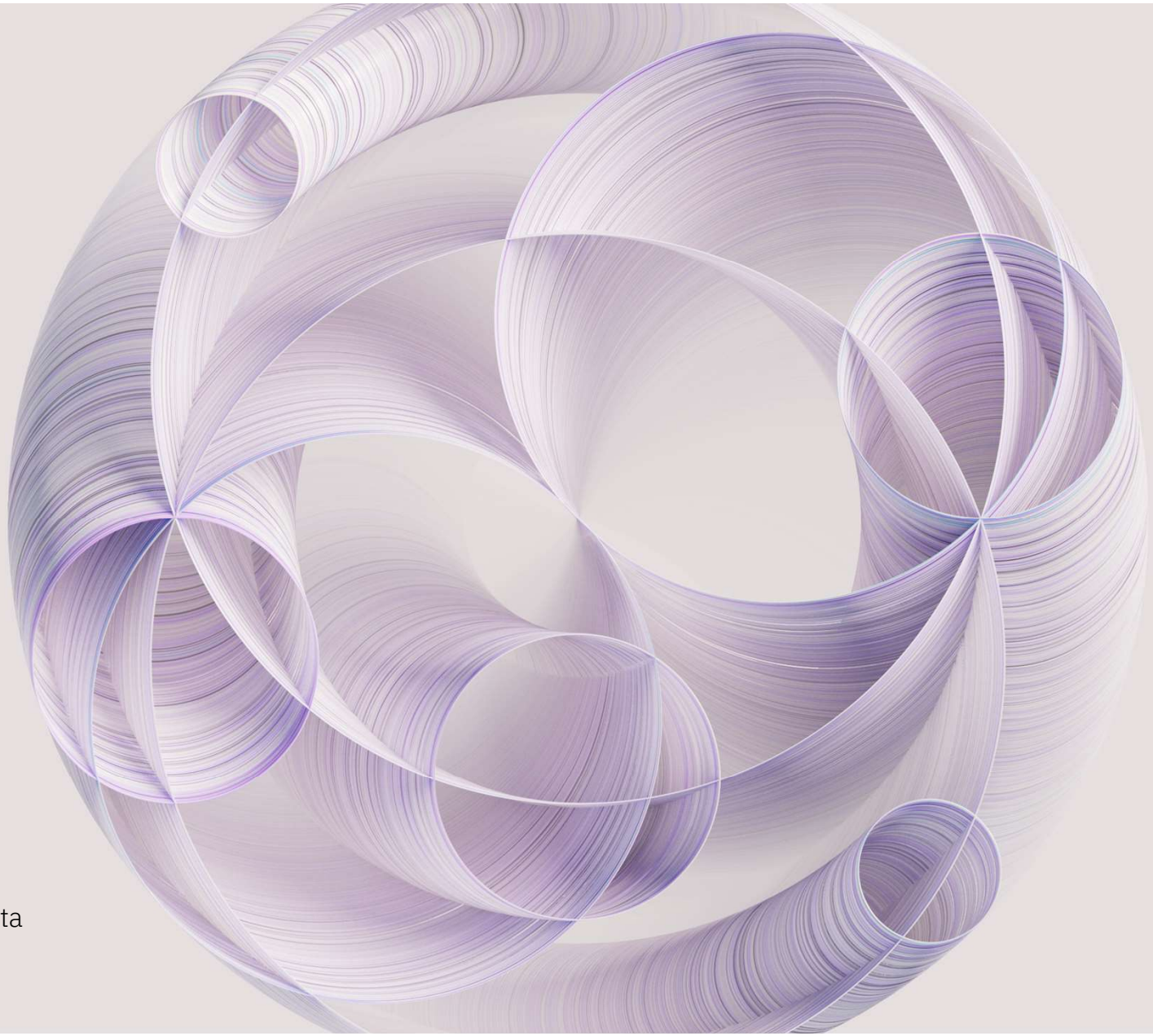
watsonx.ai

Train, validate, tune and deploy AI models



Client presentation

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

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

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

Contents

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

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

Generative AI and traditional AI

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

Generative AI

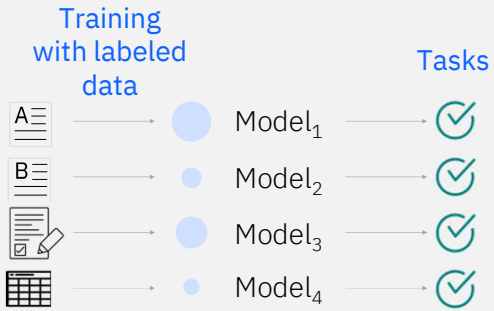
- Foundation models trained with unlabeled data
- Unsupervised
- Trained on very big data sets
- No specific task
- Transferable
- Works well for general tasks and can improve for specific tasks with less training
- Need to monitor bias and drift

