watsonx.ai

Train, validate, tune and deploy AI models

Client presentation

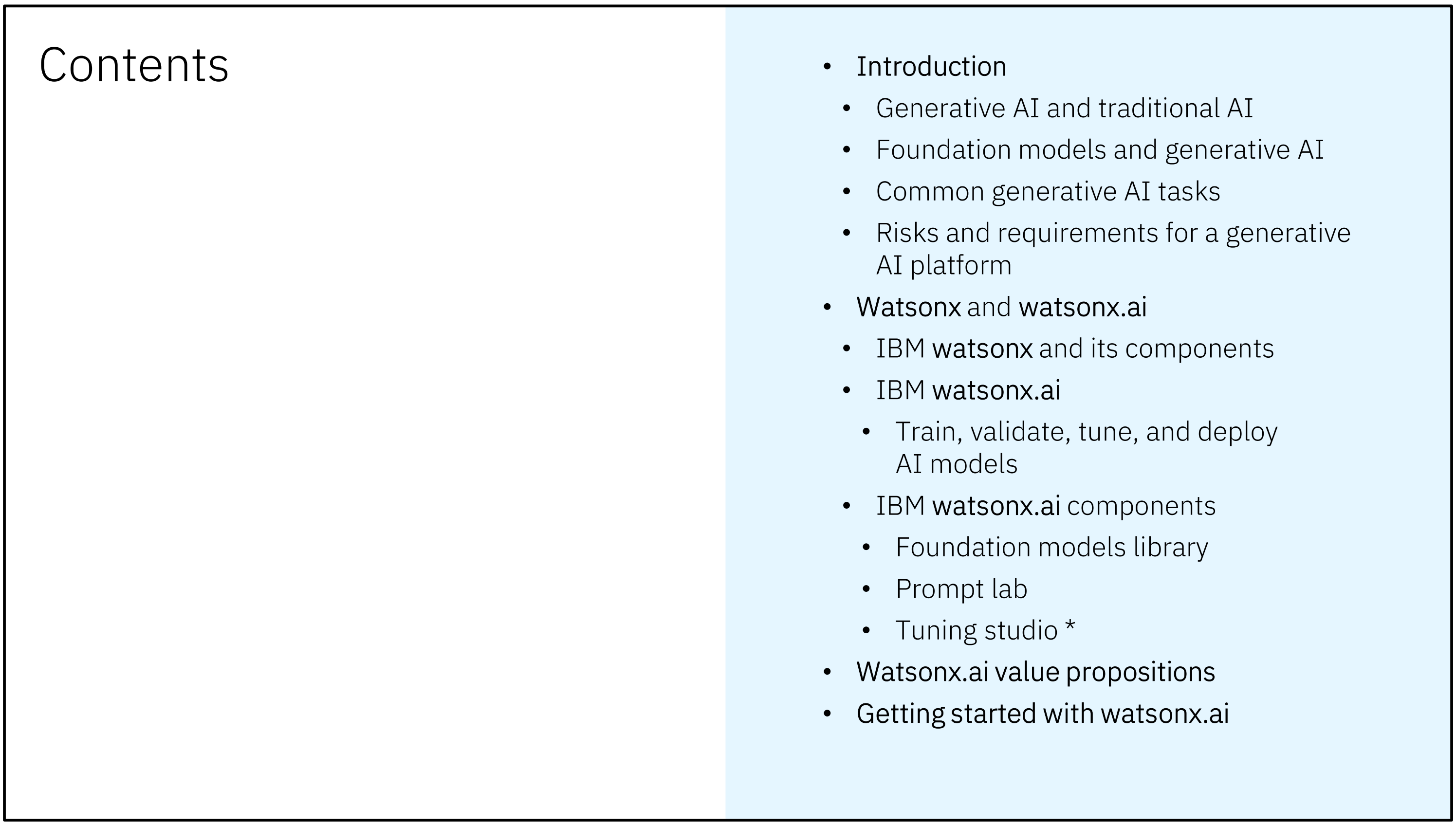
Linsay Wershaw
Lindsay.Beth.Wershaw@ibm.com
Senior Product Marketing Manager, IBM watsonx.ai

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

Watsonx is a new artificial intelligence (AI) and data platform from IBM that is designed with the three critical elements of an AI strategy in mind: AI, data, and governance. It empowers enterprises to train, validate, tune, and deploy AI across their business, leveraging critical, trusted data wherever it resides. This platform has three components that map to the three critical elements of AI platform needs: watsonx.ai, watsonx.data, and watsonx.governance

The focus of this presentation is on watsonx.ai, a studio that clients can use to train, validate, tune, and deploy both machine learning (ML) and traditional AI models, as well as foundation models (FMs) for generative AI. These models combine best-of-breed architectures with a rigorous focus on data acquisition, provenance, and quality, to serve enterprise needs.

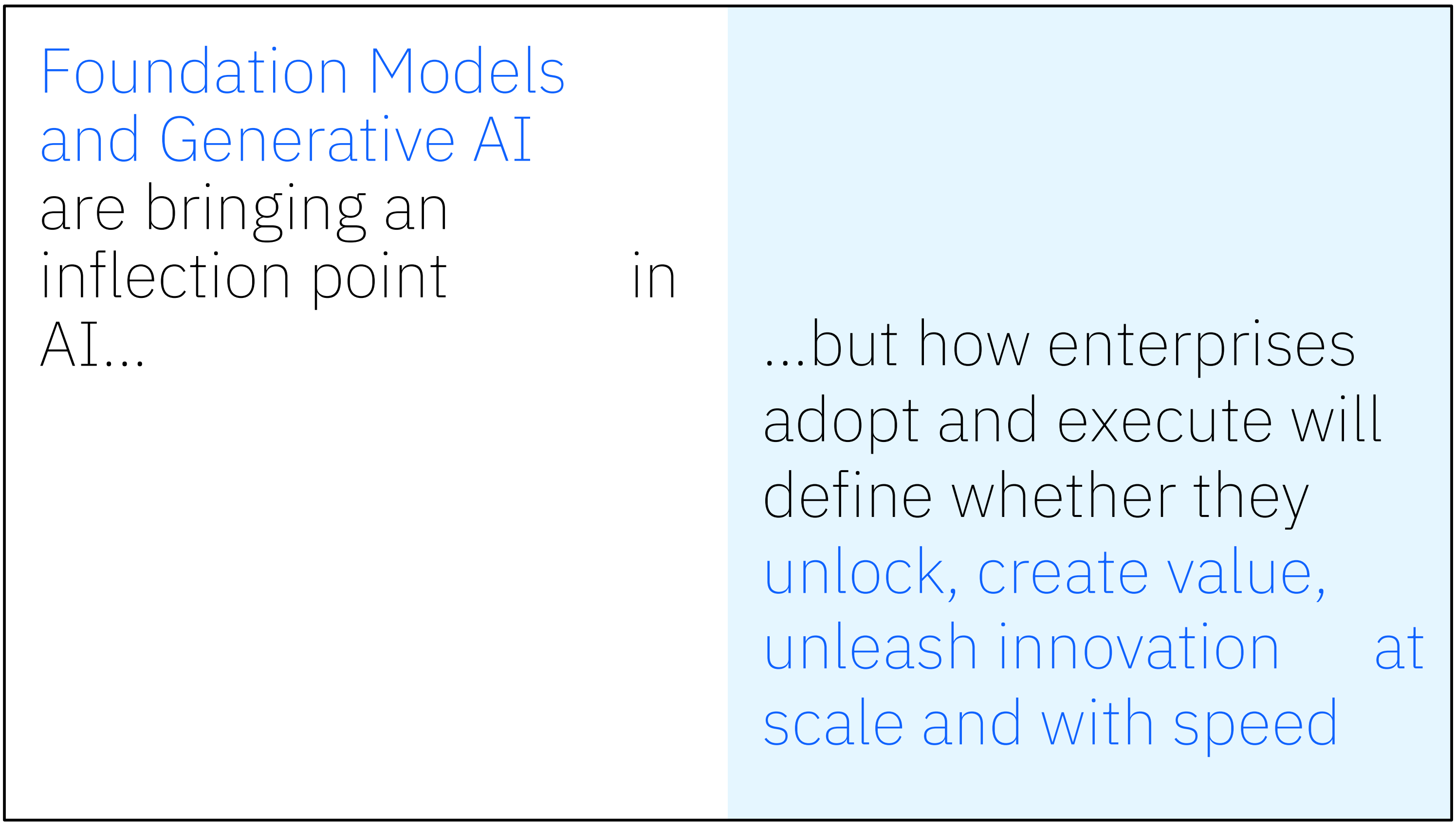


Contents

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

This self-explanatory slide provides an agenda for this presentation.

The Tuning studio is coming post-July 2023.



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

Foundation models (FMs) are different from traditional AI models. While traditional AI models require supervised training on labeled data, FMs use self-supervised training on a large pool of unlabeled data and do not require labeled data. Supervised learning is human-powered, requires intense labeling, and is log, hard, and expensive. Self-supervised leaning is computer-powered, requires little labeling, and is quick, automated, and efficient. (See backup slide for more details on the difference between supervised and self-supervised learning.)

What ChatGPT has done is to popularize the technology and showed some of its potential. FMs, in particular large language models (LLMs) are good at many things. Specifically, they

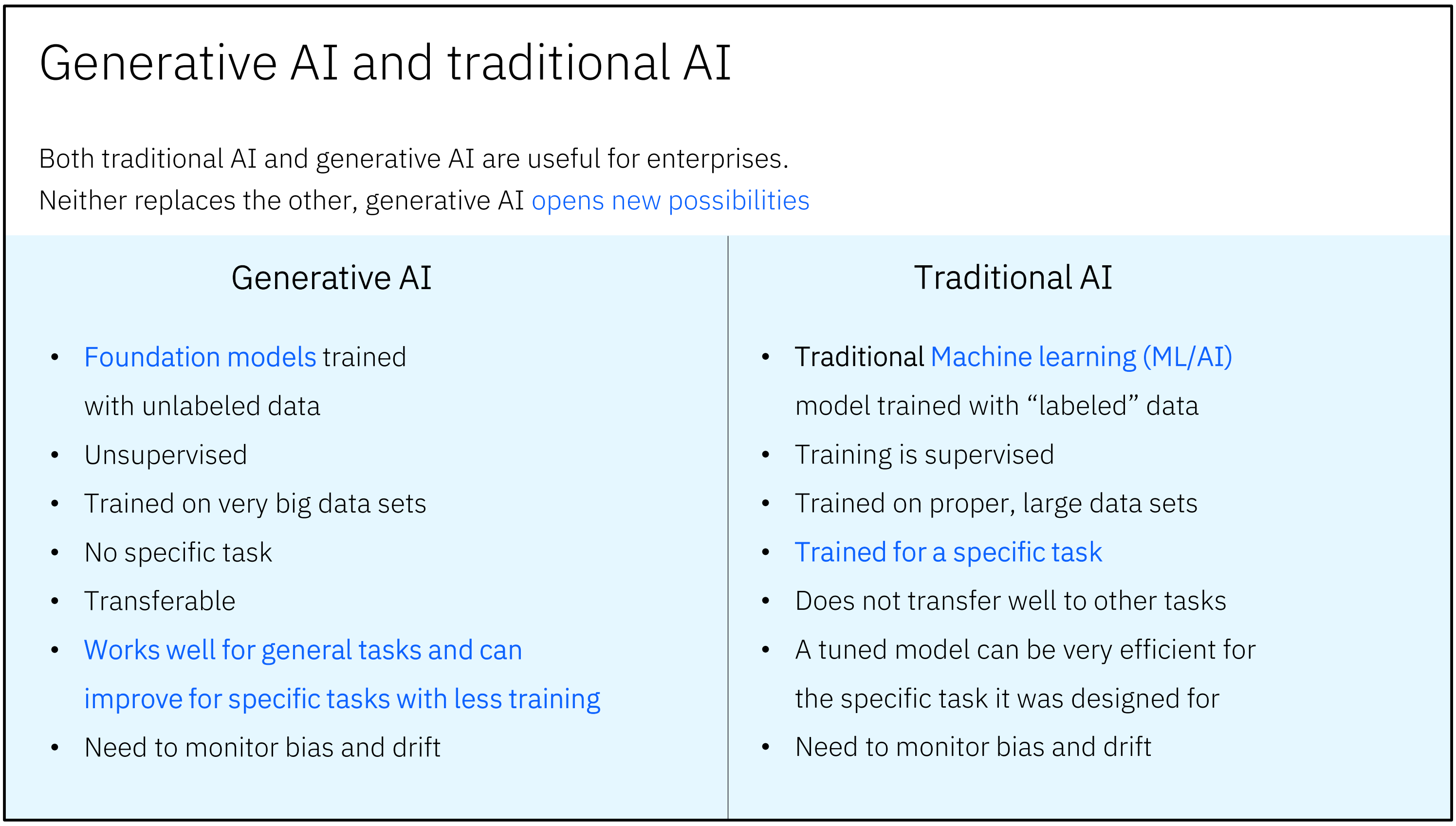
- Are good at performing question and answer (Q&A) tasks – users can raise random questions at ChatGPT and get pretty good answers.
- Can summarize a large body of texts.
- Can generate new content like a story and can be instructed to write in a particular style.
- Excel at general natural language processing (NLP) tasks.

But FMs can so much more, ranging from various NLP tasks like summarizing, extracting, Question and answering, code generating, and for general usage, to domain-specific applications in finance, cybersecurity, and much more. Given their remarkable performance and extensibility to a wide range of tasks, is bringing an inflection point in AI. How businesses adopt FM and execute generative AI will have a significant impact on future success.

It will determine:

- Whether businesses can unlock value from FM and generative AI
- Whether businesses can create value for their customers
- Whether businesses can create and unleash innovation at scale with speed to their value customers, and to

leap ahead of competitions



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

This slide contrasts traditional AI with generative AI.

First – it is important to point out that generative AI and traditional AI are both useful for companies. Clients do not have to choose between generative AI and traditional AI – both have their place and are useful for enterprise use cases. Clients do not need to give up their existing investment in traditional AI. Generative AI simply opens-up new possibilities for clients.

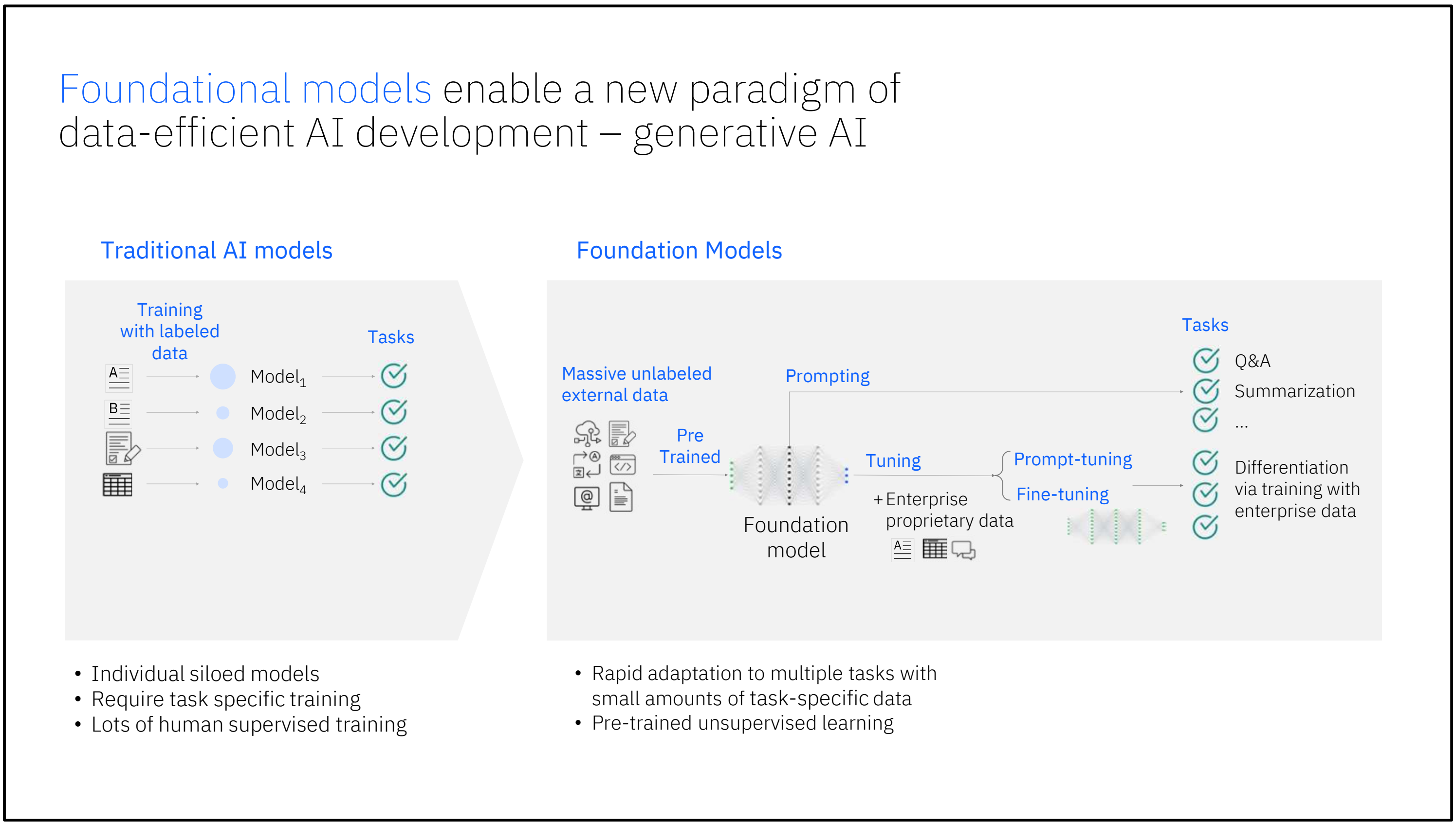
Generative AI

- Generative AI is built on a foundation model (FM). For example, ChatGPT is trained on a large language model (LLM) which is a particular type of FM specialized for natural language processing (NLP). The difference is that FMs do not require labeled data to train.
- Generative AI uses unsupervised learning.
- Training an FM requires a huge amount of data (think of traditional AI as training a person to be a biologist, or a chemist, whereas a FM would be like training someone to be a scientist who knows something about many scientific areas). This is why generative AI is difficult – it requires a massive amount of data and resources to train as the number of parameters is easily into hundreds of billions.
- Most of these models are not trained for any specific tasks. Generative AI has what is referred to as emergent behaviors – a property of FMs in which the model exhibits behaviors there were not explicitly constructed.
- FMs are therefore “general models” and are easily transferrable among different tasks as it has a very broad base. This broad base training allows generative AI models to work fairly well with general tasks. They do not work as well for specific tasks as traditional AI models (that were trained for specific task or tasks). However, they can be further trained with a much smaller set of labeled data for it to improve.
- An enterprise can take an existing FM (such as GPT) and specialize it for their use with much less effort. However, it is also more important to monitor for bias and drift as there is a huge base model that is much

harder to understand, control, or govern.

