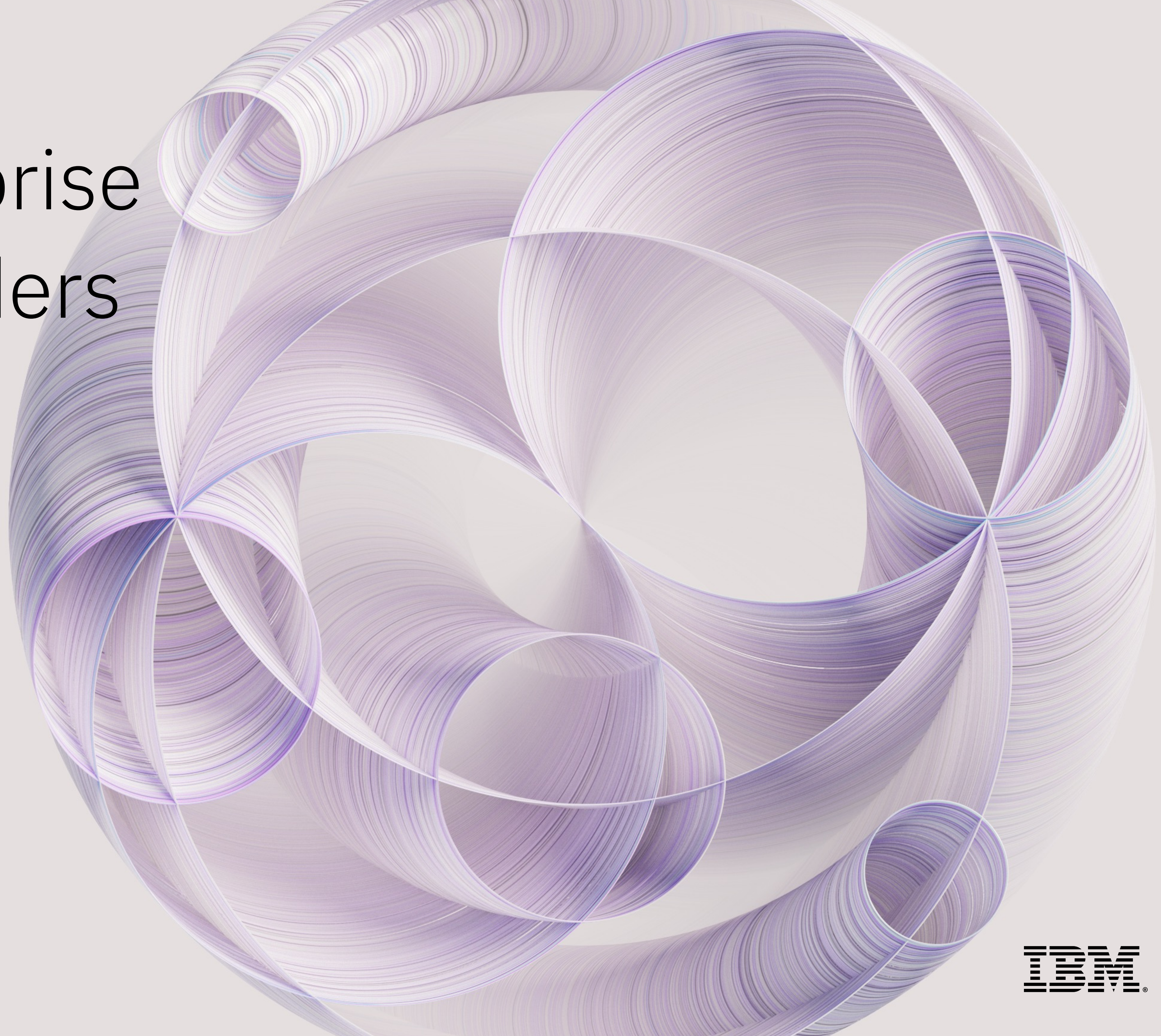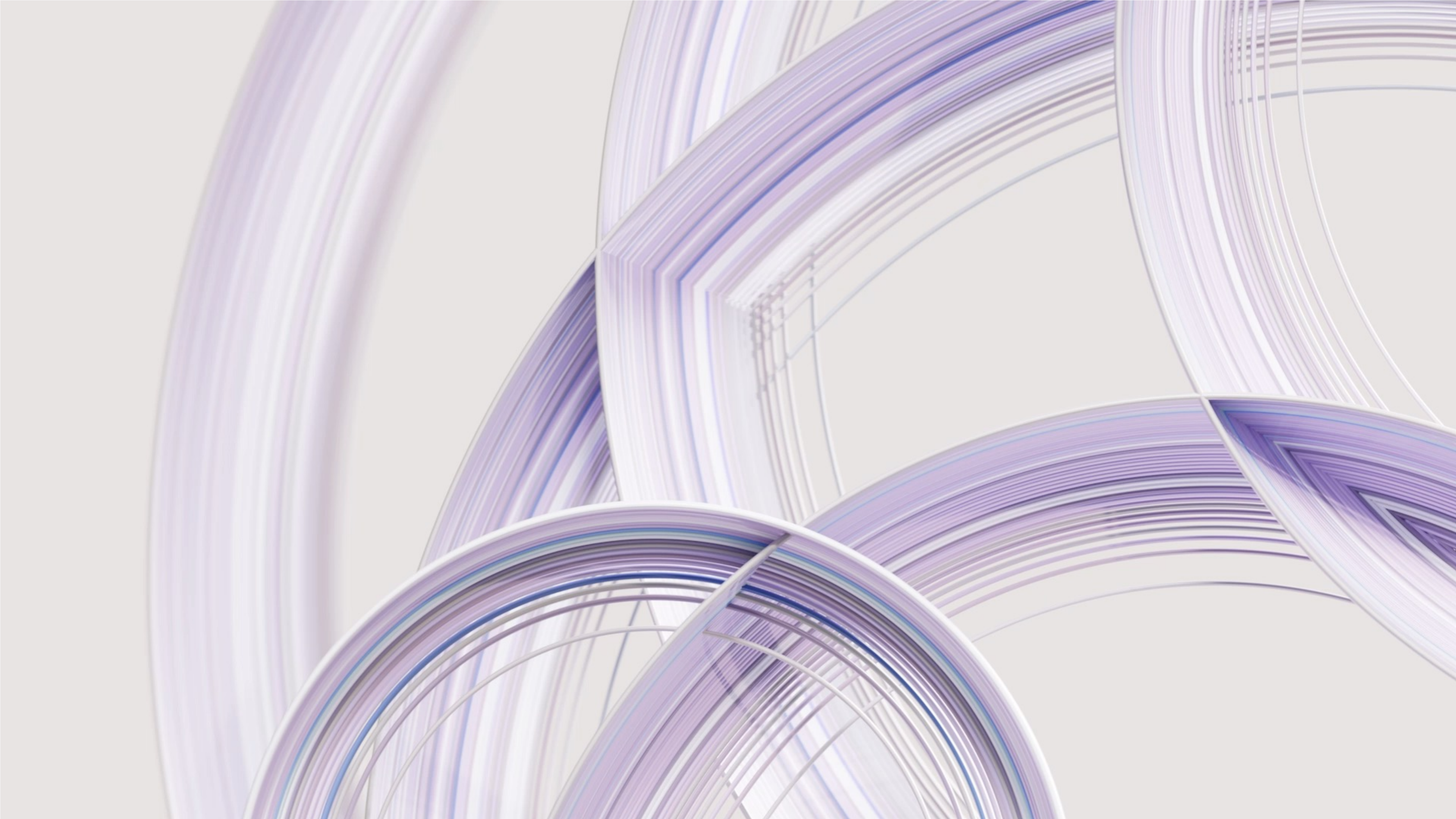# watsonx.ai: next-generation enterprise studio for AI builders
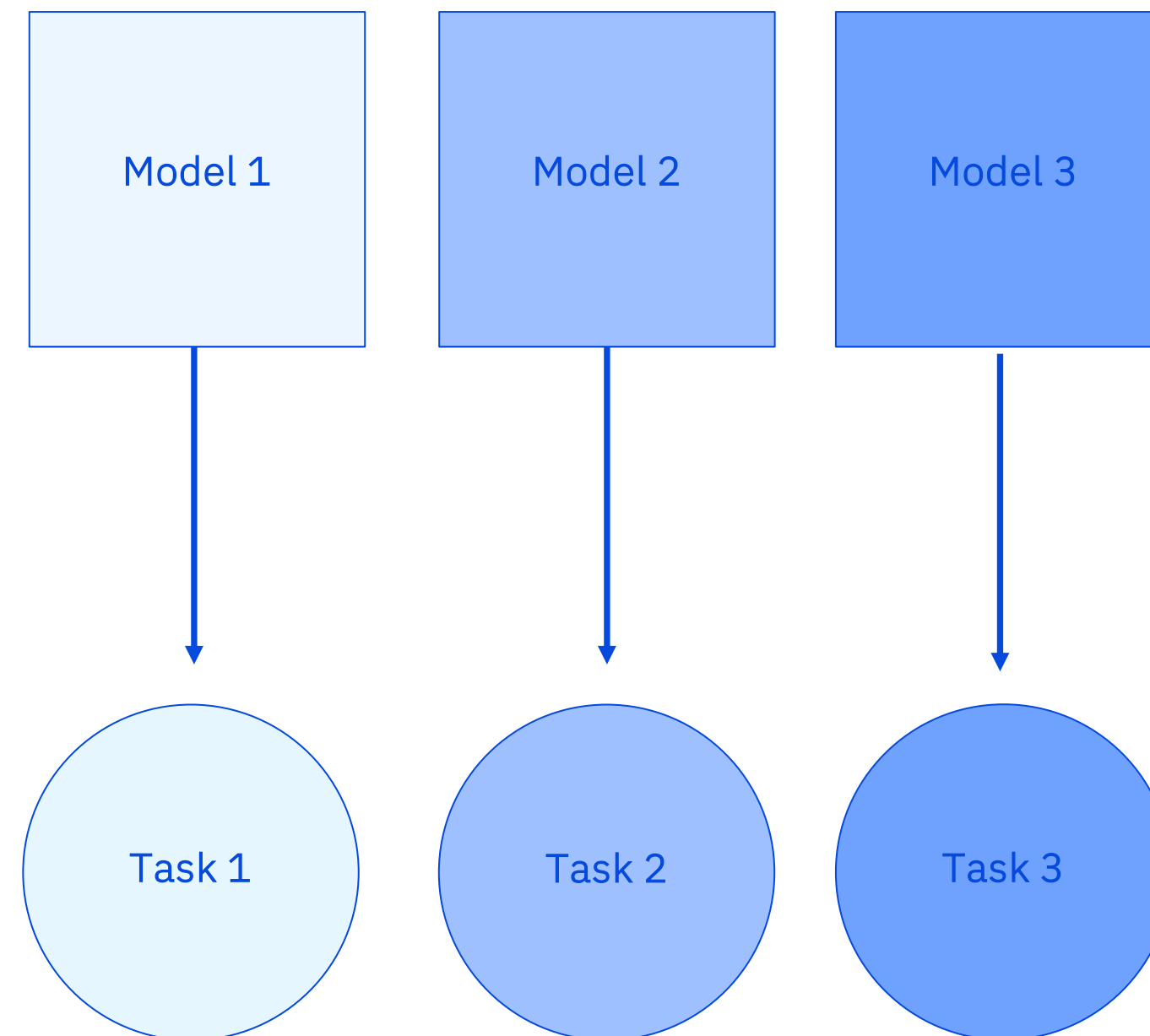
IBM