Traditional AI

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

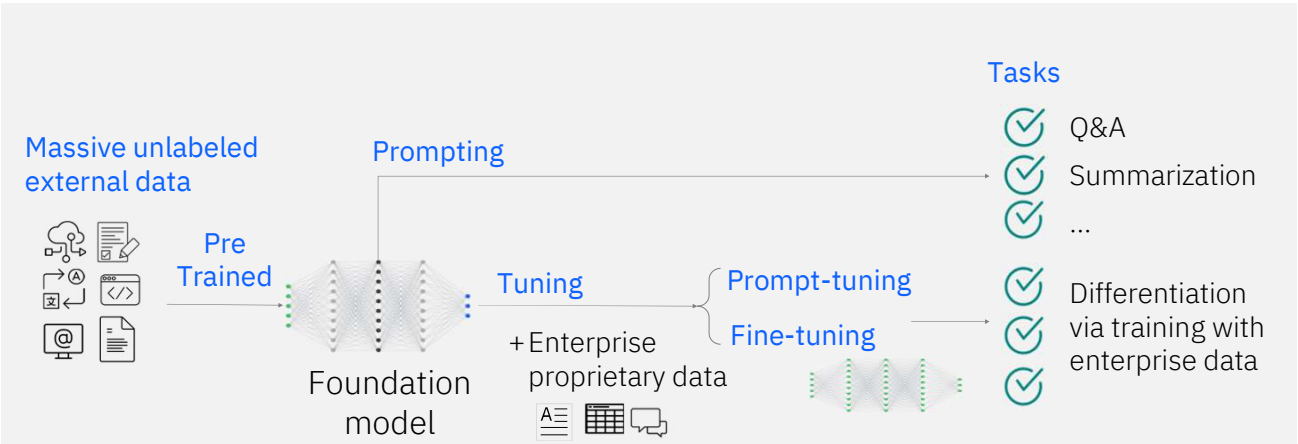
Foundational models enable a new paradigm of data-efficient AI development – generative AI

Traditional AI models



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

Foundation Models



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

Impact of generative AI

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

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

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

Most common generative AI tasks implemented today

Summarization

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

Conversation summaries, insurance coverage, meeting transcripts, contract information

Classification

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

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

Generation

Generate text content for a specific purpose.

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

Extraction

Analyze and extract essential information from unstructured text.

Medical diagnosis support, user research findings

Question-answering

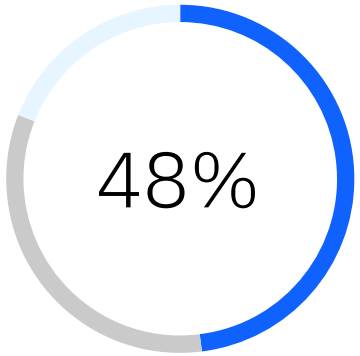
Create a question-answering feature grounded on specific content.

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

Generative AI adoption considerations, inhibitors and fears

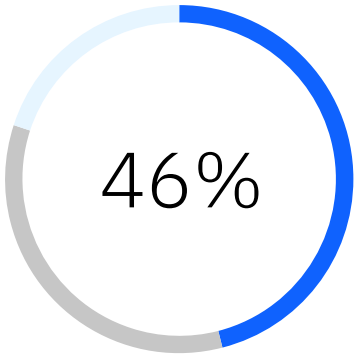
80% of business leaders see at least one of these ethical issues as a major concern

Explainability



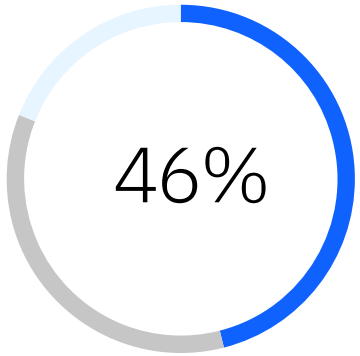
Believe decisions made by generative AI are not sufficiently **explainable**.

Ethics



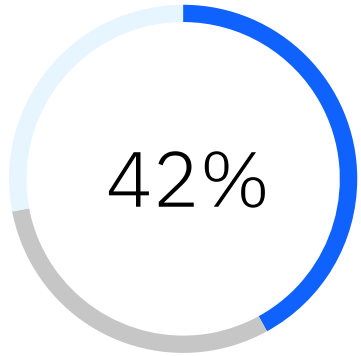
Concerned about the safety and **ethical** aspects of generative AI.