Traditional AI

- These are typical machine learning (ML) and AI models trained with a lot of labeled data. This is a human-intensive and time-consuming task.
- Training is typically supervised for best results. Users must constantly monitor, feedback, and retrain.
- Typically trained on large data sets – and they will need to be properly labeled.
- Traditional AI with labeled data is targeted for a specific task – summarizing log records to catch specific items, responding to human resource queries, and others.
- With traditional AI, clients can get transferability with their models (through a process called *transfer learning*) but this is a tedious task that involves re-training the entire and using newly labelled data it has not seen before. Traditional AI models, therefore, do not transfer well/easily to work with other tasks.
- For the task, the model was trained on (and for the particular enterprise) these AI models work very efficiently.
- AI model needs to be monitored for bias and drift to ensure that the model remains relevant.



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

Traditional AI models

Training with labeled data

Tasks

A B

Model$_1$

Model$_2$

Model$_3$

Model$_4$

Foundation Models

Massive unlabeled external data

Pre Trained

Foundation model

Prompting

Tuning

+ Enterprise proprietary data

Prompt-tuning

Fine-tuning

Tasks

Q&A

Summarization

...

Differentiation via training with enterprise data

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

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

This slide illustrates why generative AI built on foundation models provide advantages over traditional AI.

On the left is how traditional AI is built.

Traditional Ai requires:

- Individual siloed models. For every task, a particular AI model is typically built on a particular set of labeled data.
- Task-specific training for each model and its training is not generally transferable (at least not easily without re-training with a completely different set of data). It can be very time-consuming when there are a lot of tasks (even if they are somewhat similar)
- Lots of human effort and cost is spent on labeling data alone.

Given that an enterprise will require hundreds to thousands of automated tasks, this is very expensive and is often why AI adoption stalls from incubation to production.

On the right is how generative AI is built.

Unlike traditional AI, foundation models (FMs) do not begin with labeled data but are pre-trained on massive amounts of data. Specifically:

- Instead of building each model for a specific task, FMs are massive multi-tasking models that can be applied to many different tasks.
- FMs are adaptable with little or no training to support a multitude of different tasks.
- These models are pre-trained and require no supervision. Due to the amount of data used, training massive

FMs may require a lot of Graphical Processing Units (GPUs) and memory, but FMs do not need the large and costly human capital required to come up with specific labeled data for various tasks.

- FM models:

  - Are built to be very capable as is. These can be used to support many lower-level, less specific AI tasks.
  - Allow clients to modify the prompt to optimize and train the model on desired behavior and output
  - Can also be domain-specific, such as models built on financial data, cybersecurity data, geospatial data, etc. These are much more capable of supporting more specific tasks – again, without much intervention or input. IBM plans to provide these domain-specific models based on the best sets of data available.
  - Generative AI platforms (such as IBM watsonx.ai) typically offer a selection of FMs (open-source and proprietary)

- FM models support many tasks that are either not possible or much more difficult to achieve with traditional AI. They provide enhanced capabilities for many Natural Language Processing (NLP) tasks (up to 70% reduction in certain tasks) such as:

  - Summarization of text
  - Conversational knowledge
  - Content generation (email, etc.).
  - Question answering

Clients can customize and train their FMs with their own data to support the specifics they require (for example, unique business terms and proprietary data).

With IBM watsonx.ai, clients always own the data and any model they train. The key takeaways are that FMs:

- Take much fewer resources and labeled data to train (versus traditional AI)
- Can be easily used for multiple, similar tasks with no training, or with very little training, via a different set of labeled data.
- Help clients turn around new AI models based much quicker, with a lot less overhead than ever before.
- Can be trained with the business's data, and they can provide true differentiation based on the quality of domain-specific data.

Impact of generative AI

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

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

Massive early adoption

80%

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

Broad-reaching and deep impact

Generative AI could raise global GDP by

7%

within 10 years

Critical focus of AI activity and investment

Generative AI expected to represent

30%

of overall market
by 2025

Even though generative AI is relatively new, the widespread popularity of ChatGPT has created significant interest in the notion of large language models (LLMs) and foundation models (FM) and what they can do. It took quite some time for enterprises to start moving towards traditional AI. In contrast, generative AI has a massive early adoption: 80% of enterprises are already working with, or planning to leverage, FM and to adopt generative AI in their use cases and workflow.

Moreover:

- Scale Zeitgeist 2023 AI Readiness Report shows that with the companies they reviewed, 21% have generative AI models in production; 29% are experimenting with generative AI and another 31% planning to work with generative AI models; a total of 81% are either working with or planning to work with generative AI models.

- Goldman Sachs has estimated that generative AI will have a very deep economical impact – raising global Gross Domestic Product (GDP) by 7% within 10 years – which reflects its huge potential.

- Boston Consulting Group (BCG) noted that generative AI is expected to represent 30% of the overall market by 2025.

All these data point to the extremely high adoption trend for generative AI.

It’s great for consumer applications and there are some business applications it will help too, but the discussion should really be about what is the technology that's going to help businesses move forward.

Sources:

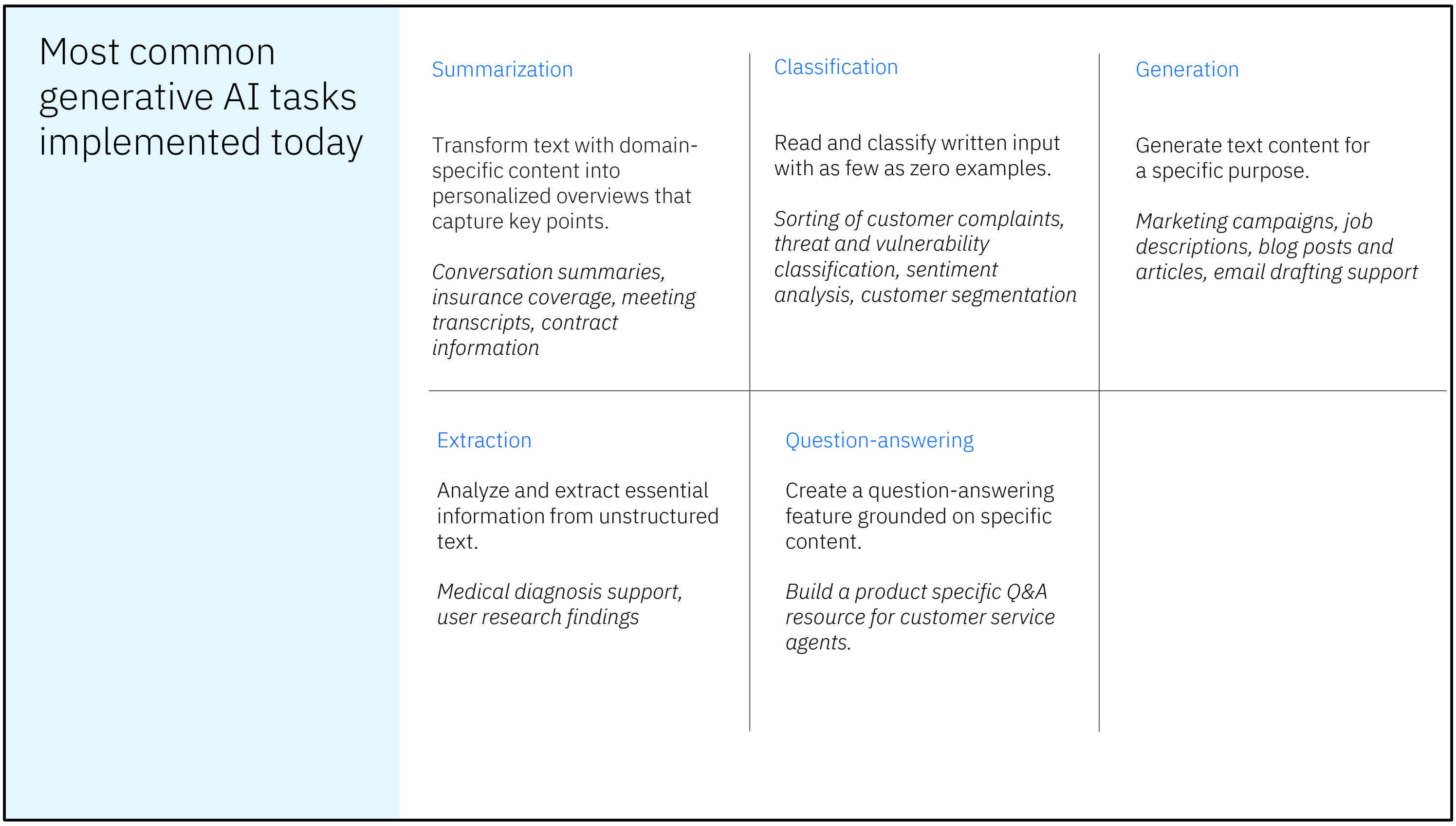
- ChatGPT took 5 days to reach 1 million users: https://tinyurl.com/yckfjxas

- Reuter on ChatGPT sets record for fastest-growing user base: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

- Goldman Sacks on generative AI could raise global GDP by 7%: https://www.goldmansachs.com/insights/pages/generative-ai-could-raise-global-gdp-by-7-percent.html

- Gartner on generative AI: https://www.gartner.com/en/topics/generative-ai#:~:text=We%20predict%20that%20by%202025,marketing%20copy%20and%20personalized%20advertising.

- Scale Zeitgeist: AI Readiness Report, a survey of more than 1,600 executives and ML practitioners: https://go.scale.com/hubfs/Scale-Zeitgeist-AI-Readiness-Report-2023.pdf



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 has a lot of applications, this self-explanatory slide shows some of the most common ones encountered today.

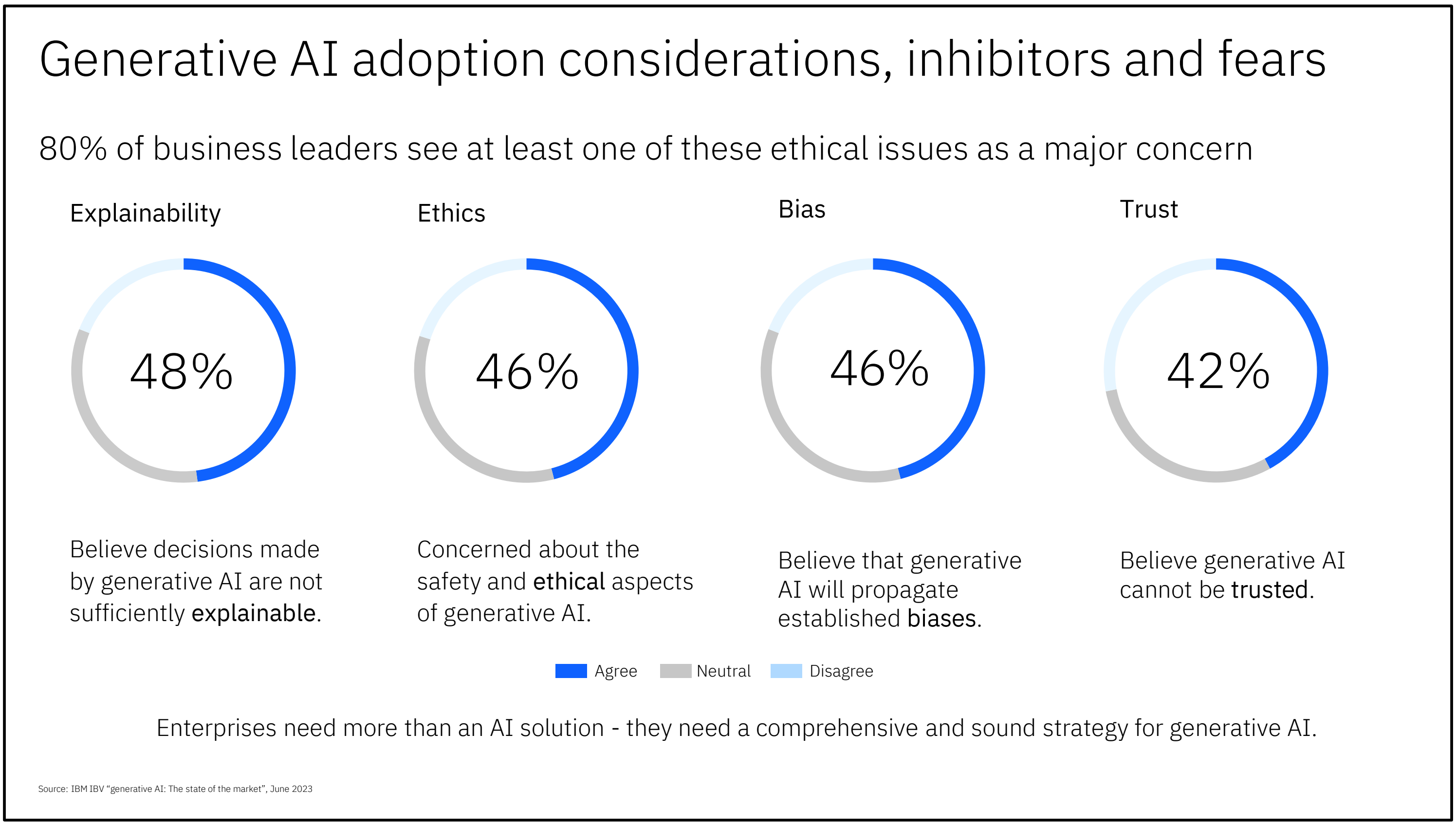
1. Summarization – Take a large body of text (such as meeting minutes, contract documents, etc.) and summarize the key points.

2. Classification – Read an input and provide classification based on business terminologies and specific categories. This can apply to sentiment analysis, security vulnerability classification, and more.

3. Generation – The main capability of ChatGPT type chatbot – this generates text content for specific purposes. In the business context, this can be for email drafting, marketing material creation, job posting, etc.

4. Extraction – Use to analyze and extract essential information. Enterprises now have lots of unstructured data from PDF, Parquet, JSON and other open sources. Generative AI can be used to extract information.

5. Question answering – generative AI will search documents or other dynamic input to support question and answer applications such as chatbots. This is useful for building a question and answer (Q&A) platform for both external (to find out about specific products or processes to join an early preview) and internal customers (for example, for looking up company policies or processes).



Generative AI adoption considerations, inhibitors and fears

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

Explainability

48%

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

Ethics

46%

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

Bias

46%

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

Trust

42%

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

This slide shows some major factors that are inhibiting the adoption of generative AI for business. While businesses recognize its power, about 80% of business leaders do see that there are ethical concerns with adopting generative AI. As a general rule, enterprise concerns can be broken down into the following categories:

Explainability

- How to understand and explain a model's output? This requires an understanding of how the model was created, the training data that was used to fine-tune the model, and who has control and access to update the model. Who might have changed the model for what reasons? There is also the need to understand the type of data used to train the model.

Ethics

- There are lots of concerns over ethical issues with generative AI and for good reasons. Companies need to uphold their hard-earned reputation and protect themselves from generative AI models that output hateful, biased, or unethical behavior which would be very damaging to the company, not to mention legal ramifications.
- Just because something "works" does not mean it is compliant with many ethics standards (corporate and government), and various government regulations, for example on data privacy.
- As generative AI becomes more prevalent, there are likely to be increasingly more regulations. Clients want to work with vendors committed to the highest level of accountability, and responsibility in complying with regulations, and in upholding high ethical standards.

Bias

- This is a general issue with all artificial intelligence (AI) models and generative AI is not immune from it. Indeed, because of the way generative AI models are trained and used, it is even more important and at the same time more difficult. There is a fear that generative AI will simply propagate established bias, or even magnify it.
- There is a strong requirement that the models must be checked for bias before being released and automatically monitored for bias once released.

Trust

- Can generative AI models be trusted? There are two aspects:
    - Data being entered into generative AI must not be used for any other purpose aside from prompting/training the model.
    - Outcomes from the model need to be trusted – trained on curated data, explainable, transparent, and monitored.
- Without trust, generative AI can be no more than a powerful and interesting chatbot.