A new era for AI

Generative AI and foundation models are making massive AI scalability possible.

A foundation model is an AI model that trains on large amounts of general data, which can be adapted easily to new scenarios and use cases.
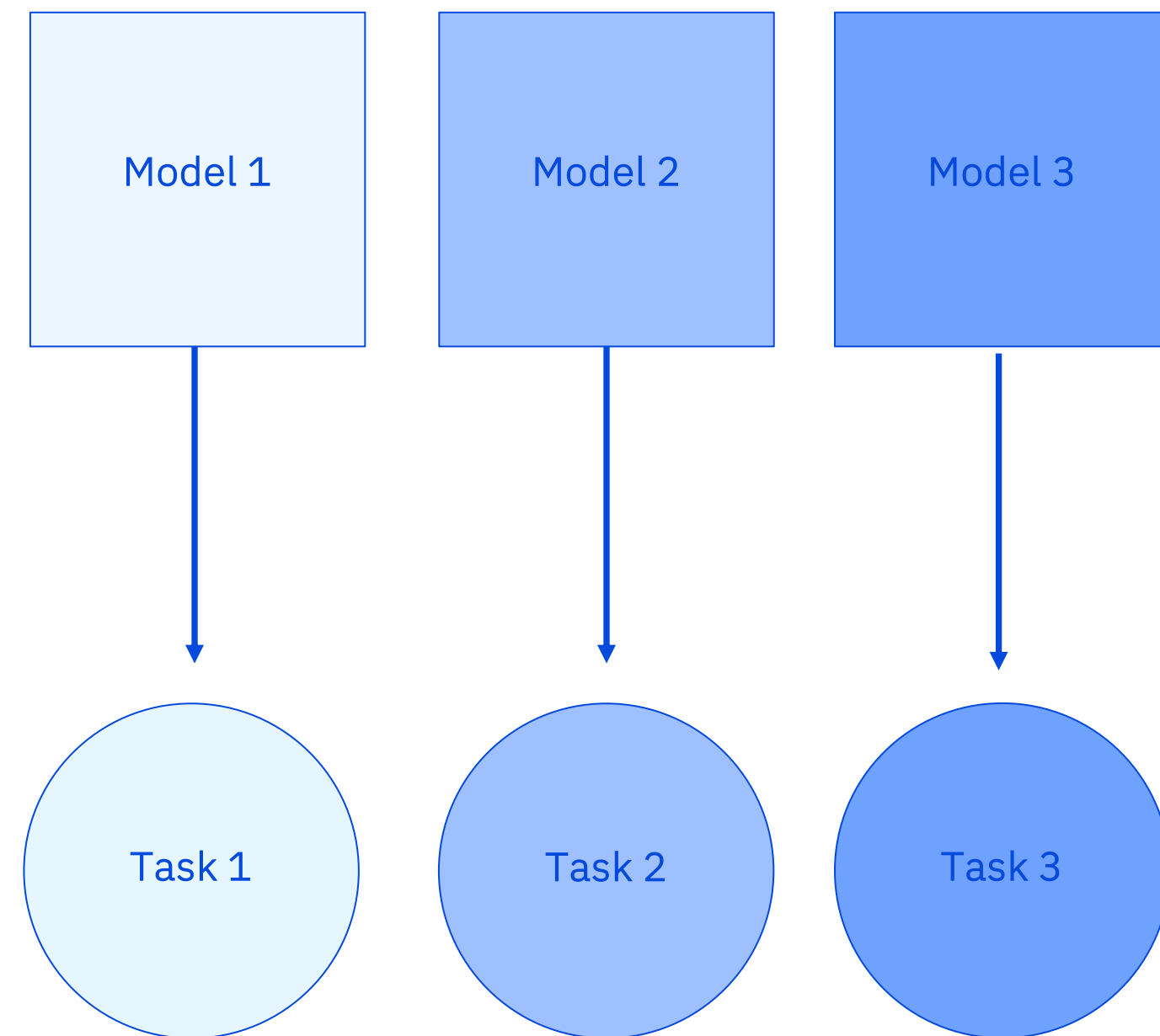
Traditional AI Models