Bias



Believe that generative AI will propagate established **biases**.

Trust



Believe generative AI cannot be **trusted**.

Agree Neutral Disagree

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

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

Generative AI must be tailored to the enterprise

Open

Based on the best open technologies available.

Access to the innovation of the open community and multiple models.

Trusted

Offering security and data protection.

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

Targeted

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

Models that can be tuned to your proprietary data.

Empowering

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

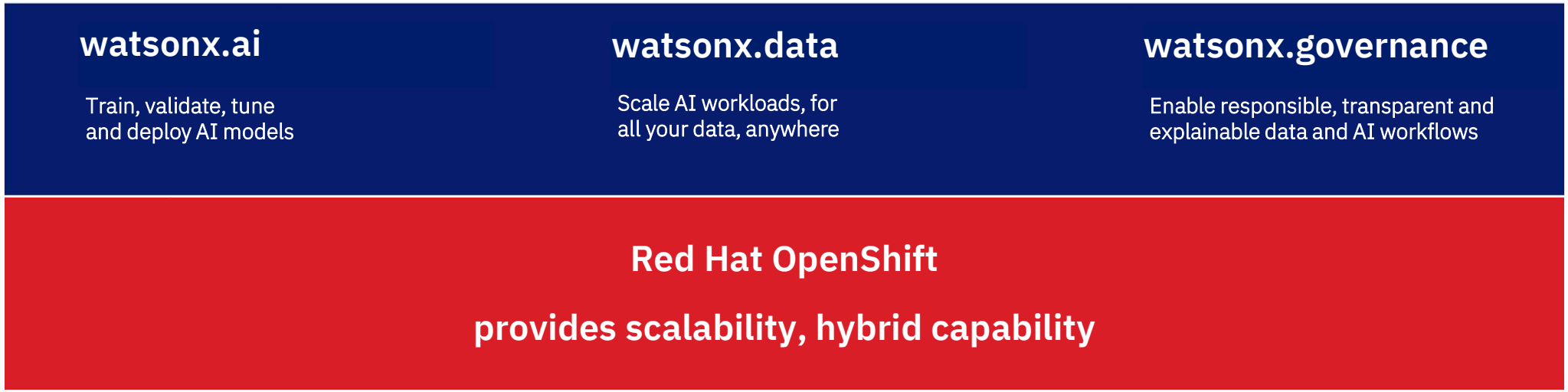
Running anywhere, designed for scale and widespread adoption.

Introducing...

watsonx.ai

Put AI to work with **watsonx**

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



watsonx

and its 3 components

The platform for AI and data

Scale and accelerate the impact of AI with trusted data.

watsonx.ai

Train, validate, tune and deploy AI models

A next generation enterprise studio for AI builders to train, validate, tune, and deploy both traditional machine learning and new generative AI capabilities powered by foundation models. It enables you to build AI applications in a fraction of the time with a fraction of the data.

watsonx.data

Scale AI workloads, for all your data, anywhere

Fit-for-purpose data store, built on an open lakehouse architecture, supported by querying, governance and open data formats to access and share data.