Other issues

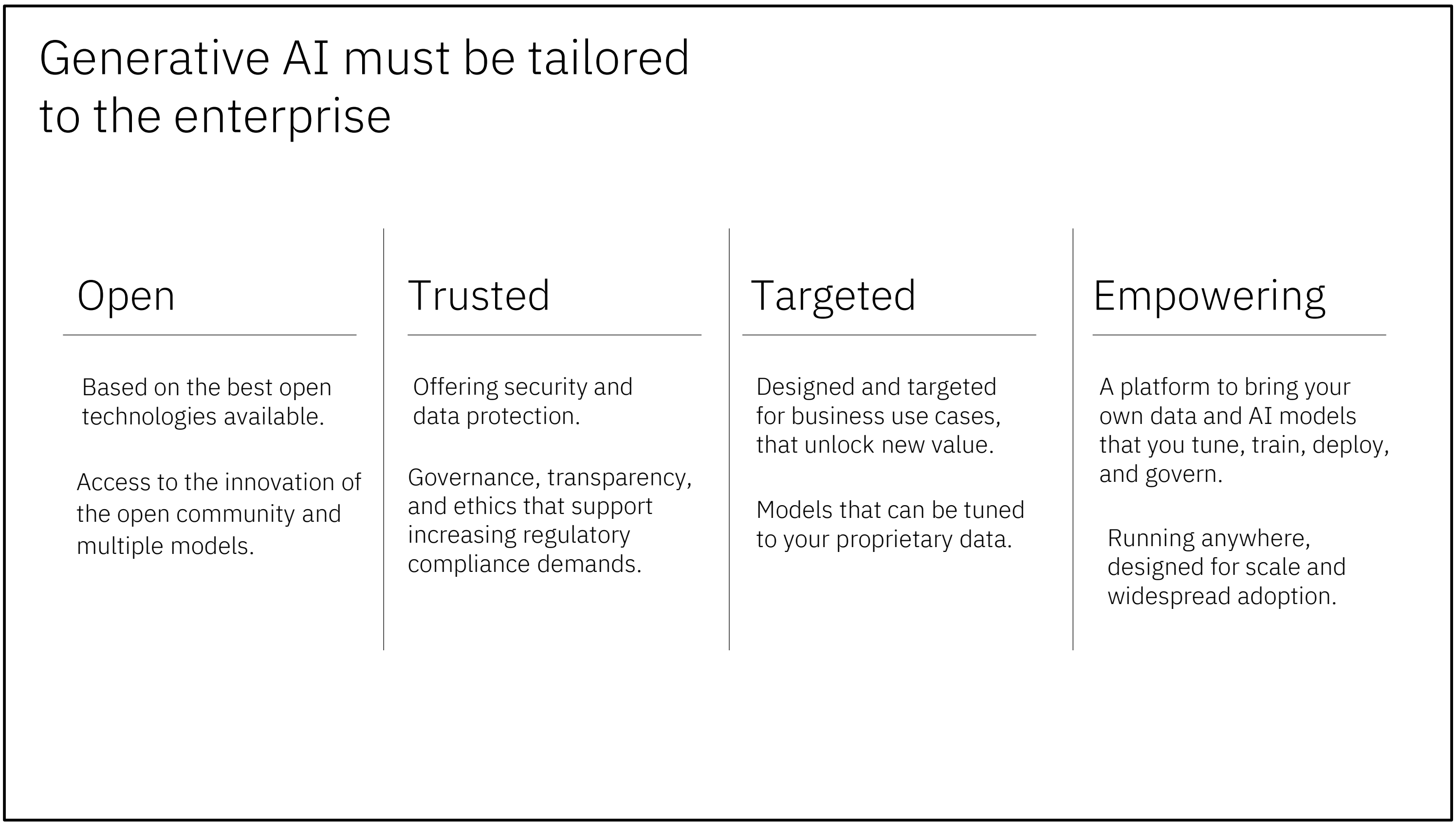
- Hybrid and multicloud capabilities: Clients that need to observe straight data/application locality will need a hybrid platform, and multi-cloud options to avoid vendor lock-in.

- Customization: generative AI on foundation models work reasonably well, but not for specific enterprise tasks as enterprises have very specific data, business terms, business processes, and other in-house rules. The first question clients must ask is: "How can the model be adapted to support specific enterprise use cases?" The second question clients must ask is: "What tools are available for the enterprise to hone/train the model?" The AI tools a company depends on must have guardrails, traceability, and transparency.

Clients cannot just simply adopt foundation models that cannot do what they need. Clients need to be able to trust an AI company with their data, reputation, customers, and above all, their businesses.

Source:

- IBM IBV "Generative AI: The state of the market", June 2023: https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/generative-ai-data-story (Click on "Find the client-ready content here" button then select "PRESENTATION DECK_Generation AI State of the market.pptx" link)



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.

This slide highlights IBM's viewpoint on some vital characteristics of generative AI and foundation models (FMs). Businesses want to have a holistic strategy – one that embraces traditional AI and generative AI – the latter must be tailored to the enterprise. There are 4 aspects as listed:

- Open – Open source is a rich source of innovation and is a rich source of many FMs (such as those available on Hugging Face). A generative AI platform needs to be open – or it will restrict what is available to its customers and lock them in. One model will not rule them all.

- Trusted – There are already well-known issues with generative AI platforms like ChatGPT or applications like Google Bard. Generative AI models are known for hallucination (making up answers). Generative AI platforms need to be trusted, offering security, and data and model protection. They also need to be architected with governance in mind from the start, not an afterthought and provide transparency, and explainability for their models to support increasing regulatory compliance demands.

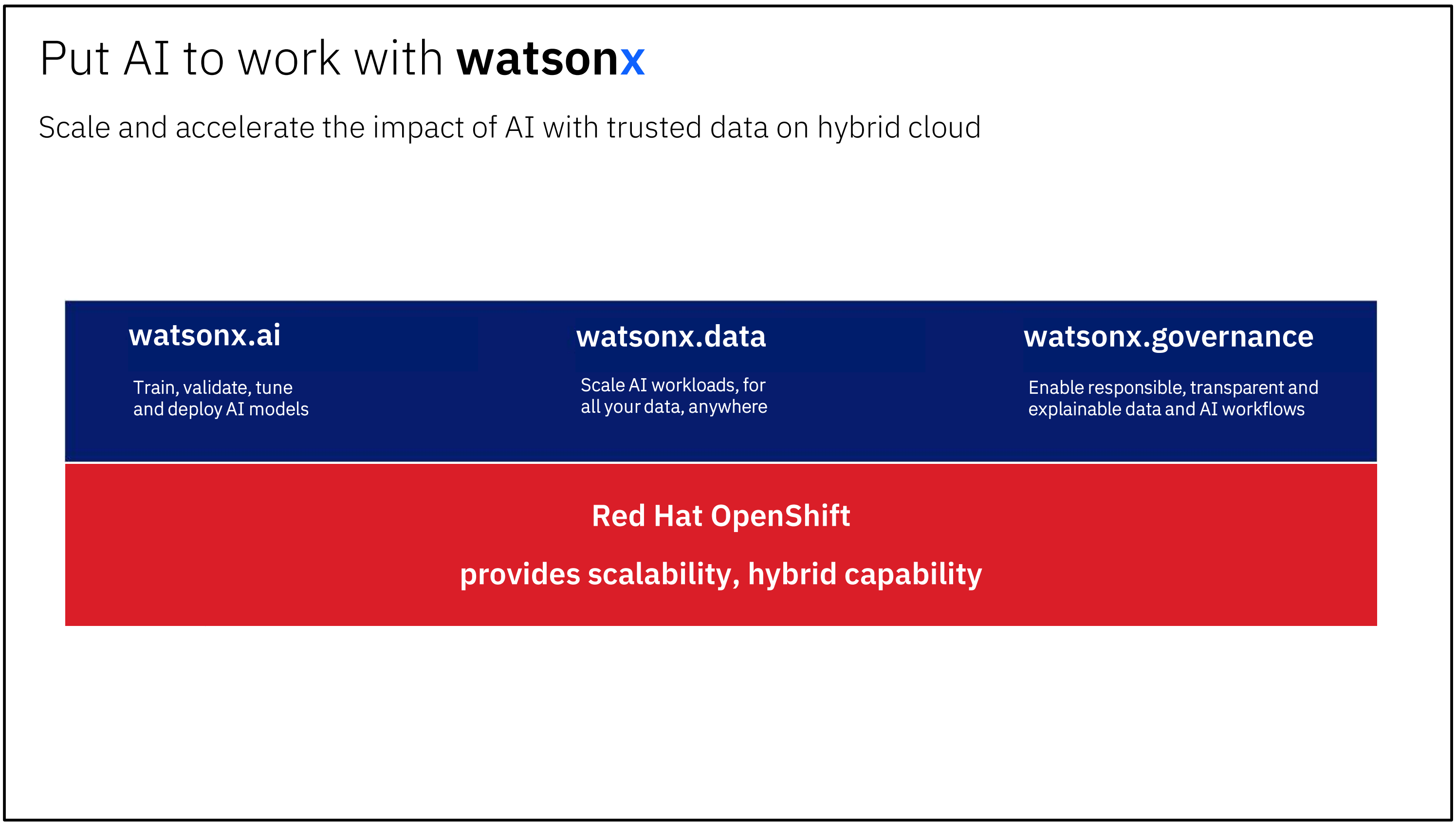
- Targeted – platforms like ChatGPT are not targeted. They are used for general consumption and are "generalists" – they are trained on a huge amount of data and require a large overhead to run (clusters of GPU nodes) and are expensive to run (inference) or to fine-tune. Enterprises should look for platforms that provide targeted models (or domain-specific models) for business use cases that can be quickly, effectively, and economically tuned with a small set of proprietary data from the business.

- Empowered – generative AI is not just about FMs, it must provide a platform that empowers enterprises to bring their own data to tune, train, and deploy generative AI models. The platform must also provide governance to ensure proper, responsible usage of the AI models. Moreover, it needs to be able to support running wherever the business wants to, at a desirable scale.

Introducing…

watsonx.ai

The second part of this presentation introduces the new IBM watsonx platform and in particular watsonx.ai, which provides foundation models for generative AI.



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

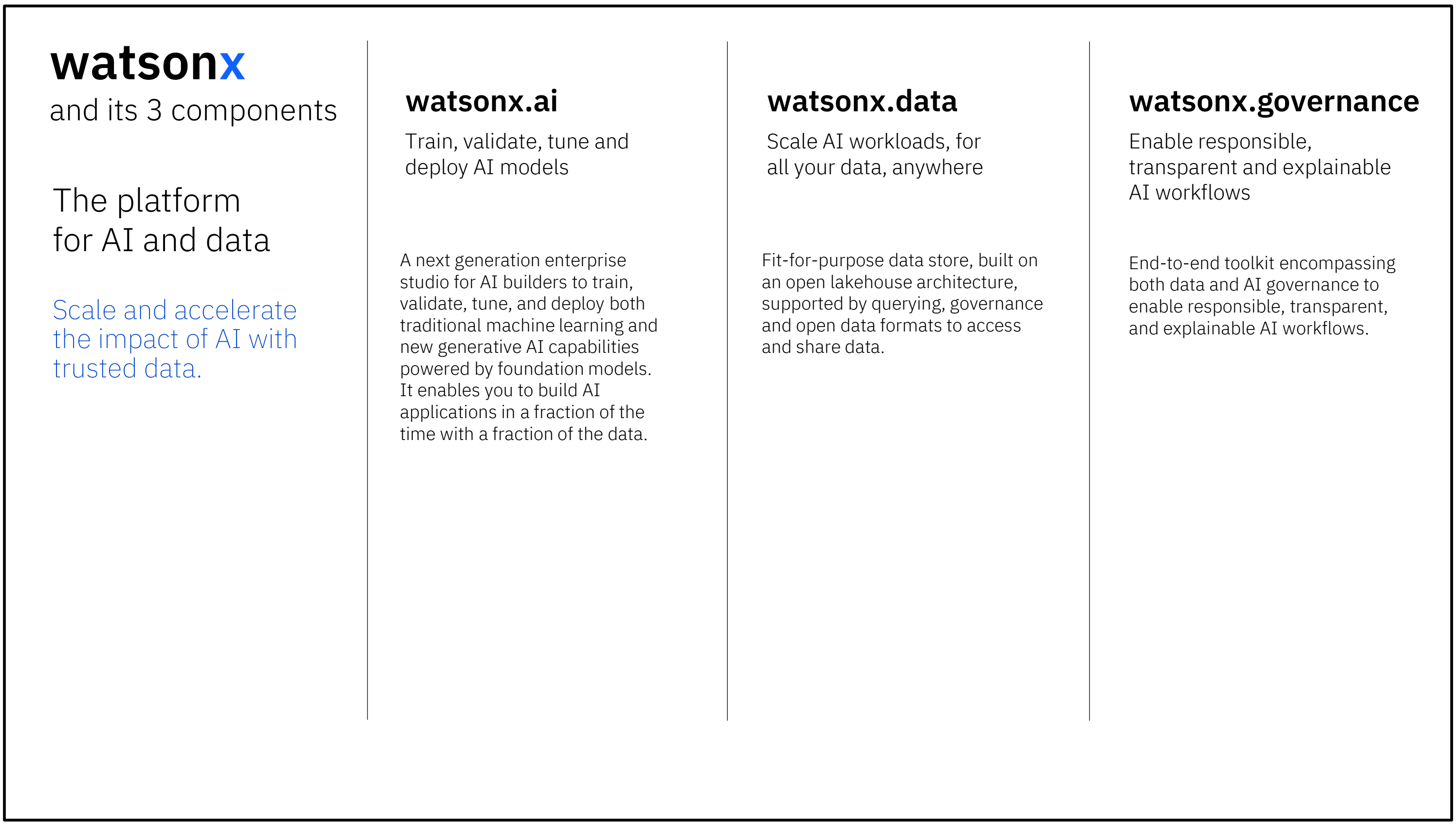
Enterprise AI needs to be portable, efficient, and sustainable. But the largest models are expensive, energy-intensive to train and run, and complex to deploy. By building on hybrid cloud technologies, IBM allows clients to optimize for performance, latency, and cost.

Watsonx is built on Red Hat OpenShift to allow clients to access and deploy their AI workloads in any IT environment, no matter where they're located. It can run on multiple clouds, including IBM Cloud, Amazon Web Services (AWS), and Azure, on clients' premises, and even on the edge.

Watsonx allows clients to move their AI workflows between public and private clouds seamlessly and safely, making it easier to securely process and store data on servers that they own or lease. This gives clients the flexibility to rapidly scale AI applications. If clients have data in the cloud, IBM can bring its stack to them. If clients' data is on-premises, they can train and tune their models there with the same cloud-native, user-friendly experience. Clients can then deploy their models and performance inference on a variety of computing environments, be it on clouds, on-premises to meet security and/or governance needs, or on the edge to meet latency requirements.

Together, watsonx.ai, watsonx.data, and watsonx.governance scale and accelerate the impact of AI with trusted data – data that is governed, and curated, with business logic and classifications applied. This allows clients to train, tune, and deploy AI across the client's businesses, leveraging trusted data where they reside.

The classic AI and machine learning (ML) capability of IBM Watson Studio is still there accessible from watsonx.ai, just not shown in the picture.



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 is a new artificial intelligence (AI) and data platform from IBM that is designed with the three critical elements (AI, data, governance) of an AI strategy in mind. It empowers enterprises to train, tune, and deploy AI across the business, leveraging critical, trusted data wherever it resides. The watsonx platform has three components:

- Watsonx.ai is a studio that clients can use to train, validate, tune prompts, and deploy both traditional AI models, as well as foundation models for generative AI. These models combine best-of-breed architectures with a rigorous focus on data acquisition, provenance, and quality, to serve enterprise needs.

- Watsonx.data makes it possible for enterprises to scale AI workloads using all their data with a fit-for-purpose data lakehouse service optimized for governed data and AI workloads, supported by querying, governance, and open data formats to access and share data. This is based on open-source technologies, including Presto and Iceberg (and more) and tests have shown incredible performance when compared to a traditional data engine.

- Watsonx.governance helps companies put AI into production by providing an end-to-end solution that encompasses both data and AI governance to enable responsible, transparent, and explainable AI workflows. AI governance helps business analysts understand the trustworthiness of their AI solutions.