Each model is trained for a
specific task

Model 1	Model 2	Model 3

Task 1	Task 2	Task 3

1,000s to 1,000,000s labeled data
points per task

Traditional AI Models

Each model is trained for a specific task

Model 1	Model 2	Model 3

Task 1 Task 2 Task 3

1,000s to 1,000,000s labeled data points per task

Foundation Models

One model that can address many tasks

Foundation Model
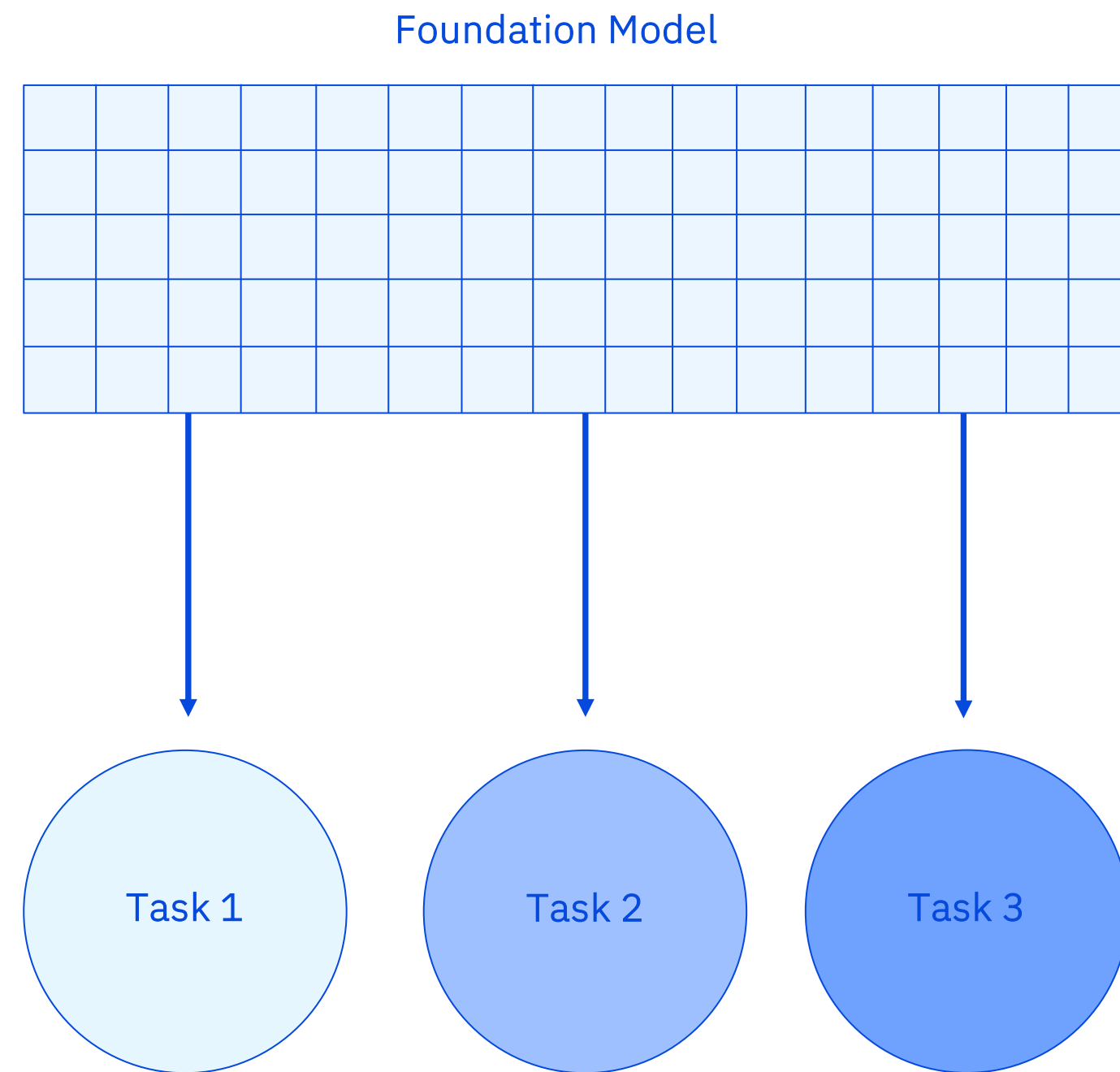
Task 1 Task 2 Task 3

0 to 1,000s labeled data points per task

Foundation Models

One model that can address many tasks

Foundation Model

Task 1   Task 2   Task 3
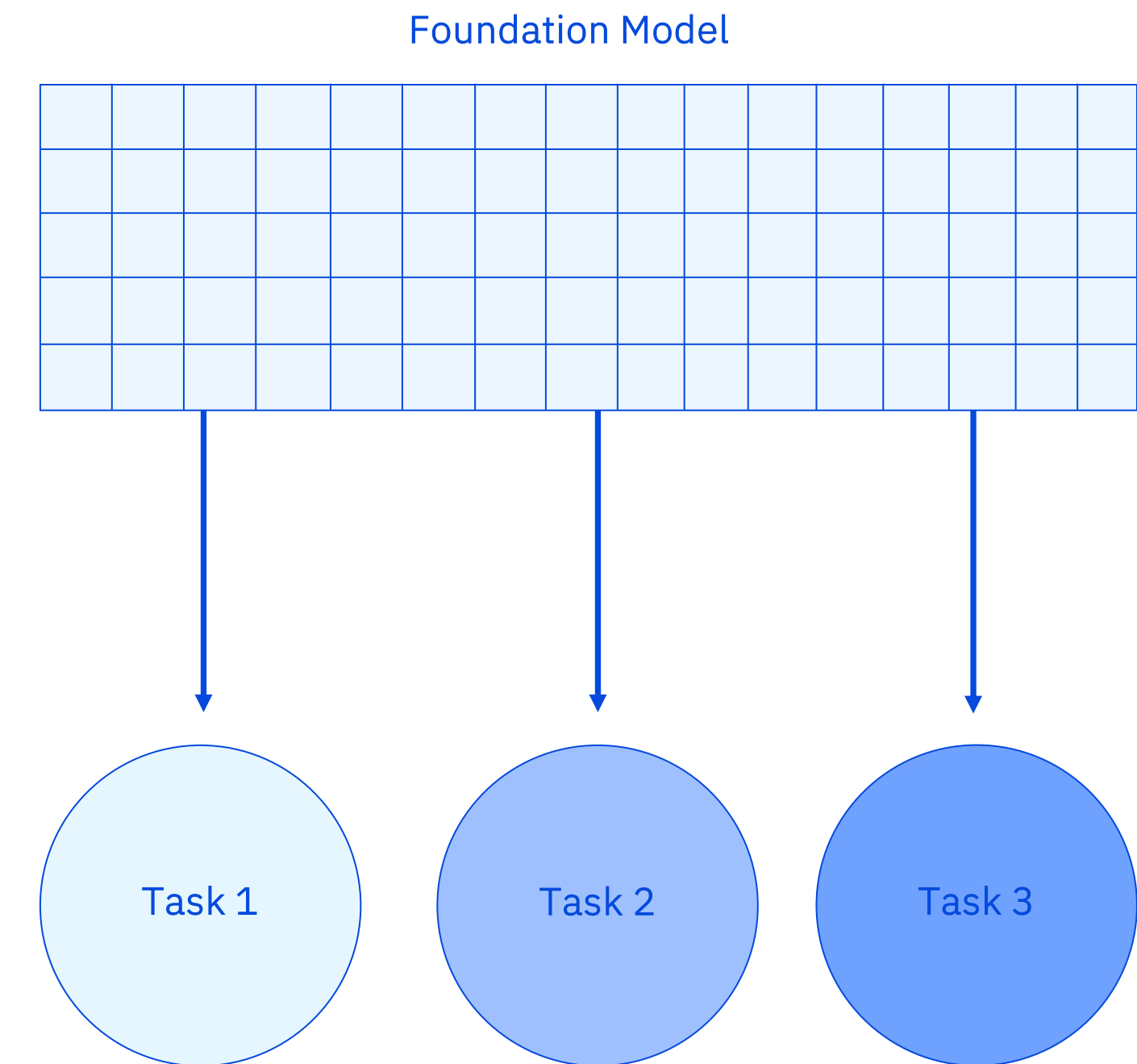
0 to 1,000s labeled data points per task

Prompting Foundation Models
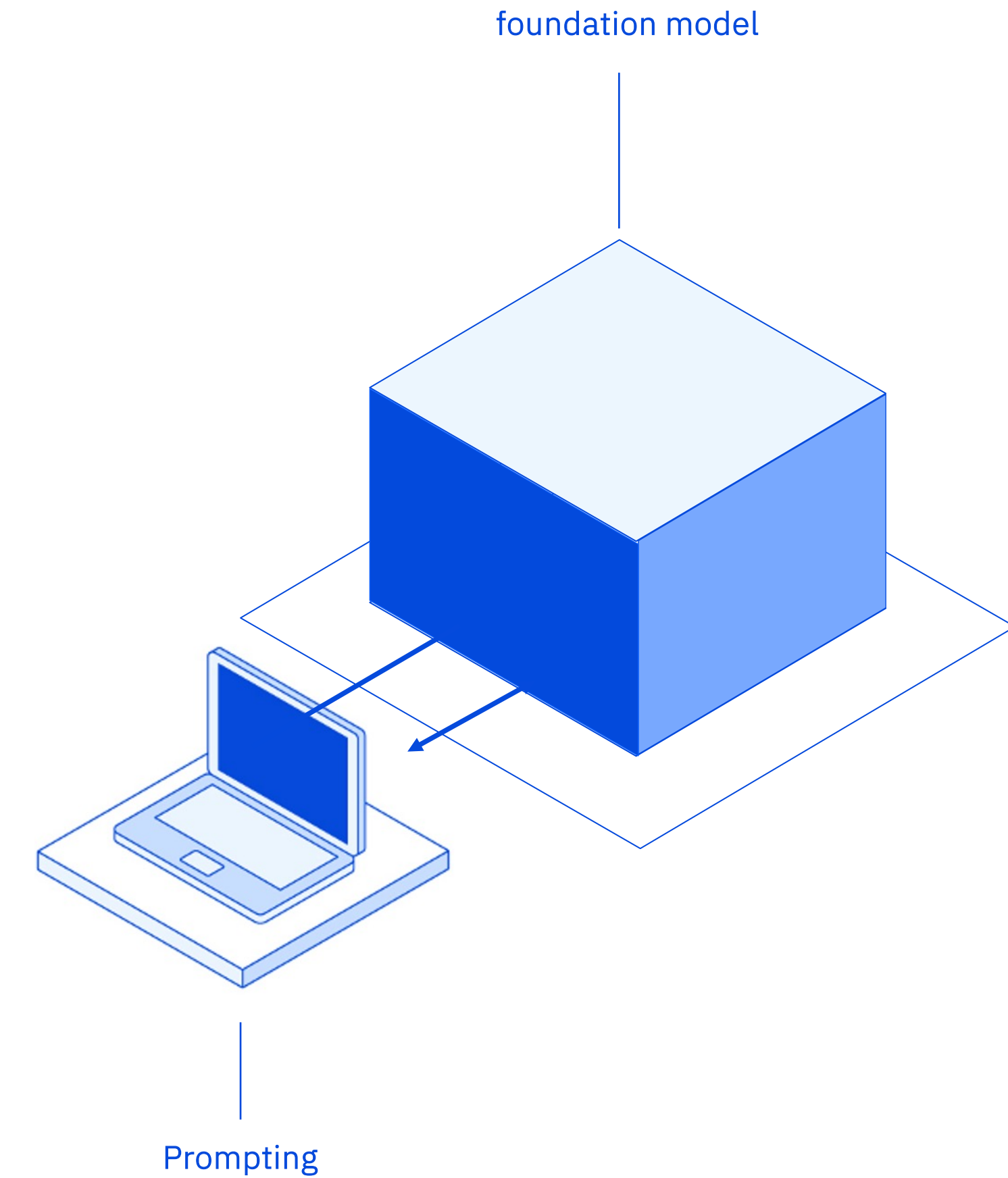
Customized for additional use cases