watsonx.governance

Enable responsible, transparent and explainable AI workflows

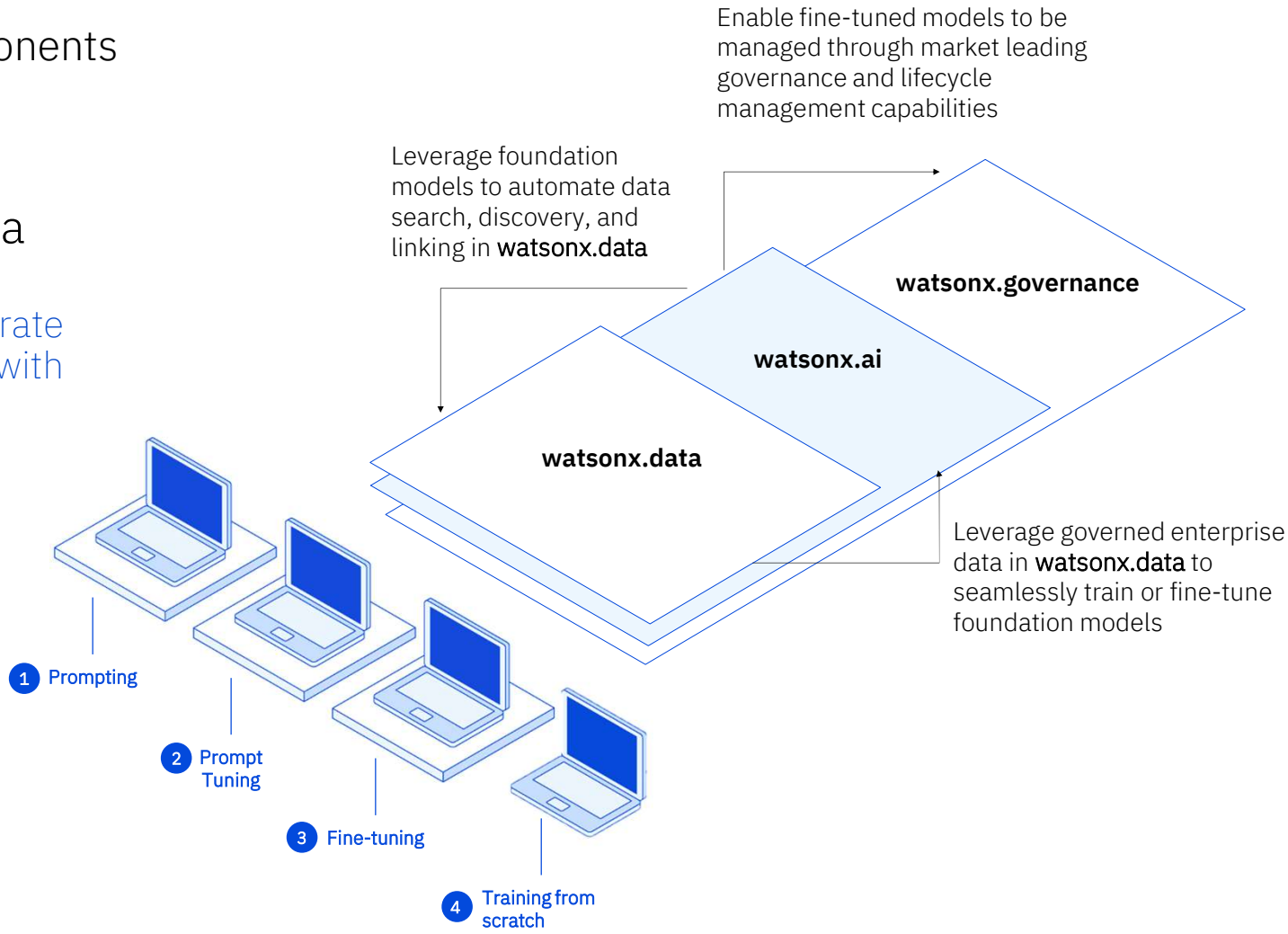
End-to-end toolkit encompassing both data and AI governance to enable responsible, transparent, and explainable AI workflows.

watsonx

and its 3 components

The platform for AI and data

Scale and accelerate the impact of AI with trusted data.



watsonx.ai

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

Bring together AI builders

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

Accelerate the full AI model lifecycle

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

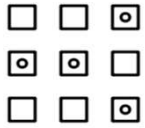
Leverage foundation models & generative AI