Watsonx is easy to remember: one platform, and three watsonx main capabilities. IBM's strength is that watsonx is not just a foundation model and generative AI offering, clients get a complete suite of tools that:

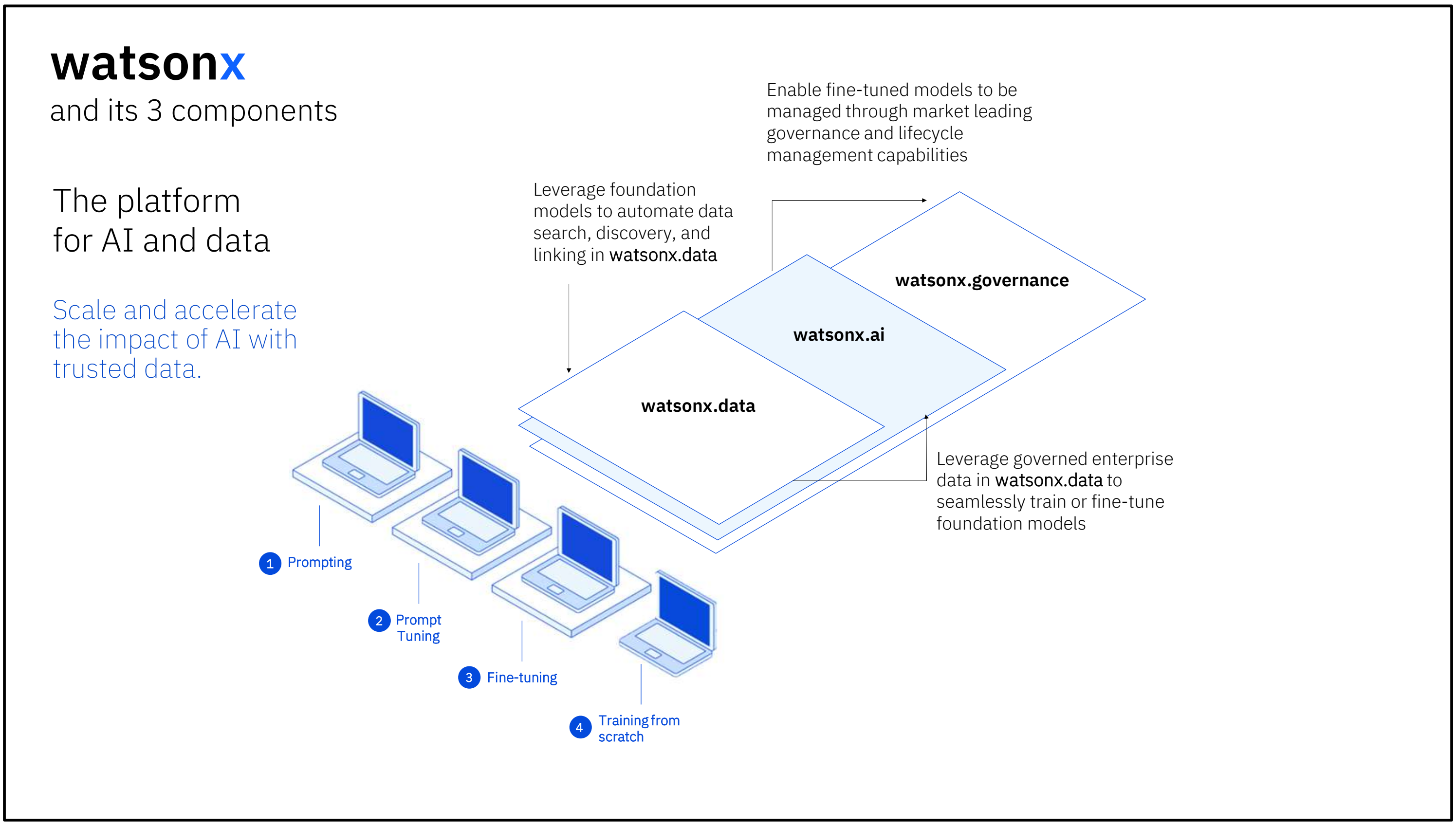
- Provide rich data capabilities: hybrid, multicloud, high performance, structure, semi-structured, and unstructured. This is the data enterprises will use to fine-tune their foundation models.

- Governance of the foundation models and modifications. Transparency, accountability, lineage, and traceability are key items that bring high confidence in the integrity, controllability, and safety of the generative AI features.

The three watsonx components work together. IBM watsonx provides a complete ecosystem that satisfies enterprises' need for transparency, safety, and the ability to customize.

In summary:

- Clients can use watsonx.ai to perform AI workloads such as data search, and discovery on their data in watsonx.data.
- Clients can use their own *governed* data in watsonx.data to fine-tune and train their watsonx.ai models
- Clients can *govern* their watsonx.ai models, providing transparency, explainability, and control over their watsonx.ai models.



watsonx

and its 3 components

The platform
for AI and data

Scale and accelerate the impact of AI with trusted data.

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

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

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

This self-explanatory slide positions the three components of watsonx: watsonx.ai, watsonx.data, and watsonx.governance.

As noted on the previous slide (and repeated here purposely) IBM's strength is that watsonx is not just a foundation model and generative AI offering. Enterprise clients get a complete suite of tools to scale and accelerate the impact of generative AI with data and models that enterprises can trust.

- Watsonx.ai leverages watsonx.data to provide rich data capabilities: hybrid, multicloud, high performance, structure, semi-structured and unstructured. Watsonx.data provides vital and necessary governance on enterprise data used by watsonx.ai to train and fine-tune their foundation models (FMs).

- Watsonx.ai leverages watsonx.governance to provide FM governance. Enterprises cannot just train, fine-tune, and roll out models, they need these models to be responsible, transparent, and explainable. A generative AI platform must include capabilities to support these requirements. With watsonx.governance, clients' fine-tuned models can be managed through market-leading governance and lifecycle management capabilities with rich experience and lessons learned from IBM Watson Studio and IBM Watson Machine Learning.

- Watsonx.data leverages watsonx.ai to provide FMs that support many important data use cases such as automated data search, and discovery, and provide important natural language processing (NLP) capabilities to simplify day-to-day data tasks for non-database specialists.

Clients can perform several tasks with FMs via watsonx.ai

- Prompting – the simplest way of interacting with FMs. Watsonx.ai provides guidance and examples to help clients craft the best prompt and test with various models.

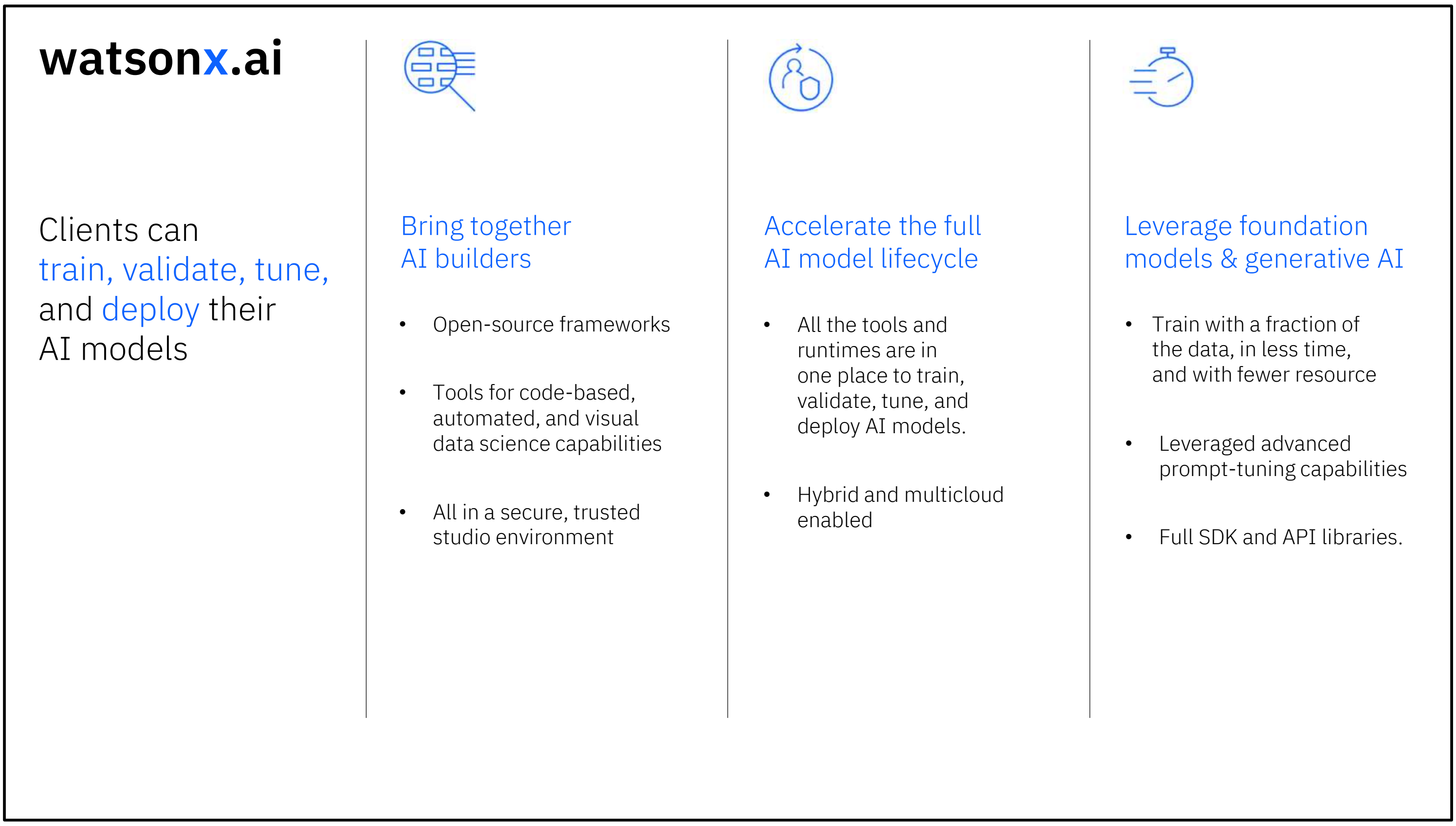
- Prompt tuning – where clients can provide a small sample set of labeled data to train the model. This requires more effort but can vastly improve performance, especially on specific tasks.

- Fine-tuning – where clients provide more labeled data and change the parameter weights of the model. This is much more labor intensive.

- Train from scratch – instead of using an existing model, clients can train their own model. This requires a large set of data as well as compute resources.

Prompting is available at GA with the others following shortly.

Similar to IBM's cloud offerings, the watsonx family is all built on RedHat OpenShift technology. Instead of a series of products with different capabilities and features, these three components will all share the same user interface/user experience (UI/UX). In addition, this allows for common connectivity and integration not just among watsonx, but also with other IBM offerings (such as IBM Cloud Pak for 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.

With **watsonx.ai**, clients have a complete AI platform that empowers clients to: train, validate, tune, and deploy their AI models, both traditional and generative. This slide focuses on the generative.

- **Bring together AI builders** because it offers open-source frameworks for clients, and it provides tools for code-based, automated, and visual data science capabilities for traditional predictive, and prescriptive AI and generative AI. Everything in **watsonx.ai** is accessed in a secure, trusted studio. The same security framework for IBM Watson Studio is in place to secure clients' models and their data.

- **Accelerate the full AI model lifecycle** because all the tools and runtime that clients require to train, validate, and deploy AI models are in one place. Clients do not need to integrate different (and potentially incompatible tools) to support their end-to-end AI cycle. What's more, **watsonx.ai** (and **watsonx** in general) is built on Red Hat OpenShift; as such it can be made available wherever the clients need it. It is both hybrid and multicloud enabled and can support any locality-related regulations.

- **Leverages foundation models and generative AI** in addition to traditional AI and machine learning. With **watsonx.ai**, IBM provides various open-source foundation models, as well as IBM's trained foundation models. Clients can start with these and train them for their specialized needs with much fewer resources and data (than building their models).

Finally, **Watsonx.ai** provides an easy interface for prompt tuning and model tuning (with the client's data and input). Application programming interfaces (APIs) are available to develop applications and interfaces to deploy and use the foundation models in generative AI tasks. A full software development kit (SDK) will also be rolled out post-July 2023.

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 part of the IBM watsonx platform that provides foundation models (FMs) and generative AI capabilities, plus traditional AI and machine learning modelling that clients have come to know (in IBM Watson Studio and IBM Watson Machine Learning).

These are the components in watsonx.ai:

The generative AI capabilities in watsonx.ai include the following:

- Foundation model library – Includes many well-known models such as Generative Pre-trained transformer (GPT), Fine-tuned Language Net (FLAN), and more, as well as IBM's own foundation models. Many of the models have been rebuilt on IBM's data sets to reduce model sizes, mitigate problems (such as bias) and other common risks.

- Prompt lab – Prompt engineering is a way of helping FMs become more accurate and precise in their response by providing some user interaction (called prompts). Prompt engineering is an "art" and requires practice. IBM provides samples of prompts and prompt templates, and the studio helps users to try different prompts and different models to get the best results.

- Tuning studio – there are multiple ways that watsonx.ai lets clients fine-tune FMs:

    - Providing additional data to further train the model

    - Changing various runtime parameters such as
        - Number of virtual tokens
        - Batch size

    - Gradient accumulation steps
    - Learning rate
    - Max input/output tokens
    - Number of epochs

  - These parameters can alter the amount of time (resource) required to train the model when prompt tuning.

Plus a proven studio for machine learning (ML)

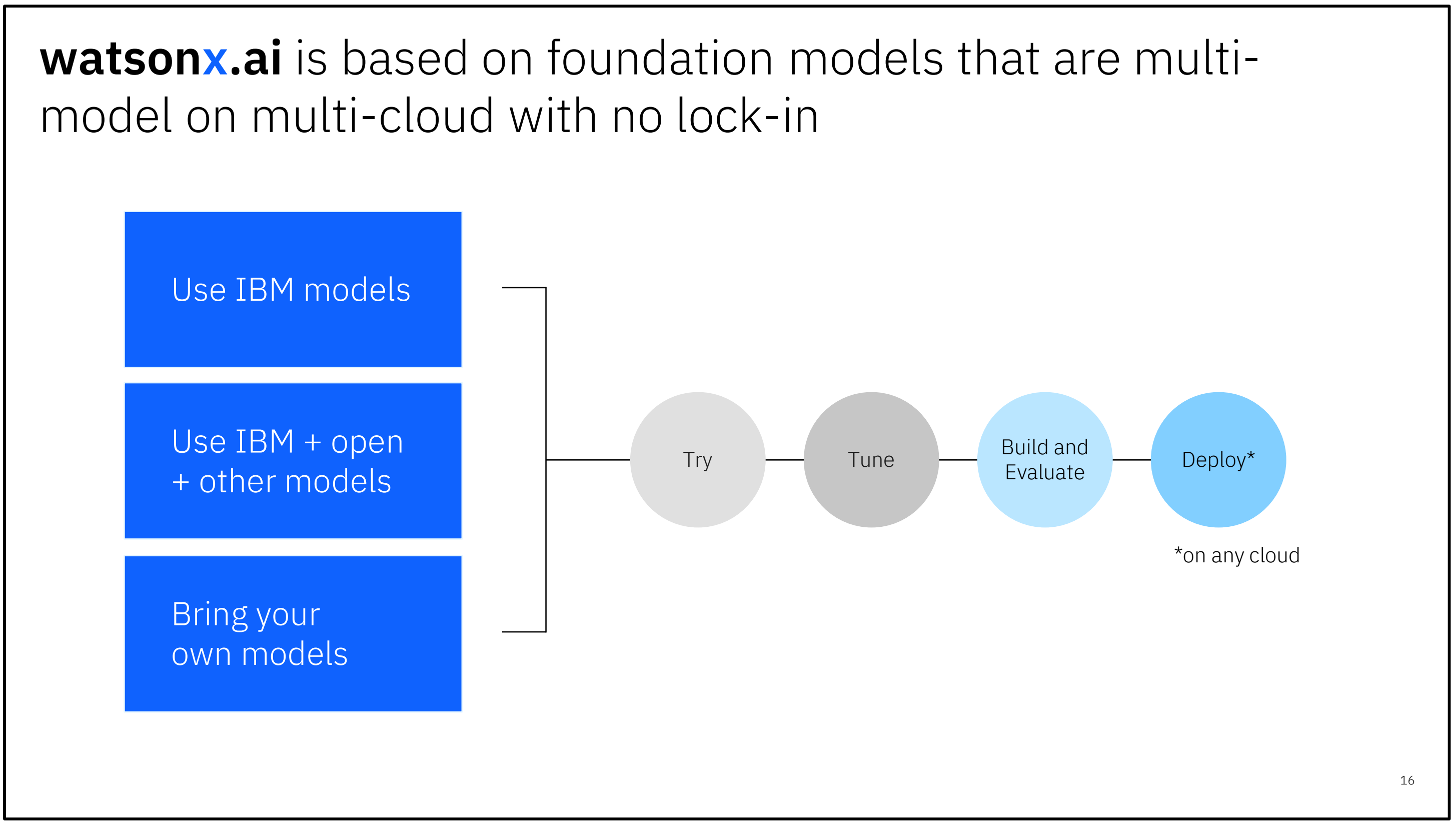
Clients continue to have access to familiar features from IBM Watson Studio in watsonx.ai such as

- Model operations: training, deploying, managing, and more
- Automated development
- Decision optimization

In addition, team collaboration and data preparation is available for all models.

Note:

- The watsonx.ai Tuning studio will be rolled out post-July 2023. However, it is important to note that the base AI and machine learning capabilities (IBM Watson Studio, IBM Watson Machine Learning, etc.) are available to clients today and client can continue to use them.



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

Use IBM models

Use IBM + open + other models

Bring your own models

Try

Tune

Build and Evaluate

Deploy*

*on any cloud

16

IBM's AI solution includes both traditional AI (IBM Watson Studio and IBM Watson Machine Learning) as well as generative AI. The focus here is on generative AI.