foundation model

Prompting

Prompting Foundation Models

Customized for additional use cases

foundation model

Prompting

Creating Value with Foundation Models

Fully deployed for enterprise use

watsonx

watsonx.ai

Foundation Models

1 Prompting

2 Tuning

3 Deploying

Creating Value with Foundation Models

Fully deployed for enterprise use

watsonx

watsonx.ai

Foundation Models

1 Prompting

2 Tuning

3 Deploying
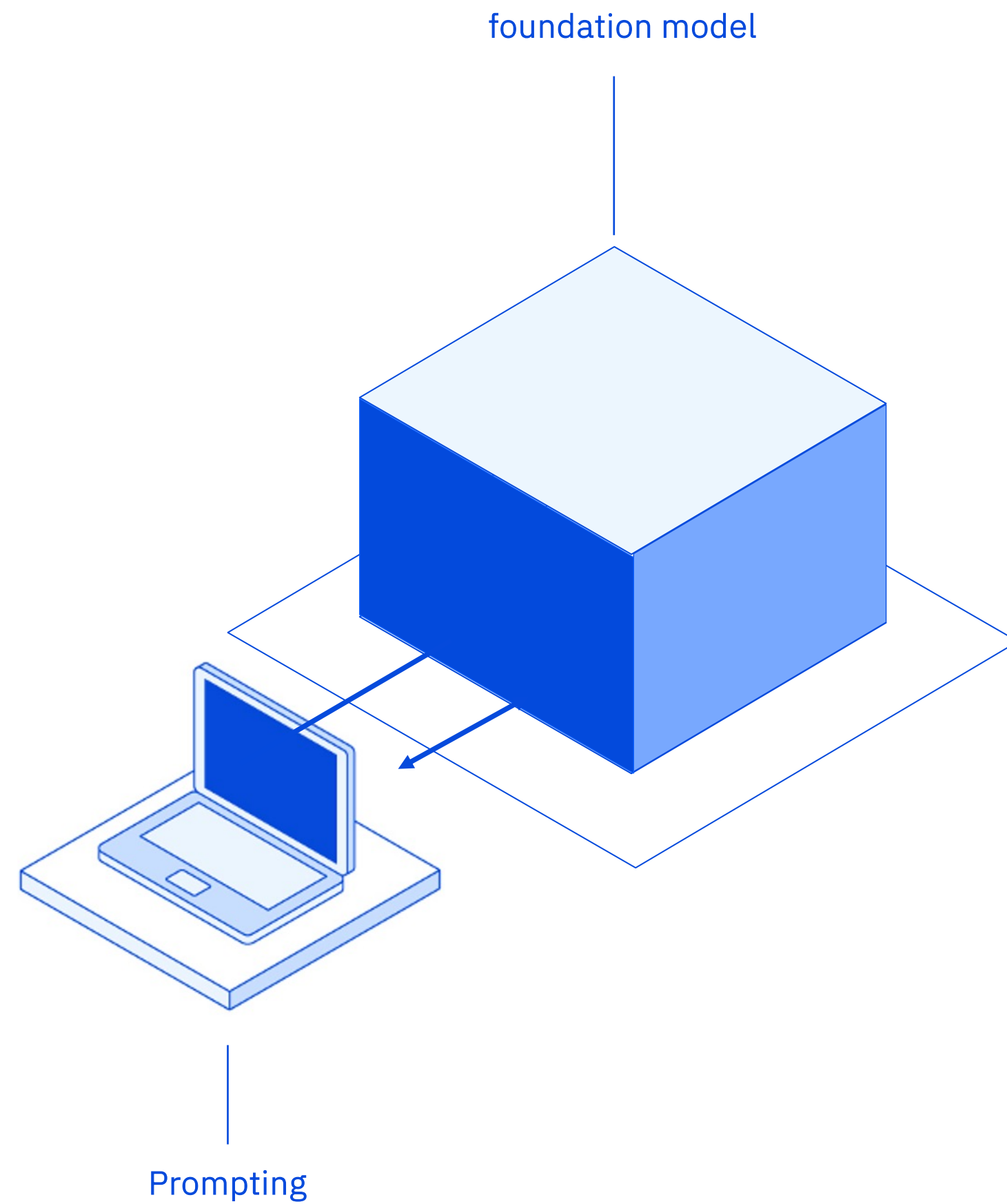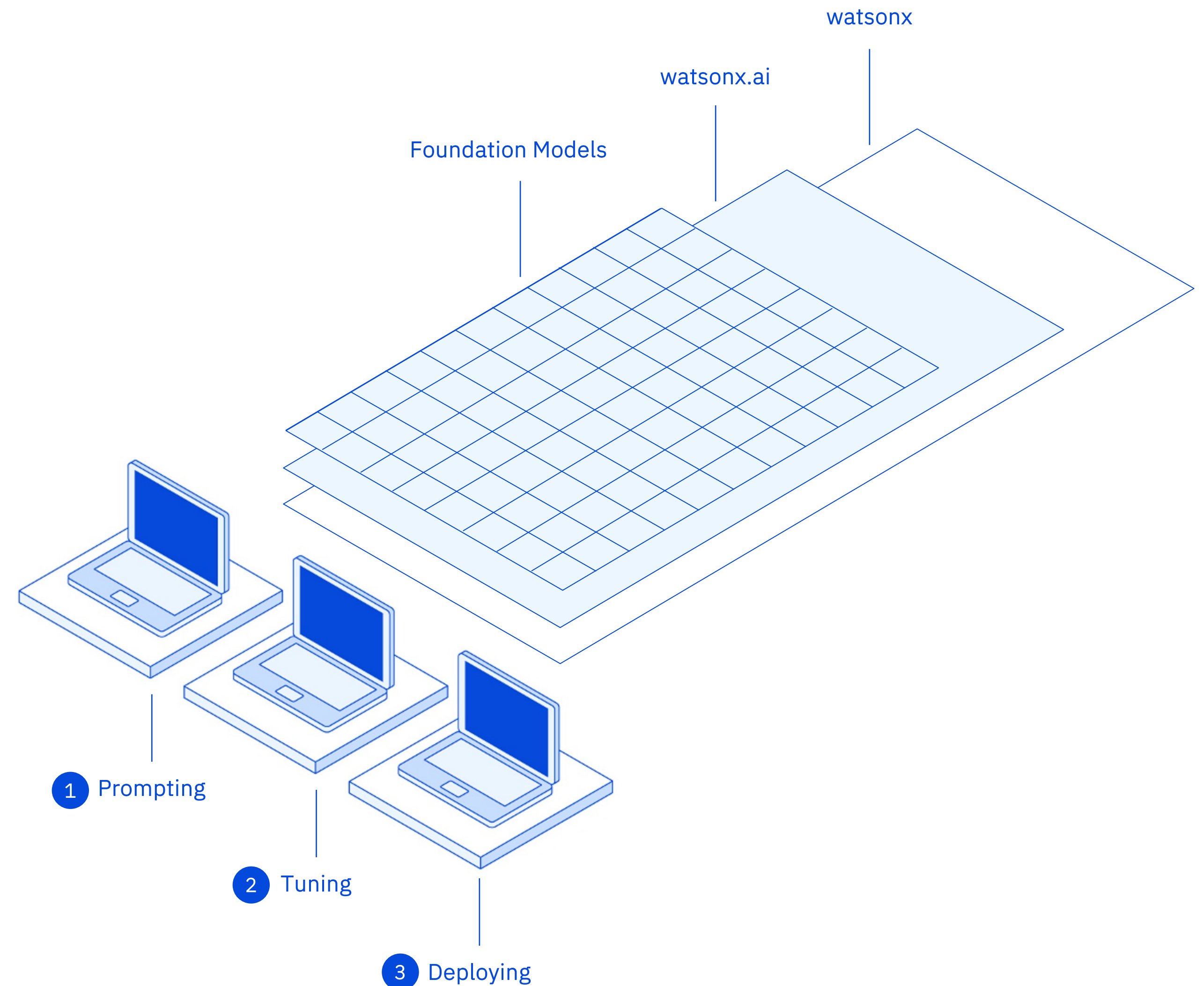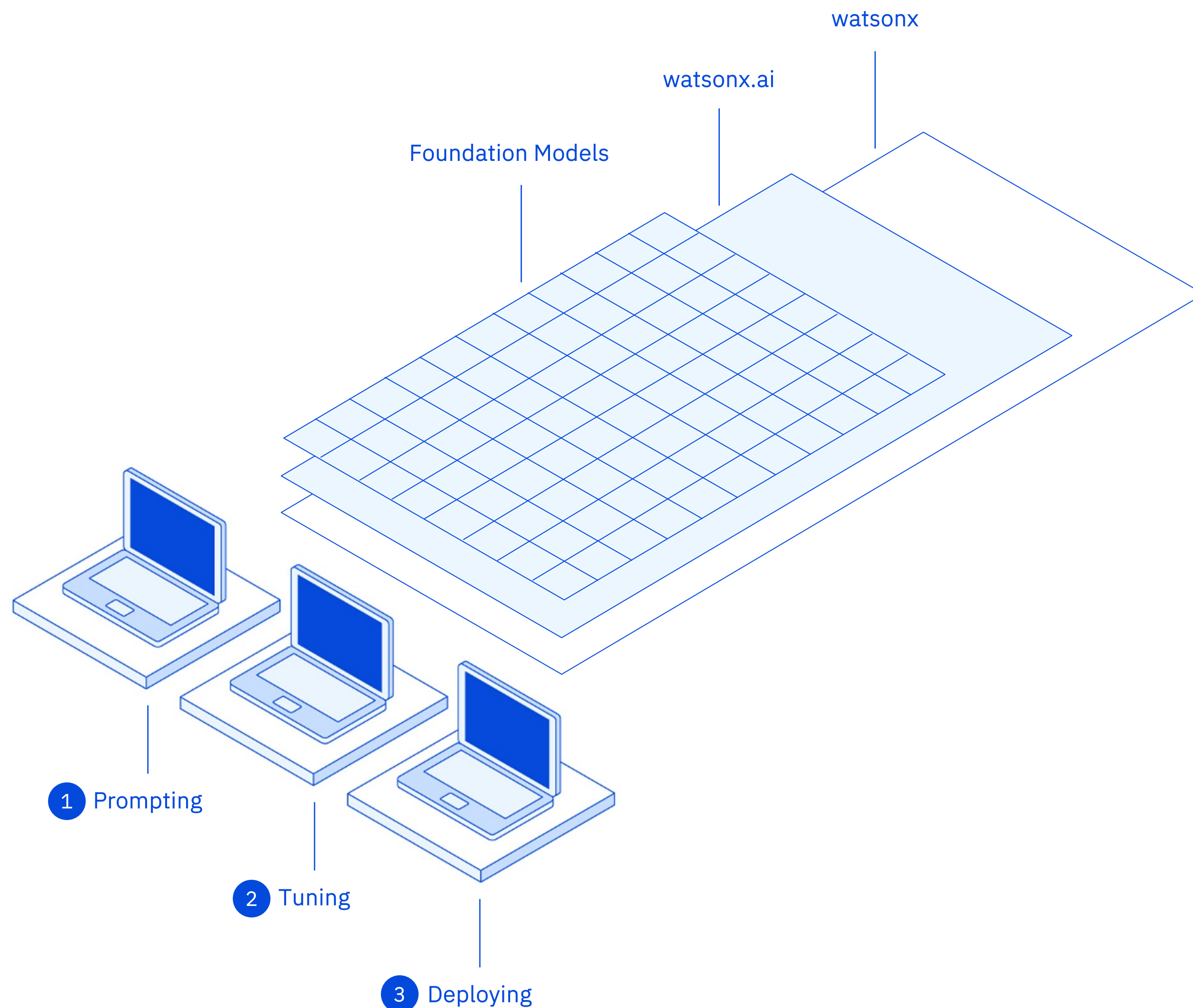
1 Prompt Lab