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

watsonx.ai – generative AI with traditional AI features

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

Generative AI capabilities

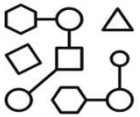


Foundation model library

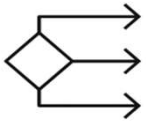
Prompt lab

Tuning studio*

Plus, a proven studio for machine learning

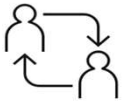


ModelOps



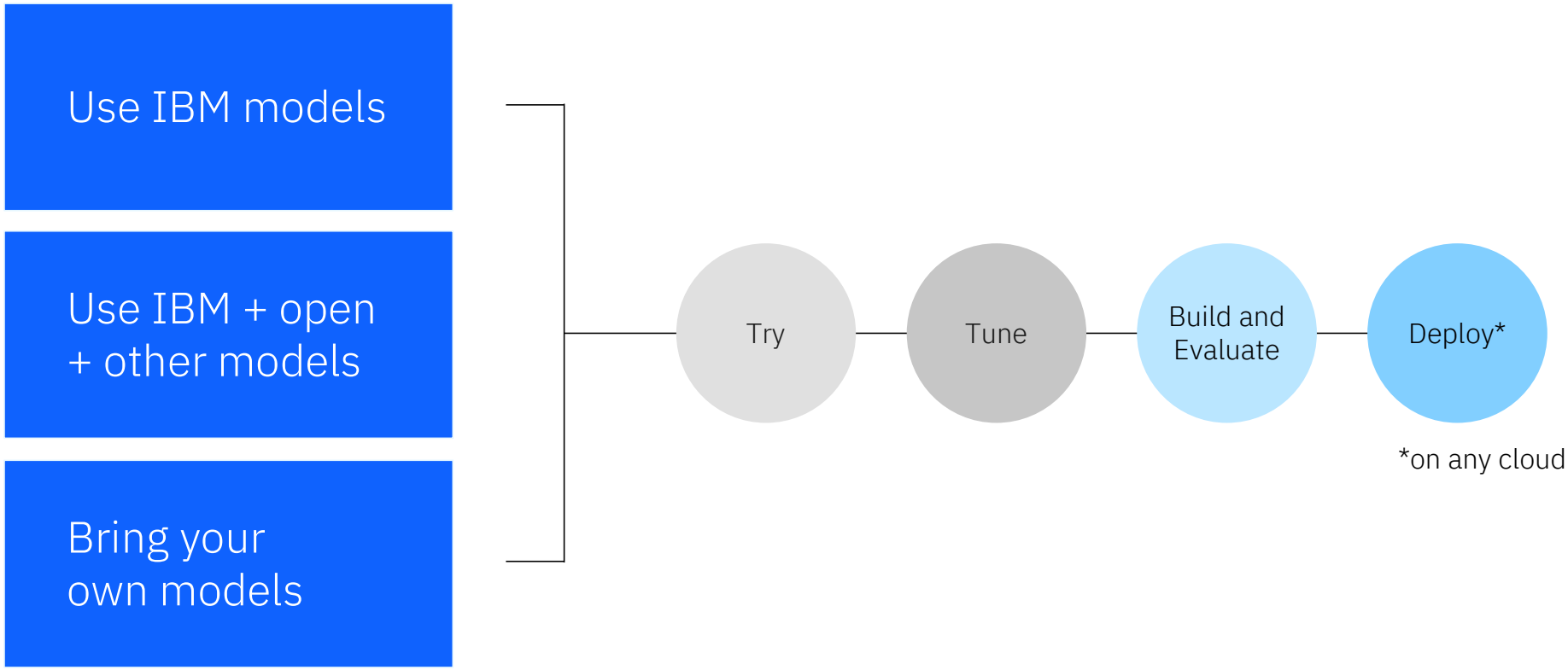
Automated development

Decision optimization



Team collaboration and data preparation

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



watsonx.ai 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 Foundation Model Library

Model variety to cover enterprise use cases and compliance requirements

IBM Foundation Models

Slate (encoder-only) NLP models
Granite (decoder-only)

Slate
multilingual distilled 153 million
Granite
trained on 13 billion parameters

Fine Tuning *Required* to support:

- Extract
- Classify

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

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	gpt-neox	mt0-xxl	flan-t5-xxl	mpt-instruct2
20 billion	*20 billion*	*13 billion*	*11 billion*	*7 billion*
encoder/decoder	decoder only	encoder/decoder	encoder/decoder	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		

Open-source models are sourced from Hugging Face

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

AI for business - IBM Granite (Decoder-only)

These are multi-size foundation models built by IBM that apply generative AI to both language and code.

These foundational models have been trained on enterprise-relevant datasets across five domains:


Internet


Academic


Code


Legal


Finance

These models are grounded in principles of transparency & responsibility...

- IBM provides the list of data sources used to train the model
- Pipeline data is rigorously cleaned for business use
- The same IP protections for IBM software are applied to this LLM

At 13 billion parameter models the Granite models are more efficient than larger models, fitting onto a single GPU.

These models can be used for...

- Text generation Summarization (condense long-form content)
- Insight extraction & classification (determinate sentiment)
- RAG (example: HR chatbot inquiry for maternity leave)

watsonx.ai: Prompt Lab

Experiment with foundation models and build prompts

Interactive prompt builder

Includes prompt examples for various use cases and tasks

Experiment with different prompts, save and reuse older prompts, use different models and vary different parameters

Experiment with zero-shot, one-shot, or few-shot prompting to get the best results