Some important notes for clients to consider:

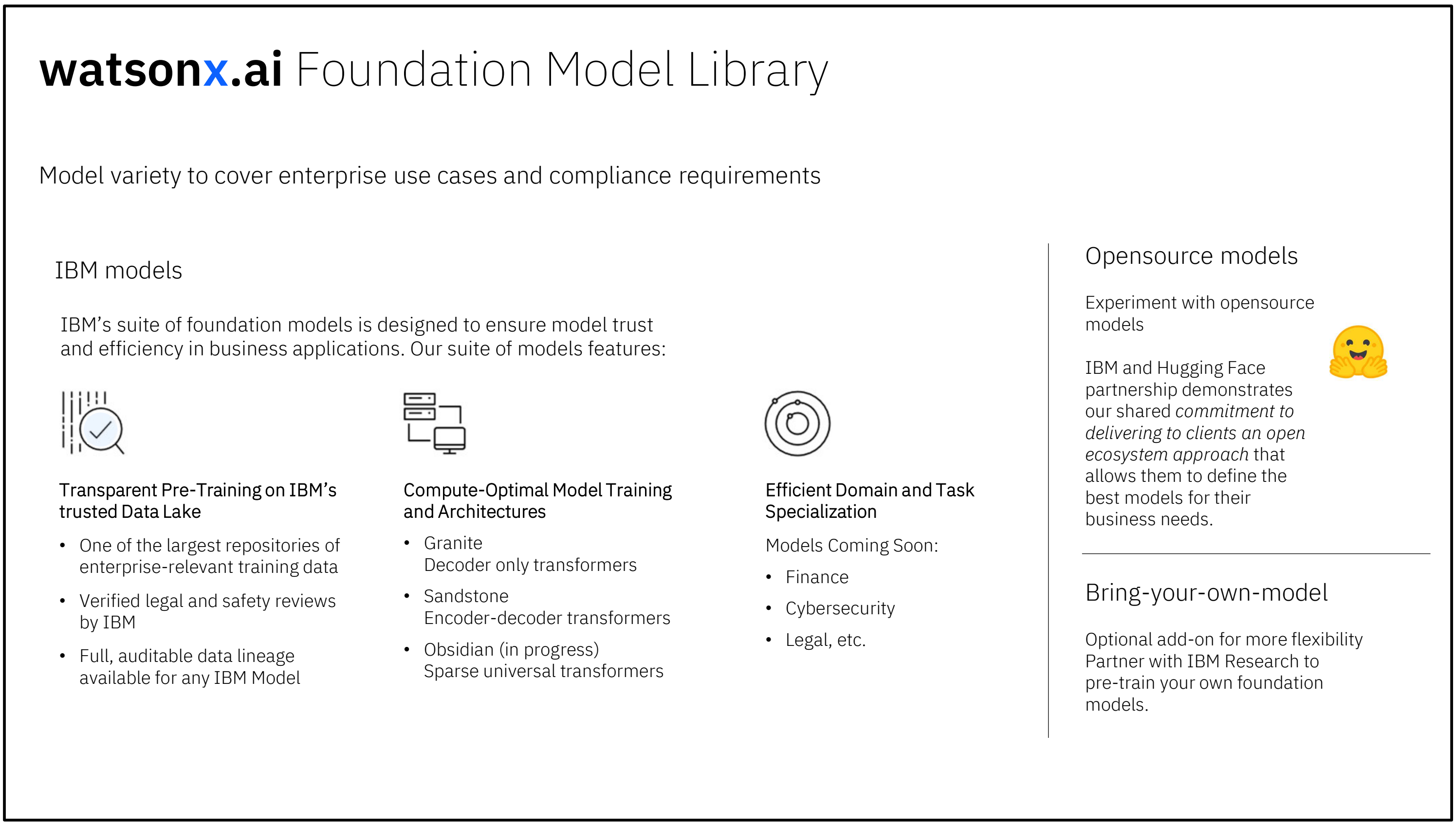
- One model will not rule them all. Instead of one model to address all use cases and performance needs, clients will benefit from many models — IBM models, open-source models, and as clients participate in the creation process, they may have their own models. The global open AI community is incredibly vibrant and creative. It offers a wide diversity of innovations and models at an accelerated pace; and through IBM's partnership with Hugging Face, IBM is incorporating the best open-source models, datasets, and libraries into the watsonx.ai platform to increase the value of foundation models (FMs)to clients' businesses.
- Because one size does not necessarily fit all, IBM is building "families" of FMs of different sizes and architectures and providing multiple tuning methods from which to choose. IBM's model architectures bring together cutting-edge innovations from IBM Research and the open research community to deliver performance, speed, and efficiency — architectures optimized for generative tasks, or for fine-tuning on specific tasks, for high inference efficiency and performance, or speed and effectiveness for enterprise natural language processing (NLP) tasks. Furthermore, they can be customized for a range of enterprise tasks in customer care, digital labor, IT operations, and cybersecurity. See backup slide on "Leveraging foundation model capabilities across various domains" for examples.
- One important IBM strategy for clients is to help infuse AI into their business processes. It is important to have application programming interfaces (APIs) and other tooling like Python notebooks to help clients exploit generative AI.

With watsonx.ai:

- Clients can leverage available foundation models – these include IBM proprietary models as well as various open-source models from GPT to Hugging Face models. Clients can also build their own generative AI applications with watsonx.ai.

- Clients can examine their prompt history – prompt engineering is an art and sometimes the more one tweaks, the less effective it may become. The prompt history allows clients to easily go back to previous prompts to find the optimal.

- Clients have access to APIs and Python libraries, starter notebooks, and a software development kit (SDK) to provide easy and simple entry points for new generative AI users, and for data scientists to quickly determine how to exploit foundation models and generative AI. This enables businesses to deploy and incorporate generative AI capabilities into business processes.



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.

IBM **watsonx.ai** provides the best of both worlds.

IBM models

Foundation models like ChatGPT are "black" (opaque) boxes and that is one big reason why trust is difficult to achieve with this kind of models. IBM's models are designed to address this gap to ensure model trust and efficiency in business applications. In particular, IBM's suite of models features:

- Transparent pre-training on trusted data

  One of the most important trust factors for models is the data used to train the model. There are 2 main issues: the quality of the data and the relevance of the data. Having random data from an internet crawl will likely inject bias, hate, and prejudice into the model. Whereas just using any data can bloat the model with parameters that may never be used. IBM has collected the largest known repository for enterprise-relevant training data covering a variety of topics like legal, finance, human resources (HR), and more. This data was then further cleansed and curated (remove duplicates, filter out bias, copyrights, etc.) to create a base pool of data used to build IBM's model. IBM models have fully auditable data lineage and transparency – allowing clients to use them out of the box with trust.

- Compute-optimal model training and architecture

  Using the trusted data, IBM is building different models that can be applied in different scenarios, some of which include:

  - **Granite** series models – are decoder-only transformers and are great for generative tasks such as generating emails and marketing materials.

  - Slate series models – are encoder-only transformers – smaller and compact but are better for non-generative use cases such as extracting and summarization and may require task-specific labeled data to fine-tune.

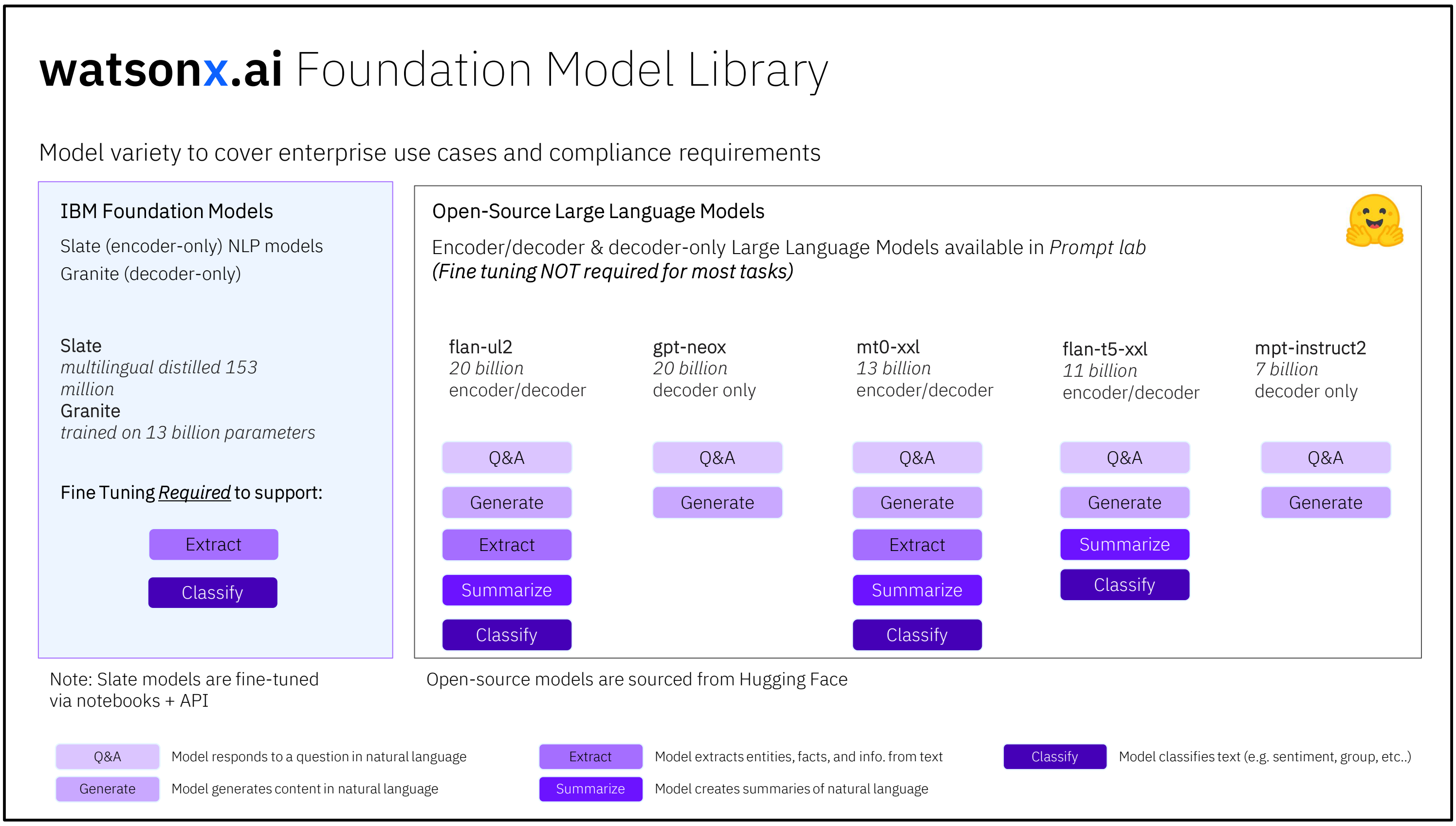
  - Sandstone series models – are encoder-decoder transformers supporting both generative and non-generative use cases. These have the best cost performance trade-off for generative use cases when the input is large but generated output is small.

- Efficient domain and task specialization

  IBM provides domain-specific models where they are pre-trained with specific relevant data. This allows the model size to be minimized with focused pre-training for specific domain use cases such as finance, cybersecurity, legal, etc.

- Open-source model

  IBM has a partnership with Hugging Face to provide access to open-source models to clients. IBM will work jointly with Hugging Face and surface the best models suited to clients' business needs.



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

Open-source models are sourced from Hugging Face

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

This self-explanatory slide describes the models available in **watsonx.ai** at general availability (GA) (July 7, 2023). Many more will be added over time.

Watsonx.ai offers:

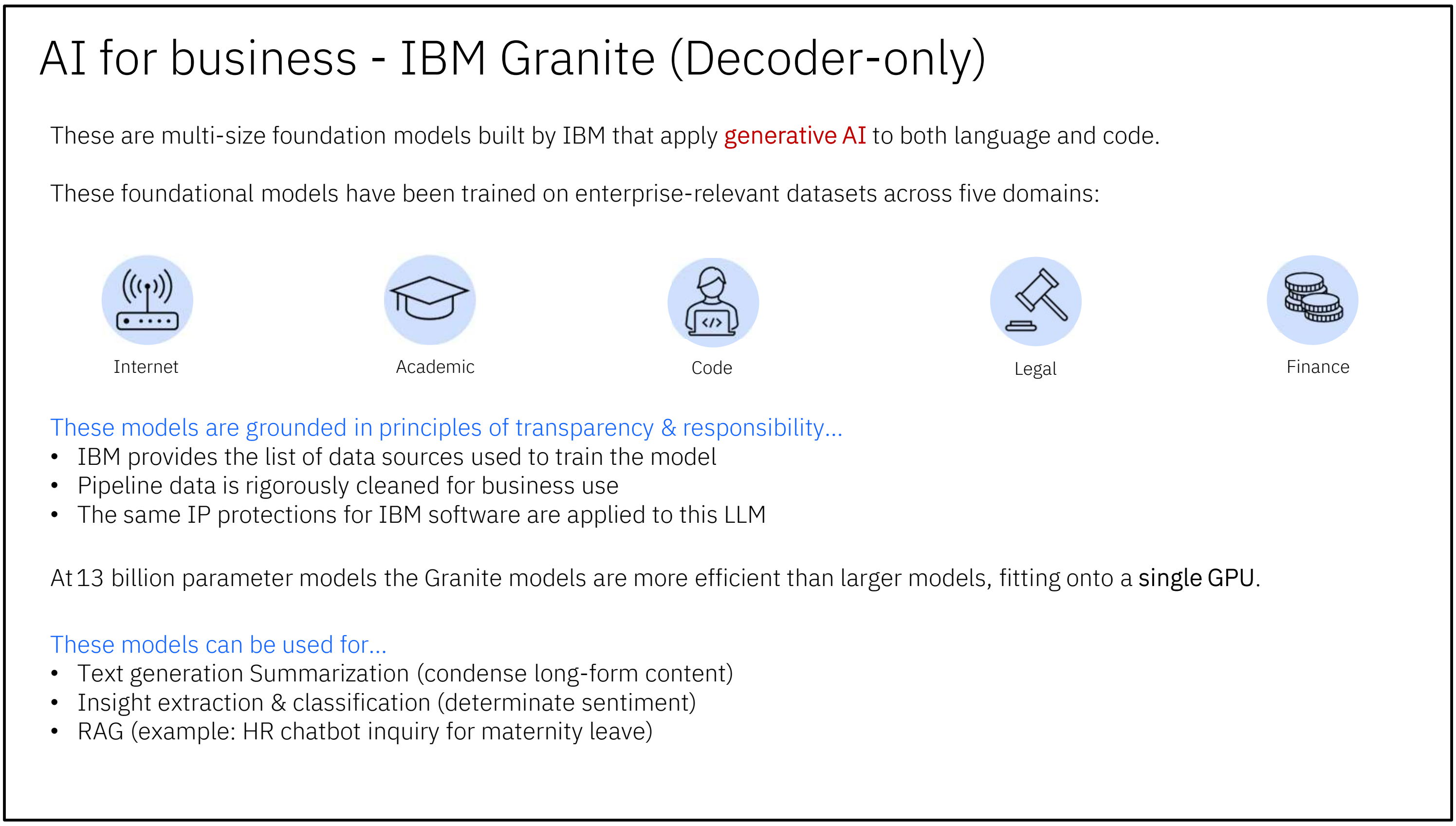
IBM foundation models

- **Slate** – Slate is an encoder-only natural language processing (NLP) model. It is a multilingual distilled model trained on 153 million parameters. Clients will need to tune this with their data to support the extract and classify use cases. Clients can fine-tune Slate via API or Jupiter notebooks.
- **Granite** – Granite is an IBM-built decoder-only model. It is trained on **7TB of data generating 13 billion parameters utilizing data from the** Internet, Academic, Code, Legal, and Finance.

Open-source large language model (LLM)

- At GA, there are five different open-source models available from Hugging Face: **flan-ul2, gpt-neox, mt0-xxl, flan-t5-xxl,** and **mpt-instruct2**. These are of different architectures (decoder, or encoder-decoder) and sizes (as signified by the number of parameters above).

- Different models are optimized for different tasks as shown above. Depending on whether the client is trying to summarize, generate, or classify a different model may perform better. In many cases, fine-tuning is not required for these models to support a task.



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)

Introduction to Granite Models

Granite models are advanced AI models developed by IBM Research. They utilize generative AI to excel in specialized business-domain tasks such as summarization, question-answering, and classification. They are built upon a "decoder" architecture, which is designed to predict the next word in a sequence without requiring knowledge of the entire sentence. This architecture is considered uni-directional, enhancing efficiency in natural language understanding tasks. The granite model comes in different forms and sizes, recognizing that a single model will not fit the unique needs of every business use case.

With these models and Watsonx, IBM enables businesses to be AI value creators. Businesses can bring their proprietary data to IBM base models and build a model that is unique to their business and use cases.

Datasets the Granite model has been trained on:

These models have been trained on diverse business-relevant datasets from five domains: Internet, Academic, Code, Legal, and Finance.

1. Internet — Generic unstructured language data taken from the public internet
2. Academic — Technical unstructured language data, focussed on science and technology
3. Code — Unstructured code data sets covering a variety of coding languages
4. Legal — Enterprise-relevant unstructured language data taken from legal opinions and other public findings
5. Finance — Enterprise-relevant unstructured data taken from publicly posted financial documents and reports

These datasets have been meticulously curated for business use.

What differentiates the Granite model series is IBM's continued commitment to advancing AI technologies that are grounded in principles of transparency and responsibility.

Granite models have an end-to-end process for building and testing foundation models and generative AI — starting with data collection and ending in control points for tracking the responsible deployments of models and applications — focused on governance, risk assessment, bias mitigation, and compliance. Since the Granite models will be available to clients to adapt to their applications, every dataset that is used in training undergoes a defined governance, risk, and compliance (GRC) review process. IBM has developed governance procedures for incorporating data into the IBM Data Pile which are consistent with IBM AI Ethics principles. Addressing GRC criteria for data spans the entire lifecycle of training data. IBM's goal is to establish an auditable link from a trained foundation model back to the specific dataset version on which the model was trained. By training models of enterprise-specialized datasets, IBM can help clients use models that are familiar with the specialized language and jargon of their business from these industries which will allow their large language models to make decisions grounded in relevant industry knowledge.