Leverage foundation models and generative AI with a fraction of the data, in less time, via no-code or full SDK and API libraries.

2 Tuning Studio

Use as few as 100 labeled data points to tune pre-trained models on your own data.

3 Deployments

Accelerate the full AI model lifecycle with all the tools and runtimes in one place.

A Next-Generation Enterprise-Ready AI Studio

Foundation Models | **IBM FMs, open-source FMs, and Bring-Your-Own-Model**

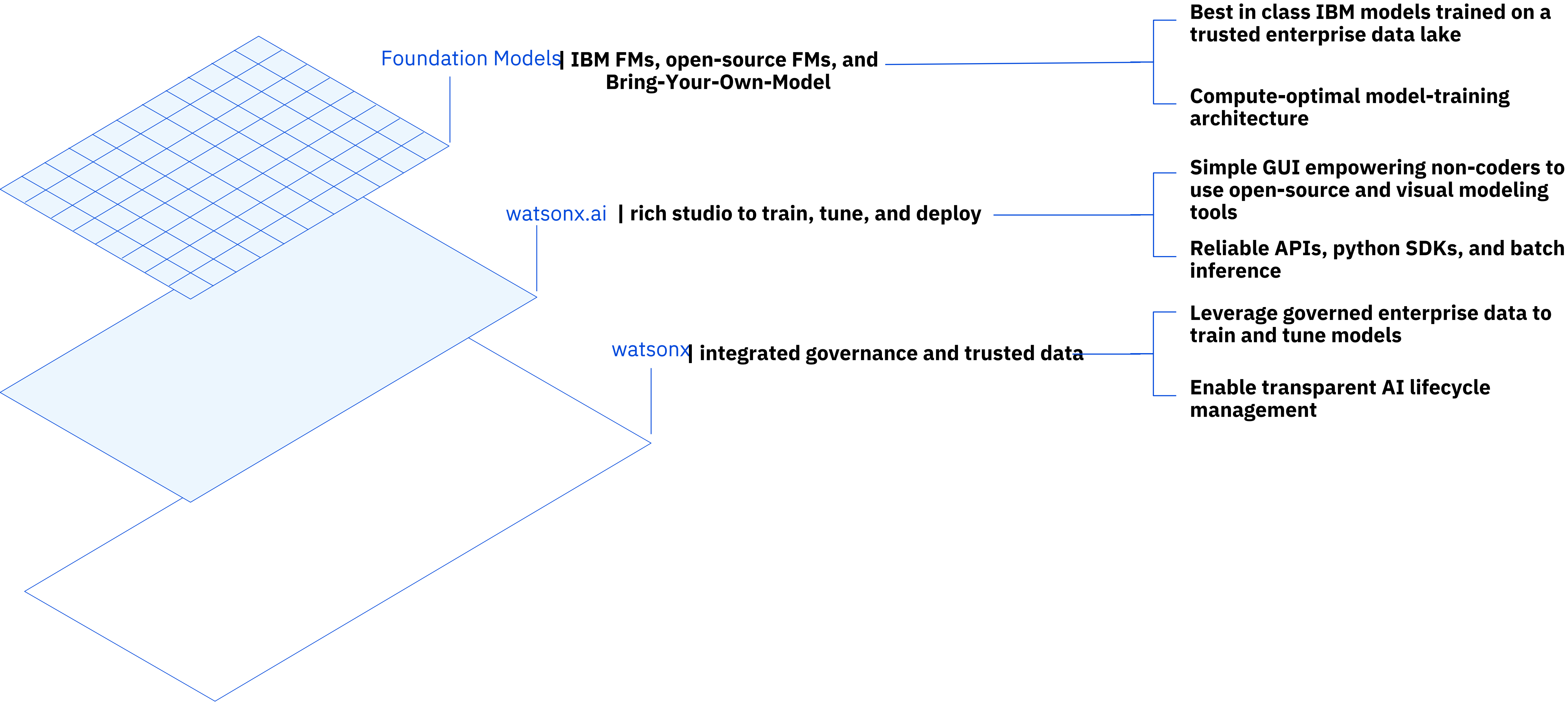**Best in class IBM models trained on a trusted enterprise data lake**

Compute-optimal model-training architecture

watsonx.ai | **rich studio to train, tune, and deploy**

Simple GUI empowering non-coders to use open-source and visual modeling tools

Reliable APIs, python SDKs, and batch inference

watsonx | **integrated governance and trusted data**

Leverage governed enterprise data to train and tune models

Enable transparent AI lifecycle management

Best In-Class Models Built on a Foundation of Trust

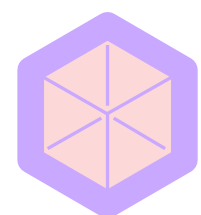**Foundation Models | IBM FMs, open-source FMs, and Bring-Your-Own-Model**

IBM Foundation Models
Models trained on the largest trusted, enterprise grade data lake

Slate
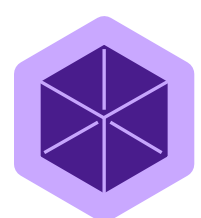Non-generative, encoder-only architecture

Sandstone
Light-weight encoder-decoder architecture

Granite
Generative specialization, decoder-only architecture

Obsidian
Novel-sparse universal transformer architecture

Moonstone
Novel architecture based on dense associated memory

And models for efficient domain and task specialization:

- Language (text, code)
- Finance
- Geospatial
- Chemistry
- Cybersecurity
- Legal
- Climate

Open source Models
Access an open ecosystem through API integrations

- Experiment with open source models
- IBM and Hugging Face partnership demonstrates our shared commitment to delivering an open ecosystem approach that allows you to define the best models for your business needs.
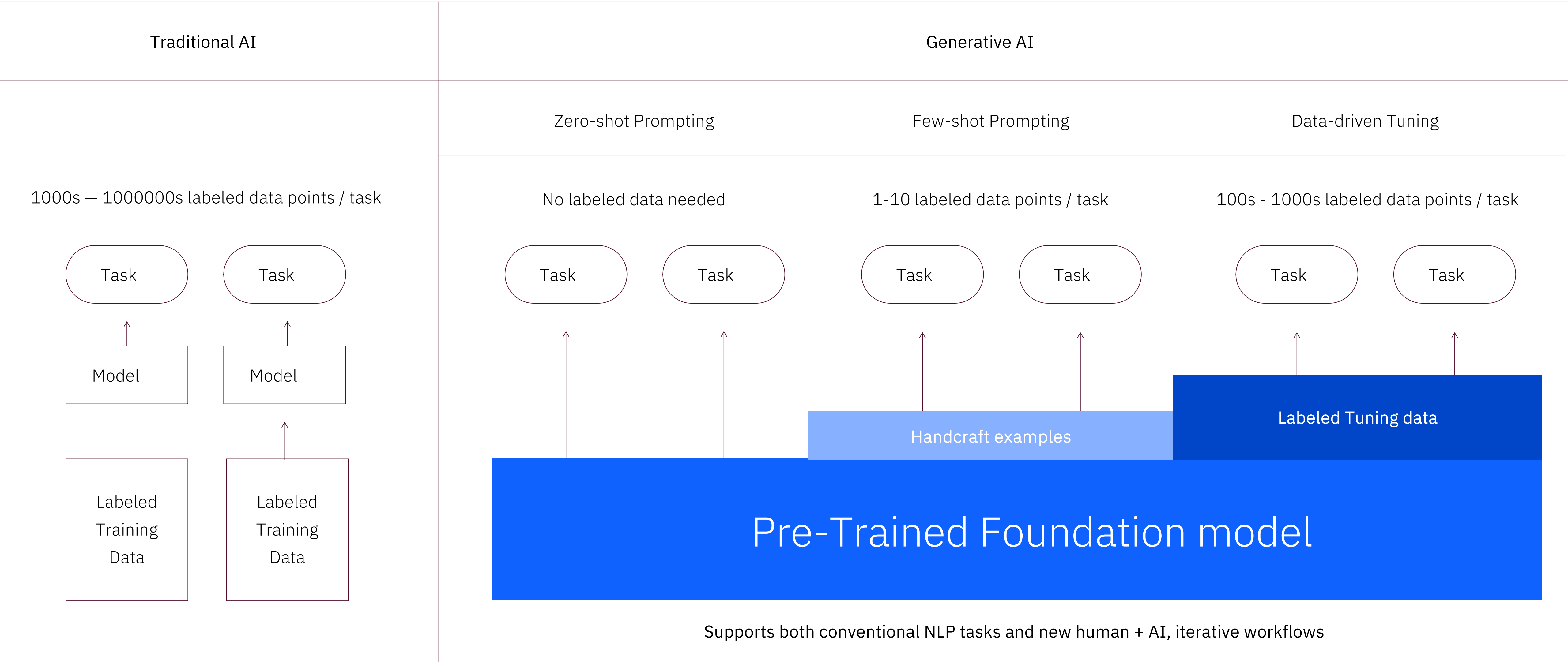
Bring-your-own-model
Optional add-on for more flexibility

- Partner with IBM Research to pre-train your own foundation models

Build for Traditional ML and Gen AI in One Place

Traditional AI	Generative AI		
	Zero-shot Prompting	Few-shot Prompting	Data-driven Tuning
1000s — 1000000s labeled data points / task	No labeled data needed	1-10 labeled data points / task	100s - 1000s labeled data points / task

Supports both conventional NLP tasks and new human + AI, iterative workflows

Search in your workspaces

Upgrade

IBM account

Dallas

MB

Welcome, Kate

Train, validate, tune, and deploy AI models.