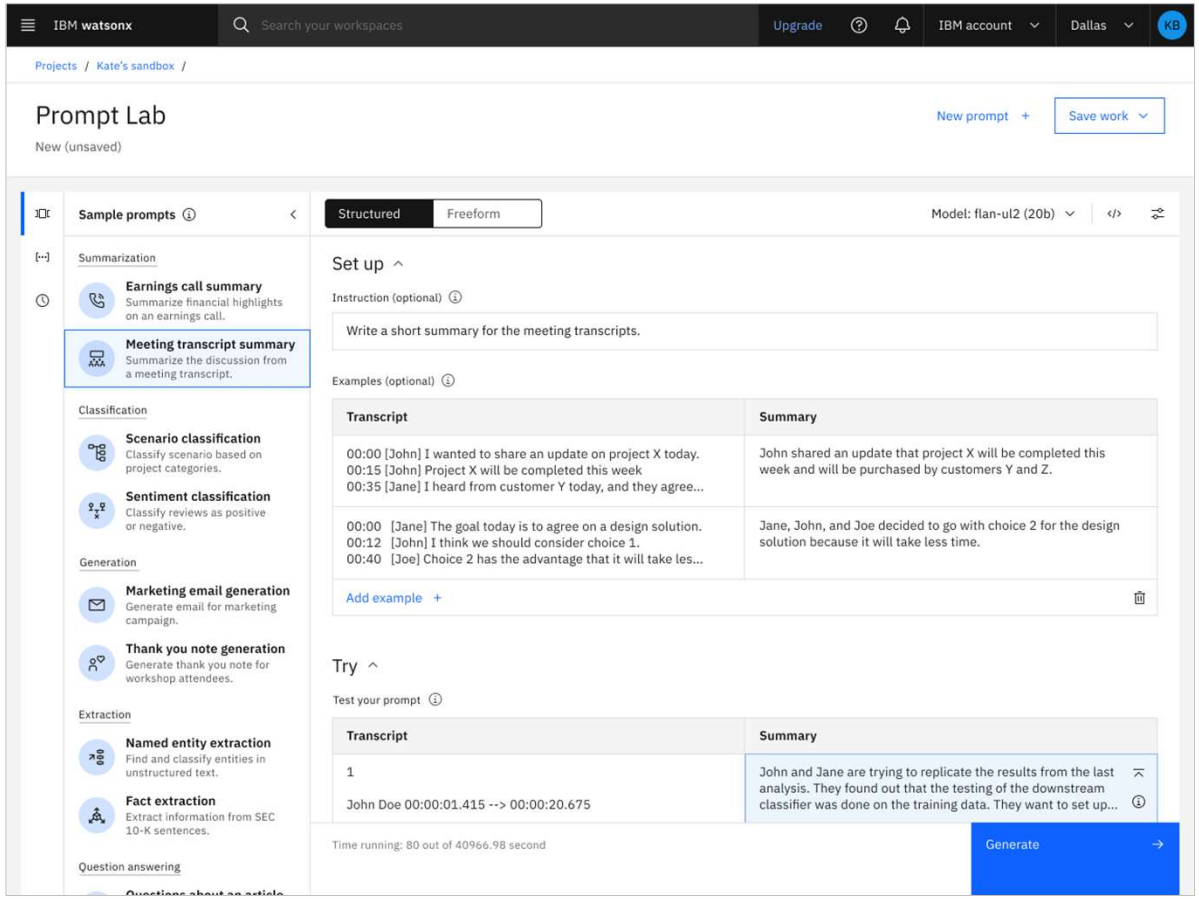
Experiment with prompt engineering

Choice of foundation models to use based on task requirements

Prevent the model from generating repeating phrases

Number of min and max new tokens in the response

Stop sequences – specifies sequences whose appearances should stop the model



watsonx.ai: Data Science and MLOps

Build machine learning models automatically in the studio

Model training and development

Build experiments quickly and enhance training by optimizing pipelines and identifying the right combination of data