Much media attention has (rightly) been focused on the risk of generative AI producing hateful or defamatory output. IBM knows that businesses can't afford to take such risks, so the granite models are trained on data scrutinized by the "HAP detector," a language model trained by IBM to detect and root out hateful and profane content (hence "HAP"), which is benchmarked against internal as well as public models. After a score is assigned to each sentence in a document, analytics are run over the sentences and scores to explore the distribution, which determines the percentage of sentences for filtering. Besides this, there is wide range of other quality measures. For example, searching for and removing duplication that improves the quality of output and using document quality filters to further remove low-quality documents not suitable for training. These granite models were trained with 7TB of data (pre-processing) and ended up with 2.4TB of data after processing to produce 1 trillion tokens (tokens are the collection of characters that have semantic meaning for a model).

Additionally, IBM's standard intellectual property protection, similar to what it provides for hardware and software products, will apply to all IBM-developed Watsonx models.

At 13 billion parameters, these Granite models are super-efficient and fit into a single GPU, making them cost-effective and environmentally friendly.

Applications of Granite models:

By using these models, clients can perform –

Text generation Summarization - Condense a lengthy piece of text into a shorter, coherent version while retaining its essential information and meaning.

Insight extraction and classification - Identify and categorize valuable information or knowledge from unstructured data, often in the form of text. Example: Sentiment determination

Retrieval Augmented Generation or RAG – Combining retrieval of information from a knowledge base with text generation to provide context-aware, high-quality responses. Example: an HR Chatbot to enquire about maternity leave

The Granite models are part of a broader series of model families, such as Granite, Sandstone, Obsidian, and Slate. IBM has also partnered with organizations like Meta, NASA, and Hugging Face, contributing to the overall AI technology stack and advancing AI innovation.

watsonx.ai: Prompt Lab

Experiment with foundation models and build prompts

Interactive prompt builder

Includes prompt examples for various use cases and tasks

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

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

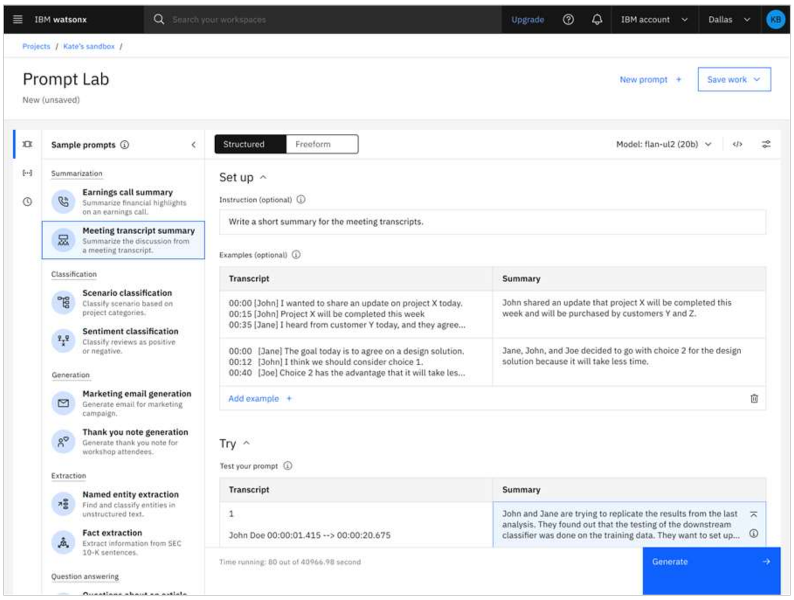
Experiment with prompt engineering

Choice of foundation models to use based on task requirements

Prevent the model from generating repeating phrases

Number of min and max new tokens in the response

Stop sequences – specifies sequences whose appearances should stop the model


IBM watsonx
Search your workspaces
Upgrade
IBM account
Dallas
KB
Projects / Kate's sandbox /
Prompt Lab
New (unsaved)
New prompt +
Save work
Sample prompts
Summarization
Earnings call summary
Summarize financial highlights on an earnings call.
Meeting transcript summary
Summarize the discussion from a meeting transcript.
Classification
Scenario classification
Classify scenario based on project categories.
Sentiment classification
Classify reviews as positive or negative.
Generation
Marketing email generation
Generate email for marketing campaign.
Thank you note generation
Generate thank you note for workshop attendees.
Extraction
Named entity extraction
Find and classify entities in unstructured text.
Fact extraction
Extract information from SEC 10-K sentences.
Question answering
Structured
Freeform
Set up
Try
Add example +
Generate
Time running: 80 out of 40966.98 second

Model: flan-ul2 (20b)

Instruction (optional)

Write a short summary for the meeting transcripts.

Examples (optional)

Transcript	Summary
00:00 [John] I wanted to share an update on project X today. 00:15 [John] Project X will be completed this week 00:35 [Jane] I heard from customer Y today, and they agree...	John shared an update that project X will be completed this week and will be purchased by customers Y and Z.
00:00 [Jane] The goal today is to agree on a design solution. 00:12 [John] I think we should consider choice 1. 00:40 [Joe] Choice 2 has the advantage that it will take les...	Jane, John, and Joe decided to go with choice 2 for the design solution because it will take less time.

Test your prompt

Transcript	Summary
1 John Doe 00:00:01.415 --> 00:00:20.675	John and Jane are trying to replicate the results from the last analysis. They found out that the testing of the downstream classifier was done on the training data. They want to set up...

This self-explanatory slide describes the features available from the watsonx.ai prompt lab. Clients can work with the interactive prompt building in many ways.

A prompt is a way for clients to interact and instruct foundation models to provide answers. In generative AI, the model is not "answering" but is "generating" a response based on input and what it has been trained on. How one "prompts" the model can bring very different results. Prompt engineering is the "art" of prompting in ways that a model can best understand and respond to client requests.

The prompt lab is a part of the watsonx.ai user interface (UI) that allows users to perform prompt engineering, and experiment with different prompts across different foundation models.

- Experienced clients can select the model of interest and experiment with different prompts to see what works best for a particular model. Clients can try different "shots" with prompts:
  - Zero-shot prompting – simply passing a request into the model without additional information. Some models (larger ones) will respond better than others, depending on the request.
  - One-shot or few-shot prompting – clients can learn to provide simple input to orient the model properly to the user's intention to provide the best answer.
- Less experienced users can select different prompt samples from the prompt lab and learn how the examples utilize one-shot or few-shot tuning.
- Clients can set various parameters such as:

- Greedy vs Sampling – with ”greedy” the word/token with the highest probability is selected. Most large language models (LLMs) operate this way, it is the simplest form of next-word prediction. In sampling, the model selects a token using a random-weighted strategy across the probabilities of all tokens, resulting in the possibility of returning a less probable answer. This is used for creativity and to avoid repeating words but can cause the LLM to “wander” into other areas.

- Sampling parameters

  - Temperature – in general, the higher the temperature the higher the randomness. Setting the temperature to 0 means the model will always return exactly the same result. A higher temperature allows the model to consider other probabilities and cause hallucinations.

  - Top P – use this to limit random sampling to predictions whose combined probabilities do not exceed the threshold (value of P). A Top P value of 1 means everything is considered. A lower Top P value weeds out the lower probability options.

  - Top K – let the model select from the top K results only.

  - Repetition penalty – to prevent the model from generating repeating phrases

  - Stop sequences - tell the model to stop generating any more output once it encounters the specified stop sequence.

  - Min/max new tokens – the minimum or maximum new tokens used in the response – controlling how “verbose” the response should be.

Note that, unlike traditional AI tasks, generative AI creates a response best suited to its understanding of the question. This requires experience on the user’s part to learn how to effectively prompt the model as it needs to understand enough of the intent of the request. Using the prompt lab develops this experience.

The best way to learn how to pass good prompts is by example. The watsonx.ai prompt lab provides many examples of various natural language processing (NLP) tasks such as:

- Summarization
- Extraction
- Generation
- Classification
- And more …

Watsonx.ai prompt lab keeps a history of prompts. As part of learning, users may want to go back to previous prompts what might have worked better or to tweak them differently. The UI keeps track of past prompts and users can easily move back and forth.

Once the data scientist/engineer has found the best prompt, it can be passed in via an application programming interface (API) or Python notebook to the model. It can be used to automate tasks for other users.

While clients can prompt tune or fine-tune their models, that requires work and resources. Prompt engineering is a way to evolve a model for the needs of the clients without model rebuilding or a lot of labeled data.

watsonx.ai: Data Science and MLOps

Build machine learning models automatically in the studio

Model training and development

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

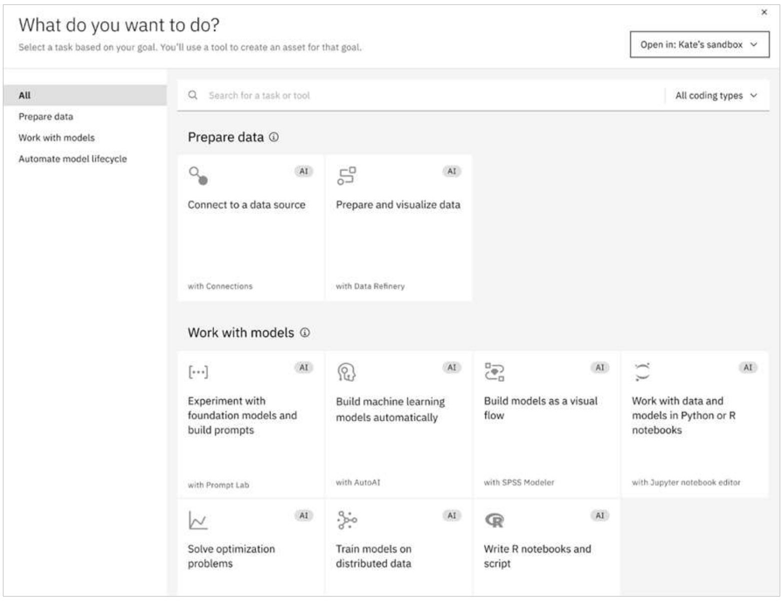
AutoAI, including preparing data for machine learning and generating and ranking candidate model pipelines

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

Integrated visual modeling

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

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



What do you want to do?

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

×
Open in: Kate's sandbox
All
Prepare data
Work with models
Automate model lifecycle
Search for a task or tool
All coding types
Prepare data
AI
Connect to a data source
with Connections
AI
Prepare and visualize data
with Data Refinery
Work with models
AI
Experiment with foundation models and build prompts
with Prompt Lab
AI
Build machine learning models automatically
with AutoAI
AI
Build models as a visual flow
with SPSS Modeler
AI
Work with data and models in Python or R notebooks
with Jupyter notebook editor
AI
Solve optimization problems
AI
Train models on distributed data
AI
Write R notebooks and script

This slide shows the AI operations available from watsonx.ai. This provides clients with familiar AI model build and operation available to apply machine learning operations (MLOps) against foundation models.

Watsonx.ai

- Model training and development

    - Clients can experiment with building with foundation models, and organize pipelines when tuning is necessary.
    - AutoAI – providing process automation – generating and ranking models candidate
    - Provide predictions for decision optimization. Clients can work with models in Python, natural language, or OPL stack (OpenAI, Pinecone, and Langchain).

- Integrated visual modeling

    - Visual modeling is an easy way to develop models and to analyze data by looking for patterns, and trends. This helps businesses to make the right decisions. Generative AI can help to extract/summarize information for comparison and provide natural language processing (NLP) to make the process even more end-user friendly.

Navigation

To get to the panel on the slide:

- On the main console of watsonx.ai, click on the appropriate project (typically named “<<user_name>>

sandbox”)

- On the subsequent window, click on the blue “New task” button on the upper right

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

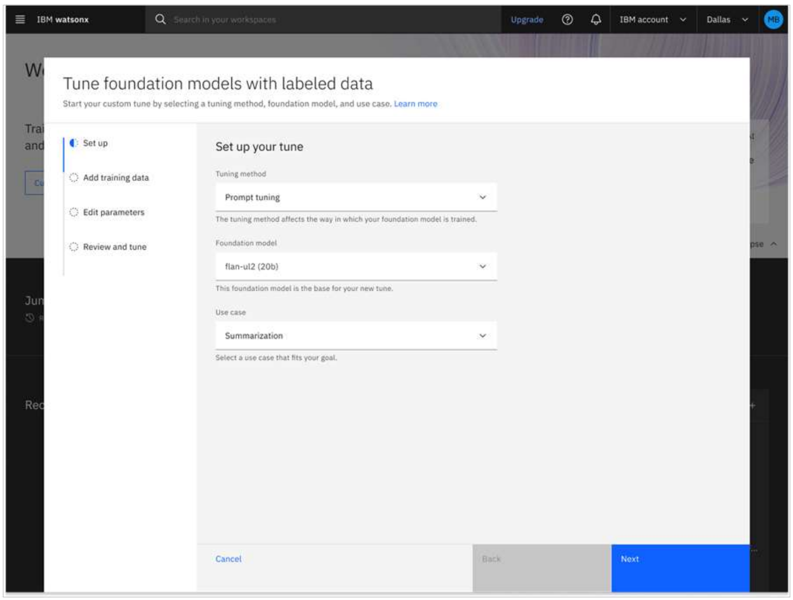
Unlike prompt engineering, prompt tuning allows clients to further enhance the model with focused, business data

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



Tune foundation models with labeled data

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

Set up your tune

Tuning method: Prompt tuning

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

Foundation model: flan-ul2 (20b)

This foundation model is the base for your new tune.

Use case: Summarization

Select a use case that fits your goal.

*Coming soon, available post-GA

The tuning studio (coming post-July 2023) allows clients to try out prompt-tuning and fine-tuning.

With prompt tuning, clients can choose a model and use their own specific data to enhance that model. While foundation models are very capable at a general level, businesses have their business terminologies, specific workflow, and specific output that they expect from queries. In prompt-tuning, clients can feed a set of <input query>-<expected output> pairs into the model. This teaches the model to better understand the intent of specific questions and how to generate the most relevant output.

As clients pass data into their model, there are also parameters they can modify in enhancing the model. Keep in mind that this (depending on the model) can be a very involved *and* expensive process. Clients can consider using larger batch_size (number of samples to work through before modifying any internal parameters) and smaller num_epochs (number of times to cycle through the sample data) to control training time/cost.

There are two important things to note with prompt tuning:

- No model weights are changed with the process. Clients are enhancing the model but keeping the original parameter weights intact.

- With the proper set of data, this can achieve near-fine-tuning results. This is important – while fine-tuning may give the best results, it is much more expensive as the model is updated (the bigger the model is originally, the more expensive it is). Prompt tuning also avoids catastrophic forgetting (i.e., the underlying model has so much of the weights modified it no longer resembles the original – this essentially destroys the reason *why* the base model was chosen in the first place).

The Watsonx.ai tuning studio will be released post-July 2023. Fine-tuning is in the roadmap for watsonx.ai –

although it is not available in the initial release. Fine-tuning is the next step of tuning where clients are providing (in most cases) a larger set of data and the underlying model will be modified.

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

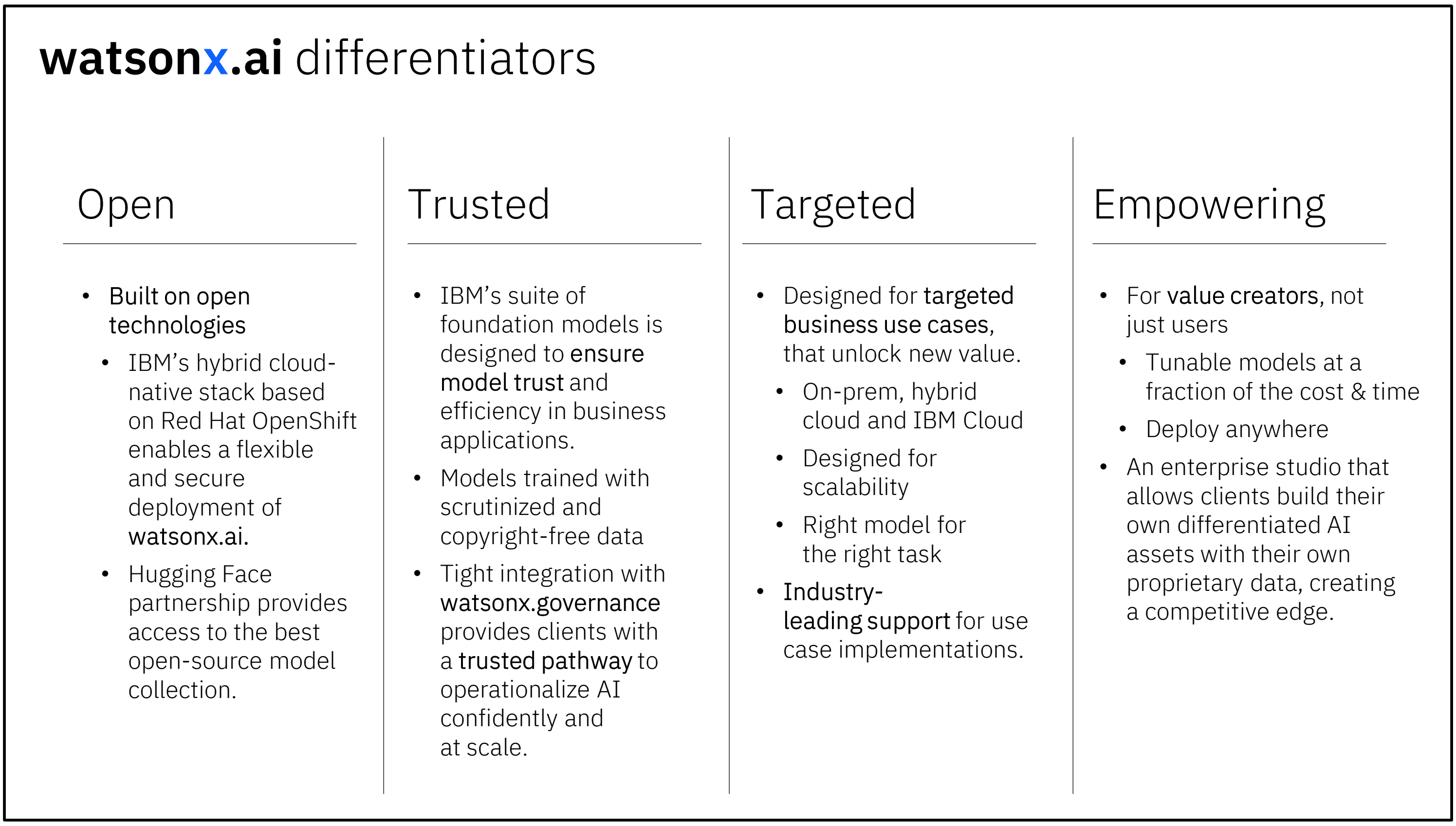
Trust is an enterprise's ultimate license to operate. The benefits of AI are moot if enterprises do not have confidence in the predictions and content generated by their models. Clients must (and they will want to) build responsibly, and transparently, and put governance into the heart of their AI lifecycle. Today, most AI models offered are trained on datasets of unknown quality and provenance. This can represent legal, regulatory, ethical, and inaccuracy nightmares. Data provenance and quality matter.