Open in: Kate's sandbox

[Customize my journey]

AI
Experiment with foundation models and build prompts
with Prompt Lab β

AI
Tune foundation models with labeled data
with Tuning Studio β

AI
Build machine learning models automatically
with AutoAI

AI
Automate model lifecycle
with Pipelines

Collapse

Jump back in

Recently visited pages

Kate's sandbox /
customer-conversations.csv

Banking conversations /
Summarization prompt

Samples /
Earnings call summary

Banking deployments /
Bank marketing model

Recent work

Projects +

Kate's sandbox   Just now
KB

Banking conversations   1 h ago
KB GB AP +3

Marketing emails   5 h ago
KB GB

Climate Change Impacts on the Global Food Supply   1 wk ago

Deployment spaces +

Banking deployments
1 h ago

Supply chain deployments
5 h ago

Data fact extraction 2023
1 wk ago

Climate Change Impacts on the Global F...
2 wk ago

Fire prediction 2023

Prompt Lab

New (unsaved)

New prompt +   **Save work** ⌄

| Sample prompts ⓘ ‹ | Structured | Freeform | Model: flan-ul2 (20b) ⌄ ‹/› ⚙ |

Summarization

📞 **Earnings call summary**
Summarize financial highlights on an earnings call.

🖥 **Meeting transcript summary**
Summarize the discussion from a meeting transcript.

Classification

📊 **Scenario classification**
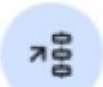Classify scenario based on project categories.

📈 **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 filing.

Question answering

Questions about an article

Set up ⌃

Instruction (optional) ⓘ

> Extract the line of credit facility maximum borrowing capacity from the 10-K filing.

Examples (optional) ⓘ

10-K filing:	Line of credit facility maximum borrowing capacity:
The credit agreement also provides that up to $250 million in commitments may be used for letters of credit.	$250M
In March 2021, we upsized the credit agreement by $50 million, which matures July 2023, to $1.725 billion.	$1.725B

Add example +   🗑

Try ⌃

Test your prompt ⓘ

10-K filing:	Line of credit facility maximum borrowing capacity:
We prepared our impairment test as of December 2022 and determined that the fair values of each of our reporting units exceeded net book value by more than 25%. Among our re . . .	$2 billion ⤒ ⓘ

Time running: 80 out of 40966.98 second

Generate →

IBM **watsonx**

Q Search in your workspaces

Upgrade   ⑦   🔔   IBM account ⌄   Dallas ⌄   MB

Tune foundation models with labeled data

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

◐ Set up

○ Add training data

○ Edit parameters

○ Review and tune

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.

Cancel   Back   Next

Bringing foundation models to enterprise AI

watsonx.ai

Train and validate

Pre-processing	Model training	Validation

Models

Suite of IBM trained foundation models

Tune and infer

Studio	Model serving

Hybrid cloud platform

CodeFlare

RAY	PyTorch
KubeRay	TORCHX

MCAD
Job dispatching, queuing and packing

InstaScale
Cluster Scaling

Scalable optimized distributed training autoscaling from 10's to 1000's of GPUs

Model Mesh

KServe	PyTorch
Hugging Face	ONNX

Caikit Runtime
Dev APIs, Prompt Tuning, Inference

TGIS
Optimized Text Generation Inference Server

Optimized model serving and state-of-the-art support for prompt-tuning

Hybrid cloud-native stack built with open technologies.

Train, tune and deploy foundation models at scale over all kinds of enterprise data.

Built for the enterprise with trust and governance at heart.

Red Hat OpenShift

Use Cases

AI Task	Use case / Demo
Q&A	Watson Assistant and Speech integration for Agent conversation as Amelia alternative
	Cisco Webex Contact Center and Watson Assistant for voice conversation using WA streaming API and handover to Live Agent with key insights about the conversation, e.g. summary, sentiment, entities, classification.
	Q&A with a tech support document repository for agent assist
	Q&A for product queries, recommendation possibly with a WA front end
	Document assistant: Q&A on PDFs and also help to fill forms using content retrieved from documents
	Knowledge management building single source of truth across PLM and EAM resources and enable Q&A
	Procurement advisor: Support, guidance for procurement process.
	Q&A for wealth management, for eg. financial/ earnings reports.
	Q&A on HR data, Digital Labor support for HR
	Operator assist using knowledge augmentation for application deployment to ABB platform
	Customer Assist for technical support during hardware ordering
	Assist for requirements engineering, find existing features for re-use
	Q&A across a large document set of various types in a Box folder
Classify	OCR with Document Understanding tech preview library - need to adapt the demo for invoice based on latest ask from Finacle
	Classification of support tickets, agent conversations, email threads to understand topics of conversations and accordingly determine next step, e.g. route tickets
	Classification of documents to enable tagging documents for quick navigation, search. Need to know labels in advance.
	Classify product feedback on product area, sentiment, tone with data in Airtable
	Classify PDF documents, text categorisation
Summarise	Summarise support tickets, email threads, agent conversations for agent assist
	Summarise tech support documents - need to work on large document summarization
	Summarization of docs in PDF, show in context of legal contracts
	Summarise customer feedback, account plan, performance in sales scenario with data in Airtable
	Summarise documents in a Box folder and also generate document tags, possible using Extract, Classify patterns

Use Cases

AI Task	Use case / Demo
Extract	Entity / PII extraction from support tickets, agent conversations, text documents
	Extract entities/PII from PDF, also show topic, concept, relation extraction
	PII extraction for various format documents in folder and mark documents as confidential and mask information
	Extract insights from logs for preventive maintenance. Show how we can move from rules based to an LLM based approach find log patterns and anomalies.
	Extract insights from logs for threat detection. Show how we can move from rules based to an LLM based approach find log patterns and anomalies.
	Create insights from invoices by extracting content in different formats
	Extract action items from content for sales rep
	Extract entities and create summary from healthcare data
Code generation	Examples of SQL to MongoDB Query Language with prompt engineering in workbench
	Basic code translation example in workbench
	Synthetic data generation to train robots
Content generation	Content generation for marketing emails, creative briefs, blogs, posts, responses etc.
	Generate job description from open roles in HR
	Generate product spec from data in Airtable
	Auto-generate and summarise technical documents / user manuals specific to custom order
Other	Federated SQL query across data sources, e.g. PostgreSQL, COS, using watsonx.data platform.
	Model lifecycle management
	Build predictive models
	Hosting watsonx.ai

watsonx.ai

Build in our new studio for foundation models, generative AI and machine learning. See what's coming.

Join the waitlist →   Watch the video (1:49) ▷