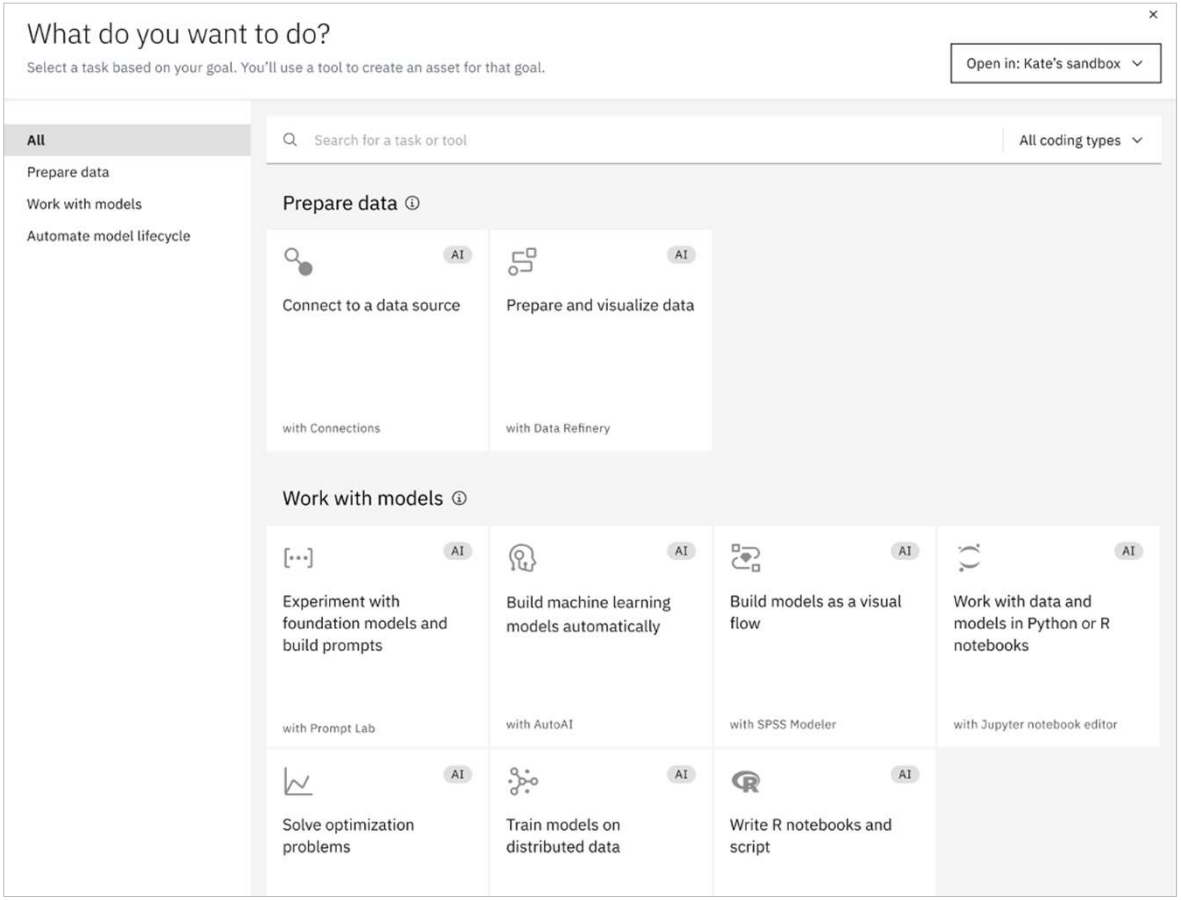
AutoAI, including preparing data for machine learning and generating and ranking candidate model pipelines

Use predictions to optimize decisions, create and edit models in Python, in OPL or with natural language

Integrated visual modeling

Prepare data quickly and develop models visually to help visualize and analyze enterprise data to identify patterns and trends, explore opportunities, and make informed, insightful business decisions

- Uncover correlations
- Insight for hypotheses
- Find relationships and connections within the data