Besides bringing their own curated datasets to watsonx.data, clients will want to know that IBM is carefully curating domain-specific and internet datasets as a first step to training trustworthy models. IBM cleans those datasets and filters them for hate, profanity, biased language, licensing restrictions, and copyright information before training. And IBM continues to develop and refine new methods to improve data quality and controls. Leveraging watsonx.data capabilities, IBM tracks and manages every step of the process from data acquisition to cleansing, filtering, processing, and training that helps clients to react and meet an evolving set of legal and regulatory requirements.

In watsonx.governance, IBM tracks the curated data, the methods used to curate it, and the models that each data point has touched so if anything changes in the future, it is easier to identify affected models and any data that may need to be removed, remove it, retrain the models, and repeat the lifecycle. Simply put, watsonx.governance helps clients govern the training data clients use and the AI clients deploy so that they can operate, scale, and succeed with trust as their insurance.

IBM is developing and applying techniques in reinforcement learning from human feedback (RLHF) to align models with human values, reduce hallucinations, and build AI guardrails. Finally, for machine learning, IBM has developed methods to find and fix AI biases before AI models are tuned and deployed for enterprise tasks.

IBM's Center of Excellence (CoE) for generative AI—with its 1000's of consultants specialized in generative AI—is not only ready to actively build and deploy watsonx for clients, helping them implement enterprise-grade foundation models and generative AI in their operations but also helps clients operationalize AI governance. IBM's CoE for generative AI can bring solutions and services that help clients establish the organizational structure to adopt and scale AI safely, detect and mitigate risks, and provide education and guidance while developing new solutions and assets that leverage the pipeline of innovations in watsonx. The CoE's consultants have also developed an open, collaborative approach to quickly ideate and prioritize use cases, plan, build, implement, operate, and scale ethically responsible generative AI solutions, leveraging multiple models on multiple clouds to best meet clients'' unique business needs.



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.

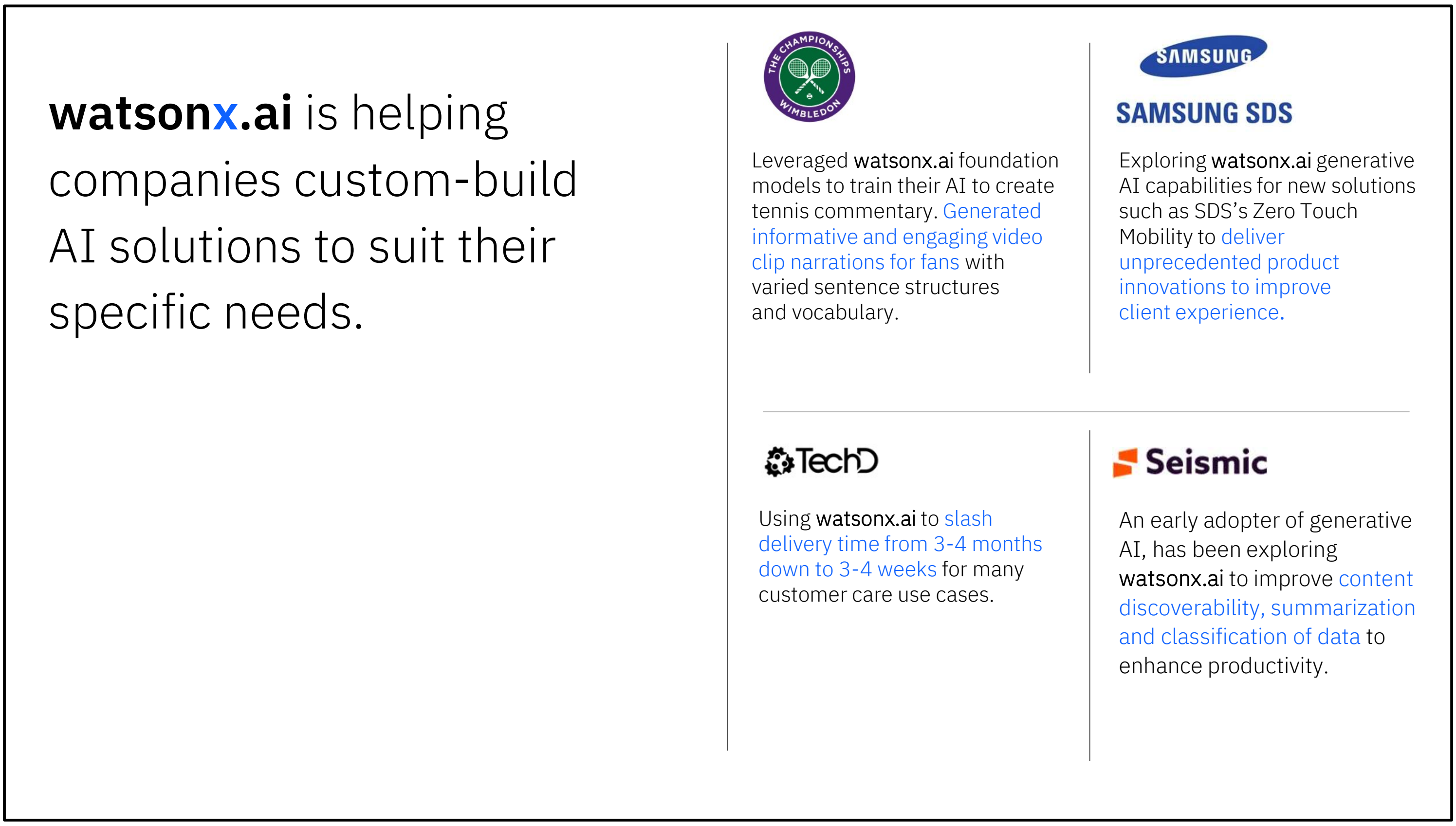
This self-explanatory slide revisits an earlier slide on the key factors that drive a successful generative AI platform and shows the watsonx.ai differentiators. To recap that slide, this slide shows some more details clients should know about watsonx.ai.

- Open
  - IBM's watsonx.ai is a hybrid cloud-native stack built on Red Hat OpenShift. This allows watsonx to integrate easily with other IBM cloud services such as IBM Watson Studio, IBM Watson Knowledge Catalog, IBM Watson Query, etc. It also means watsonx.ai is hybrid and multicloud enabled. IBM's partnership with Hugging Face provides clients with access to the best collection of open-source models available.
- Trusted
  - IBM has a long history of building secure AI and data platforms. This focus is carried over to watsonx.ai to ensure IBM's suite of foundation models is built on the principle of trust, making them ready for business applications. As a start, IBM's models are trained on highly curated, scrutinized, and copyright-free data that was also filtered to weed out inappropriate content. The integration with watsonx.governance provides a trusted pathway to operationalize AI confidently and at scale. Watsonx.ai is an AI platform that lets clients run responsible, transparent, and explainable AI.
- Targeted
  - Watsonx.ai is designed for targeted business use cases to unlock new value for customers. Specifically:
    - It can be on-premises, hybrid cloud, IBM cloud, or on the edge.
    - It is designed to be scalable – matching the needs of clients

    - Supports various models of different sizes and different architectures (encoder, decoder, and encoder-decoder) and domain-specific models (finance, cybersecurity, etc.) will be available for clients to pick the best model to use or tune for their business use cases. Industry-leading support to help clients to implement their use cases.

- Empowered

  - Watsonx.ai is designed for AI value creators, not just users. Models can be tuned at a fraction of the cost (clients pick the right size, architecture, and domain) and they can be deployed anywhere. What's more, the Watsonx.ai console allows clients to build, tune, and deploy their models. Clients can use their own data to create differentiated AI assets.



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

THE CHAMPIONSHIPS WIMBLEDON

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

SAMSUNG

SAMSUNG SDS

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

TechD

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

Seismic

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

This slide is self-explanatory and provides a number of client responses with the **watsonx.ai** early access program

Sources:

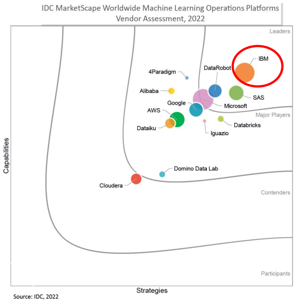
- Wimbledon: https://newsroom.ibm.com/2023-06-21-IBM-Brings-Generative-AI-Commentary-and-AI-Draw-Analysis-to-the-Wimbledon-Digital-Experience

- Information for other clients is from the early access program.

Acronym:

- SDS: Samsung Data Service

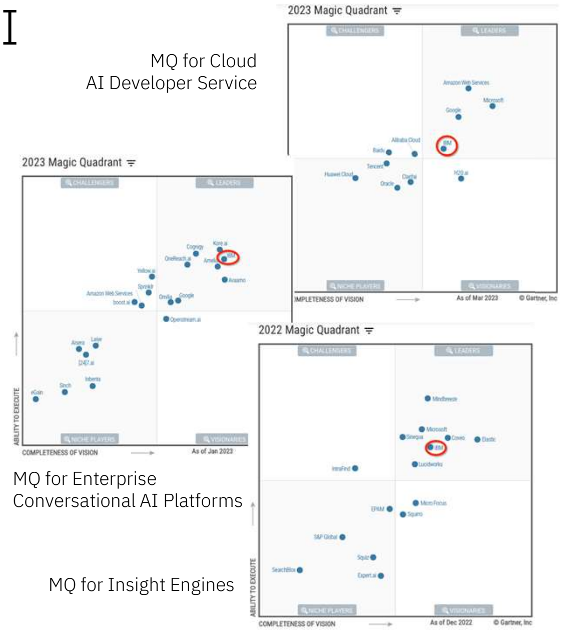
IBM is a leader in AI



Source: IDC, 2022

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

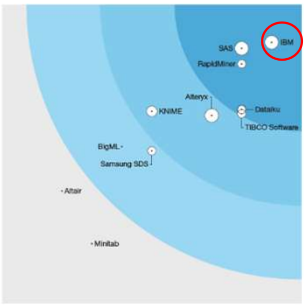
MQ for Cloud
AI Developer Service



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

IBM is recognized as a leader in artificial intelligence (AI) in the marketplace according to International Data Corporation (IDC), Gartner, Inc., and Forrester Research.

The Gartner Magic Quadrant (MQ) for AI includes the following:

- MQ for Cloud AI Developer Services
- MQ for Insight Engines
- MQ for Conversational AI

Sources:

- IDC Marketscape: Leader in Worldwide Machine Learning Operations Platforms 2022 Vendor Assessment: https://www.ibm.com/account/reg/us-en/signup?formid=urx-51950

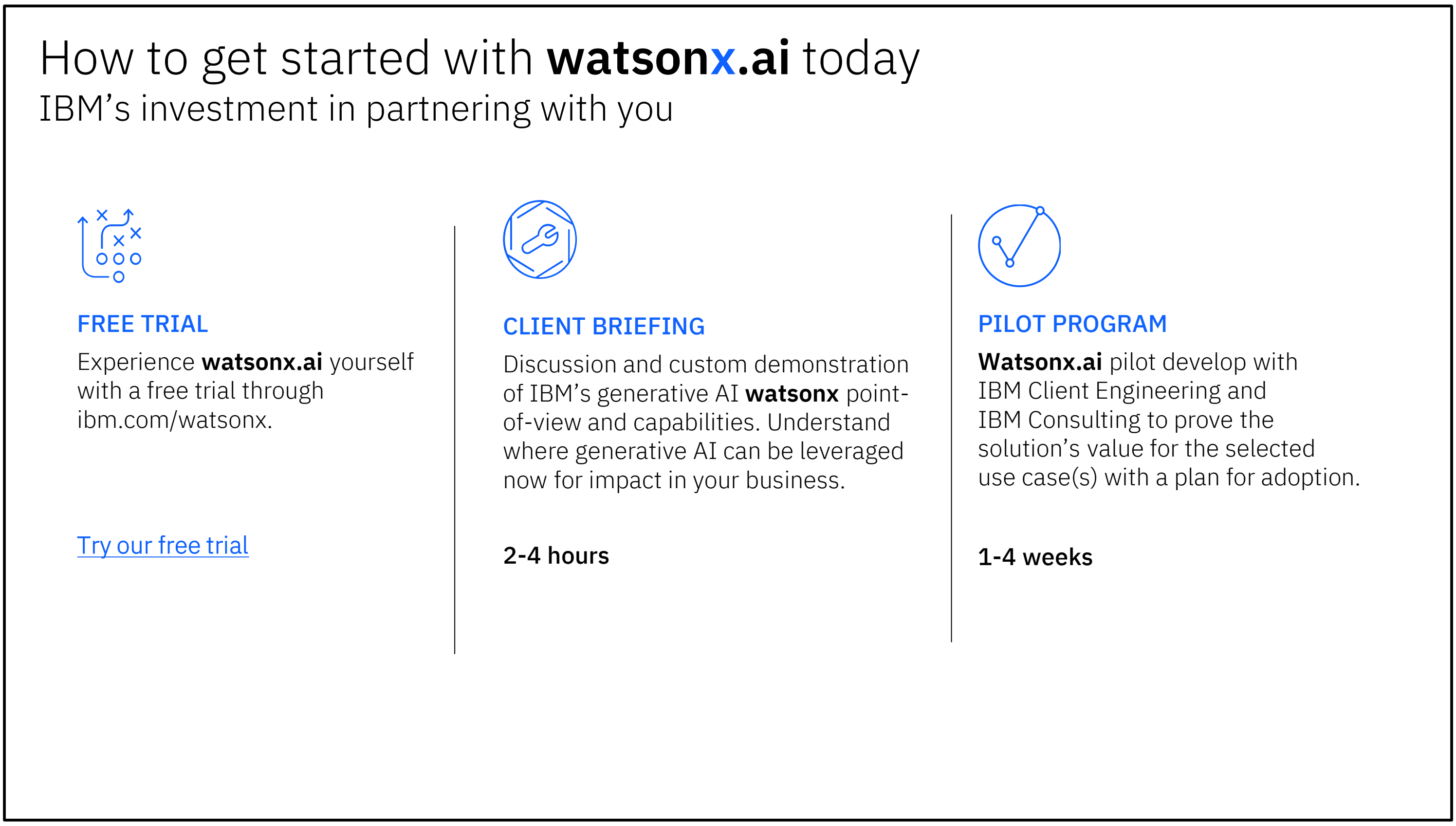
- Forrester Wave 2023: Data Management For Analytics: https://www.ibm.com/blog/ibm-named-a-leader-in-the-forrester-wave-data-management-for-analytics-q1-2023

- Forrester TEI for IBM Data Management: https://www.ibm.com/downloads/cas/9X4AX5WK

- Gartner MQ for Cloud AI Developer Services: https://www.gartner.com/document/4372099?ref=solrAll&refval=378158450&

- Gartner MQ for Conversational AI: https://www.gartner.com/document/4154599?ref=solrAll&refval=367313557

- Gartner MQ for Insight Engines: https://www.gartner.com/document/4022148?ref=solrAll&refval=367313691



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

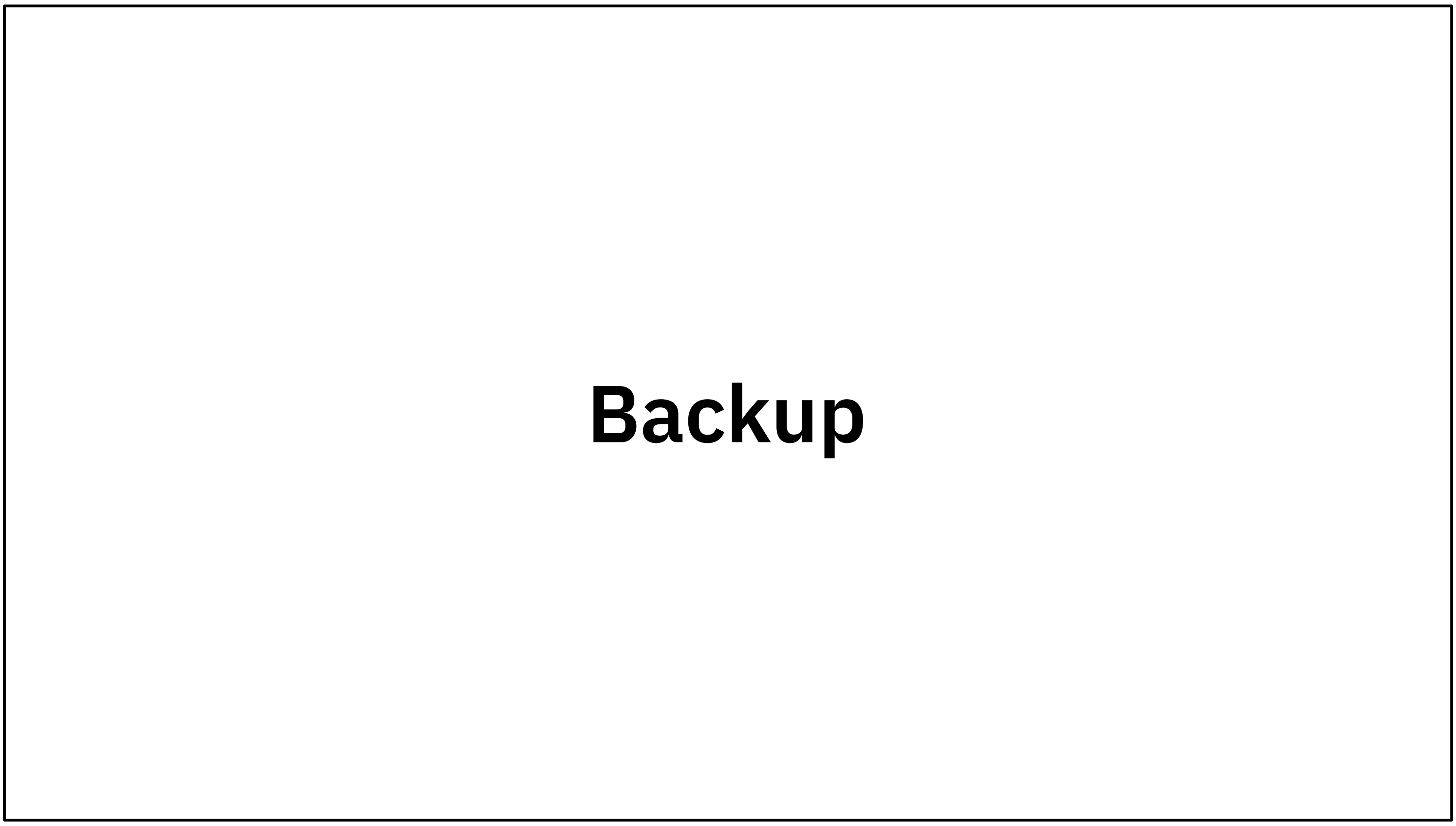
PILOT PROGRAM

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

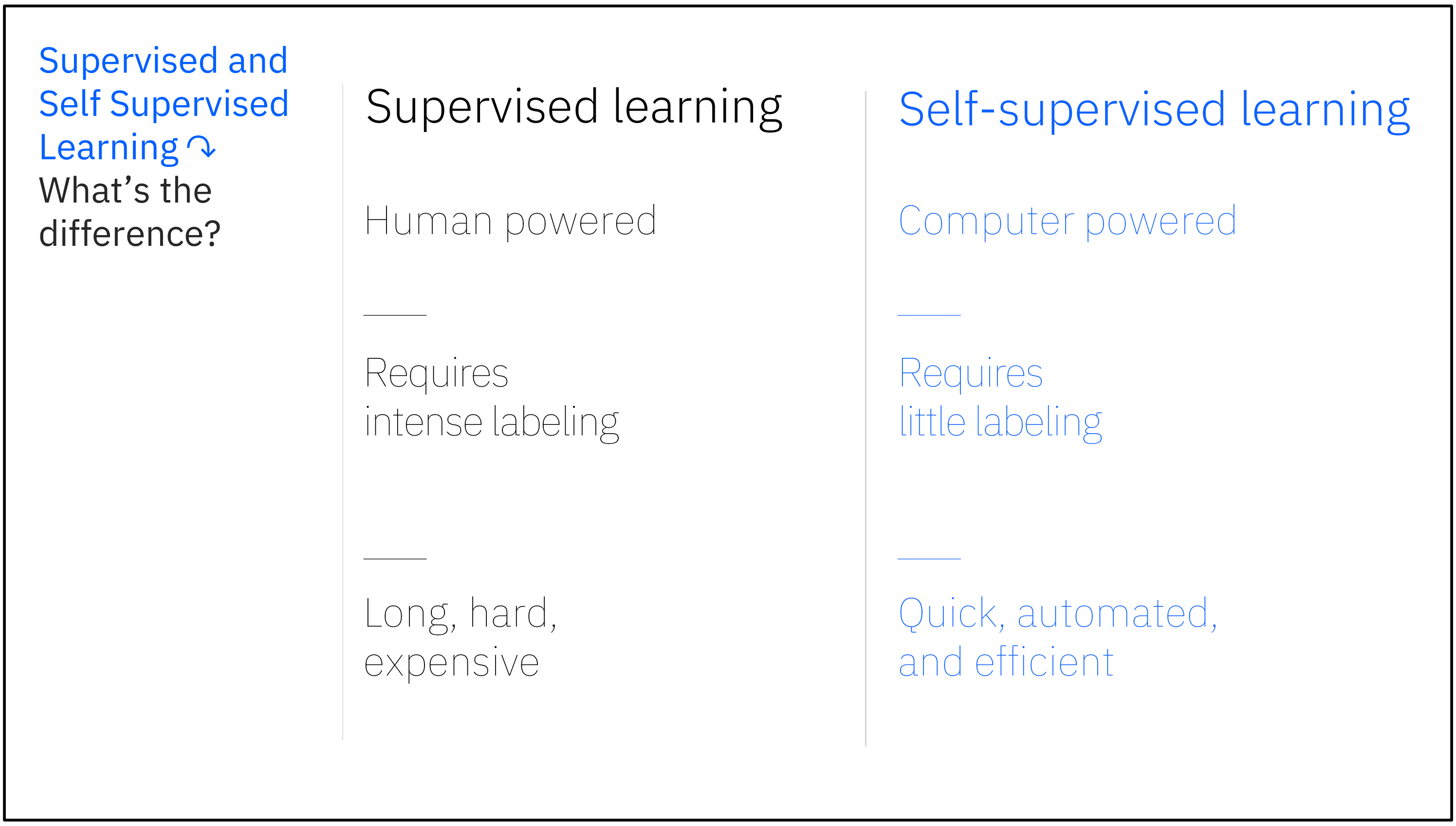
This self-explanatory slide provides information on how clients can get started with the watsonx platform and watsonx.ai. Clients can try out the platform for free, request a client briefing, and join the pilot program.

Pilot program details can be found here: https://ibm.biz/watsonxPilotProgram



Backup

This section contains backup slides.



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

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

The breakthrough with large language models (LLMs) is the ability for the AI to do self-supervised learning and **this changes the game when it comes to AI.**

This slide shows the difference between supervised learning (the slow and expensive way of yesteryear's AI) and self-supervised learning (the breakthrough that's behind foundation models (FM)).

Supervised learning

With supervised learning, humans are in the loop... labeling data. For example, if someone was building a speech translation AI that could translate English to French, that labeled data would include English sentences and their French equivalents (referred to as 'pairs') that a human would verify was true. An insurance adjuster would assert that a broken headlight picture was indeed a broken headlight, or a cat was a cat, and so on. This is the most trusted form of data (since a human labelled it). But it's also very expensive to do and labor intensive; and since AI needs lots of data to build robust predictions, getting accurate labeled data really steepens the time to value curve. To help in this area, various companies have created out-sourced labeling (like Amazon's Mechanical Turk) that would get people to do these tasks at a lower cost... but this still costs money and takes time. The bottom of the supervised learning column perfectly summarizes this process: **long, hard, and expensive. (Note: The hard part isn't really training the AI, it's getting all the labeled data.)**

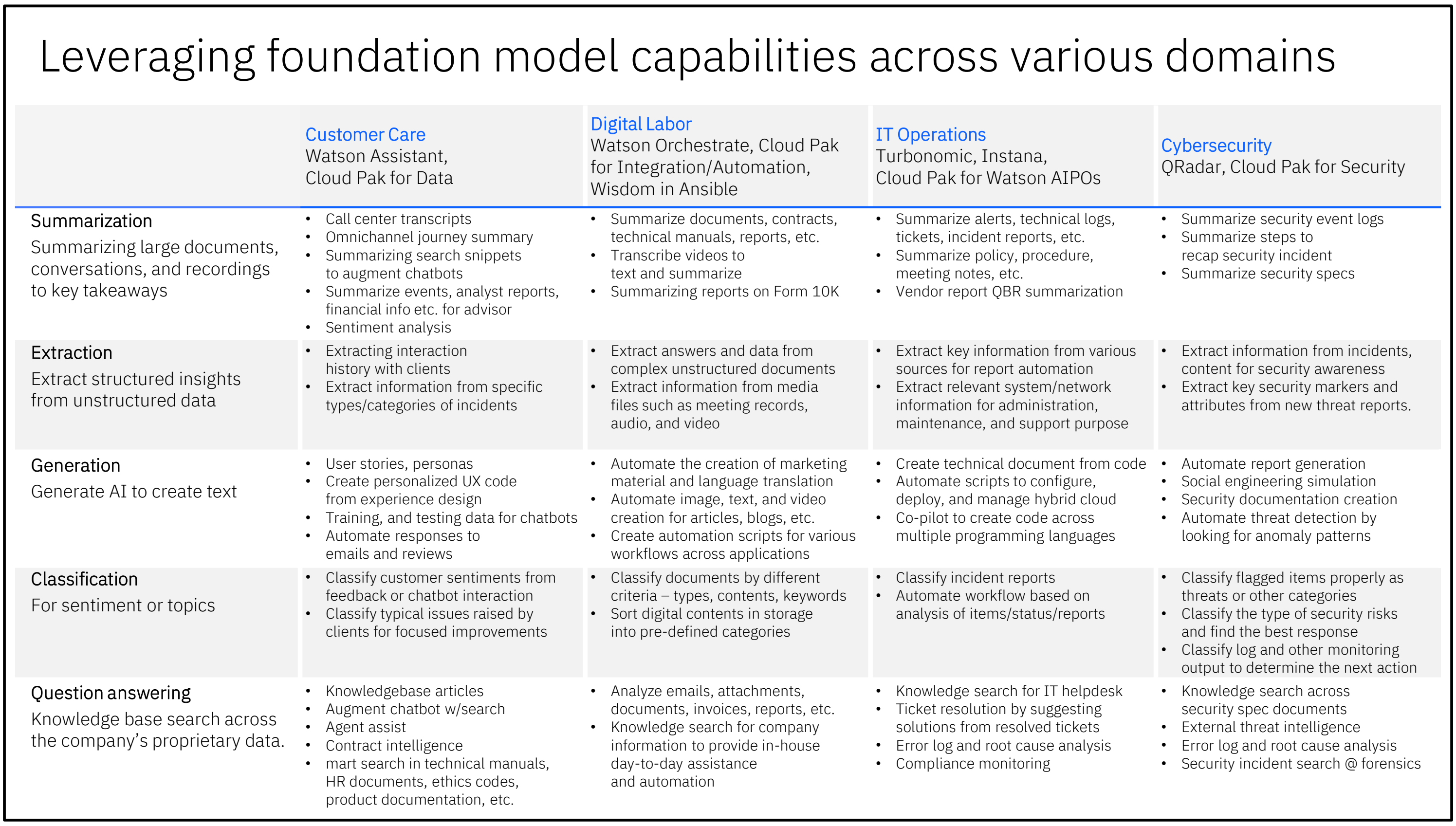
Self-supervised learning

FMs used self-supervised learning. In this case, the labels are machine generated and machine is making the decisions about the labels. In this case, AI is used to make decisions about data (so AI is helping create new AI).

This process lets a practitioner take large amounts of data without a whole lot of pre-processing work (they still must collect the required data and deal with sparsity and such, but it requires very little, or next to no labeling)... it's all AI-computer powered! There's obviously a human in the loop as it refers to the overall process (training the model, getting the data, testing accuracy and bias, and so on), but the human is pretty much out of the loop on the labeling part.

Obviously, self-supervised learning makes the whole process much more efficient, reduces cost (it's still expensive to train these models because of the sheer mass amount of data and compute that's required), and accelerates time to market (because humans aren't doing the labor-intensive task of labeling in the pre-processing phase).

It's easy to see how self-supervised learning opens-up the world of AI to more people (kind of .... the true value will be in the fine-tuning phase that most clients will likely do on top of these models).



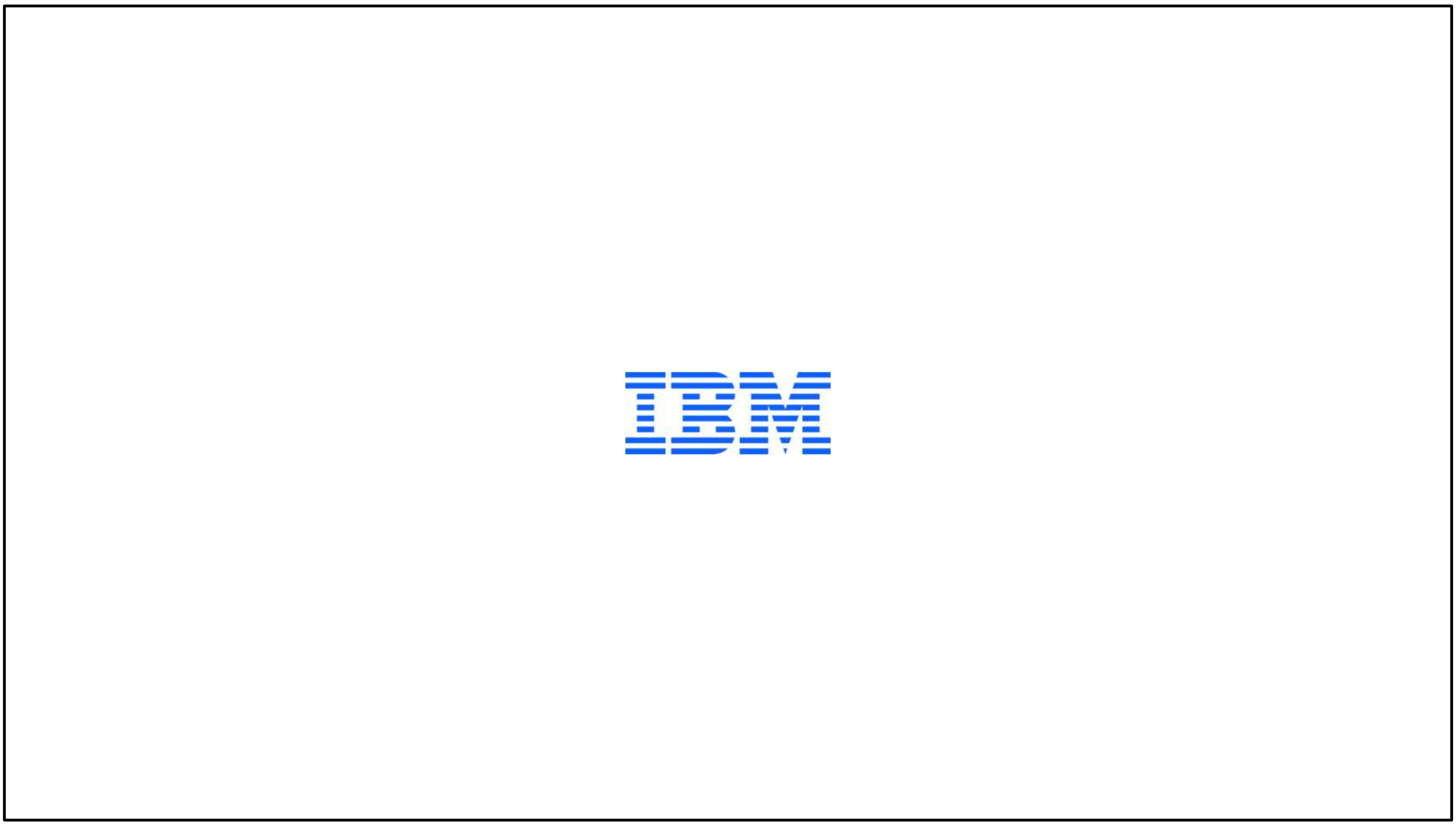
Leveraging foundation model capabilities across various domains

	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

This self-explanatory slide details how clients can leverage foundation models (and generative AI) to support these five popular use cases:

- Summarization
- Extraction
- Generate
- Classify
- Question answering

Use cases are described for 4 business areas: customer care, digital labor, IT operations, and cybersecurity. Of course, there are other opportunities not detailed on this slide.



IBM

This self-explanatory slide does not require speaker notes.