watsonx.ai: Tuning Studio*

Tune your foundation models with labeled data

Prompt tuning

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

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

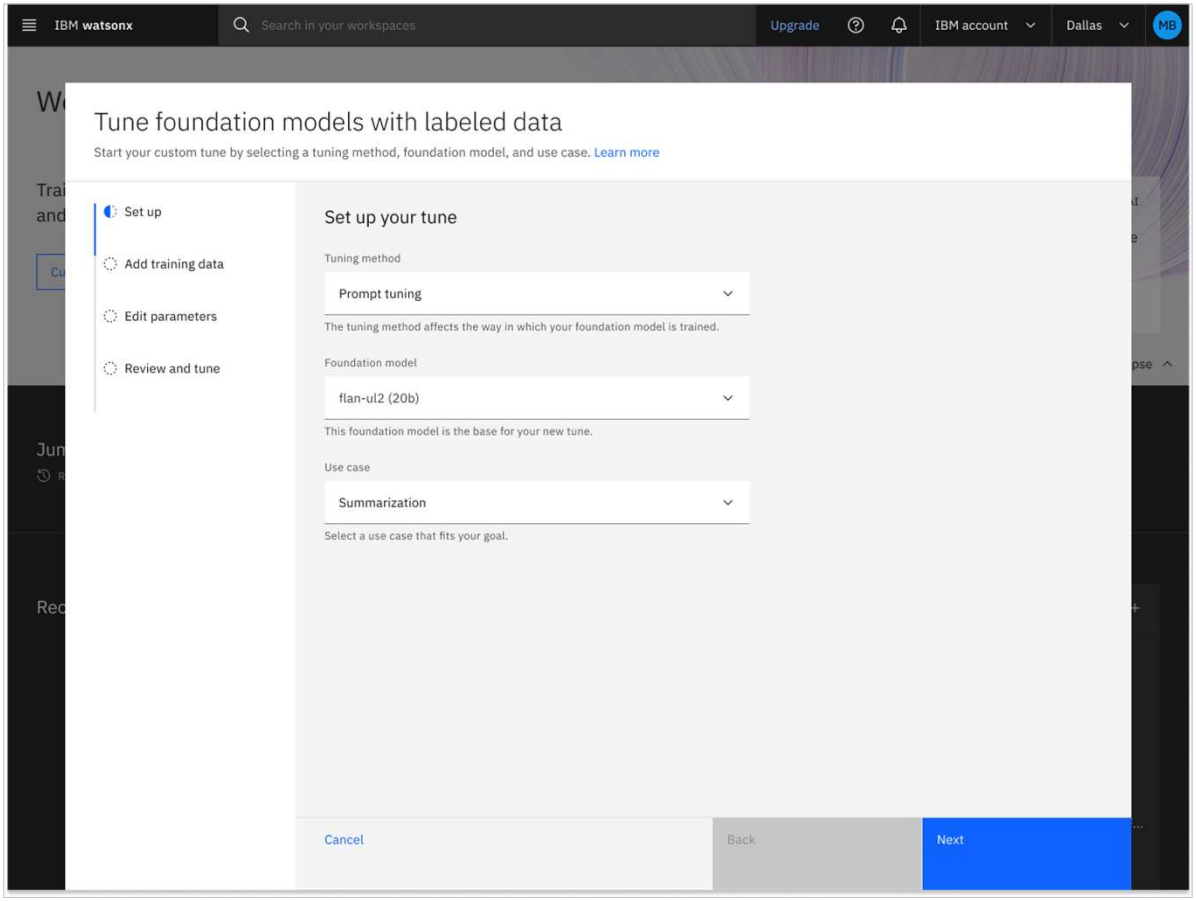
Unlike prompt engineering, prompt tuning allows clients to further enhance the model with focused, business data

Task support in the Tuning Studio

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

Requires a small set of labelled data to perform specialized tasks

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



*Coming soon, available post-GA

watsonx.ai is transparent, responsible, and governed

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

watsonx.data

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

watsonx.governance

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

IBM's Center of Excellence for Generative AI

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

watsonx.ai differentiators

Open

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

Trusted

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

Targeted

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

Empowering

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

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

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.

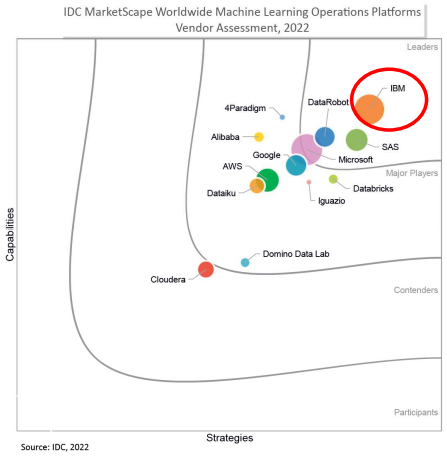
SAMSUNG SDS

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.

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

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

IBM is a leader in AI



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

MQ for Cloud
AI Developer Service

MQ for Enterprise
Conversational AI Platforms

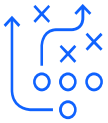
MQ for Insight Engines

Multiple Gartner Magic Quadrants
for AI-related capabilities

Forrester Wave:
Multimodal Predictive
Analytics and
Machine Learning

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

IBM’s investment in partnering with you



FREE TRIAL

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

Try our free trial

CLIENT BRIEFING

Discussion and custom demonstration of IBM’s generative AI **watsonx** point-of-view and capabilities. Understand where generative AI can be leveraged now for impact in your business.

2-4 hours

PILOT PROGRAM

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

1-4 weeks

Backup

Supervised and Self Supervised Learning ↷
What's the difference?

Supervised learning

Human powered

Requires intense labeling

Long, hard, expensive

Self-supervised learning

Computer powered

Requires little labeling

Quick, automated, and efficient

Leveraging foundation model capabilities across various domains

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