# ETSI AI Conference 2024

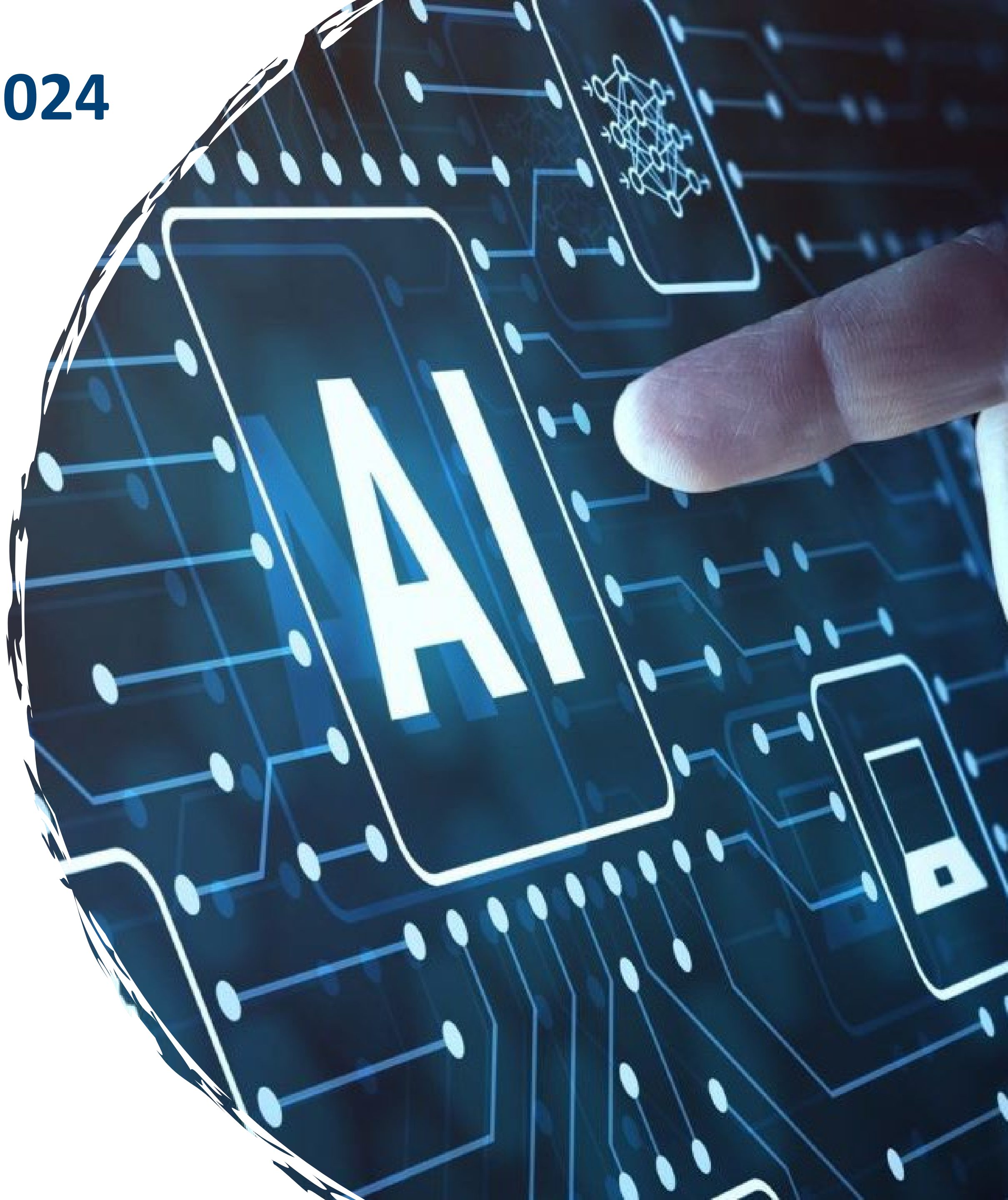
ETSI The Standards People

# Generative AI – Can that be done responsibly?

Presented by:

IBM

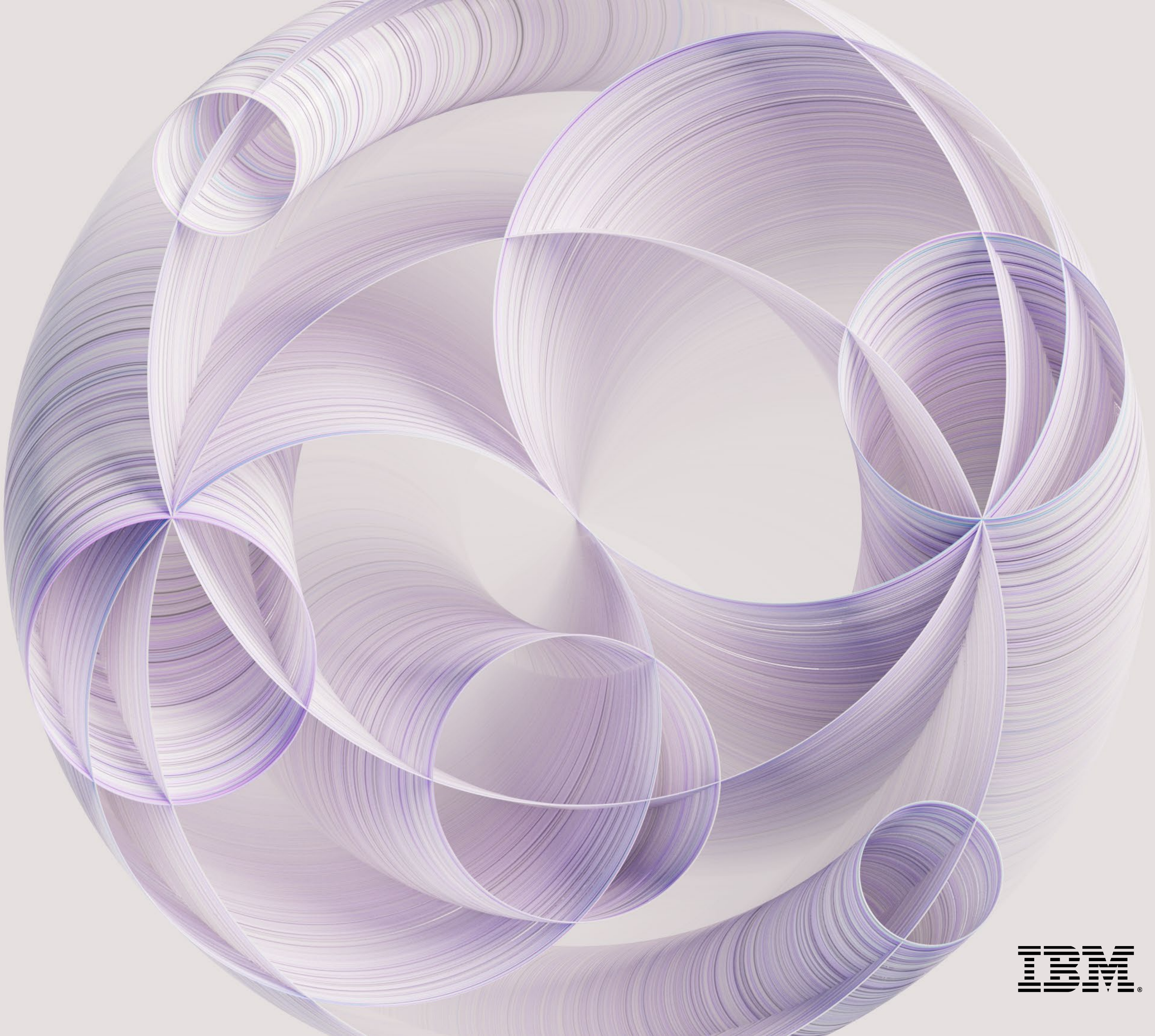
5th February, 2024



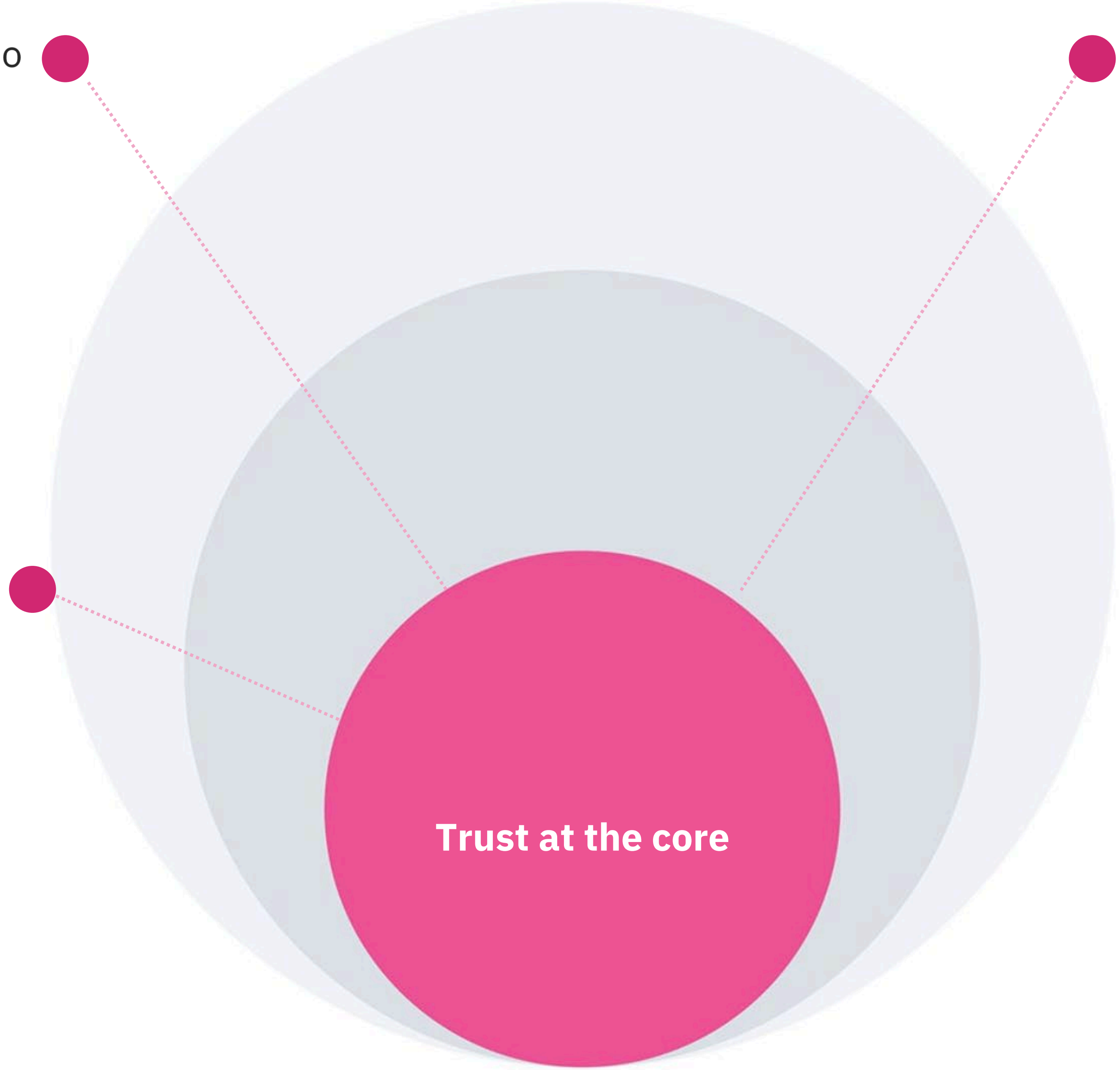
# Governance of AI

**Hans Petter Dalen**
Business Leader EMEA

watsonx.governance



**Watsonx.governance**



Trust at the core

The purpose of AI is to augment human intelligence

Technology must be transparent and explainable

Data and insights belong to their creator

# Key findings from IBM Institute for Business Value CEO Study October 2023

Customer service has leapfrogged other functions to become CEOs' #1 generative AI priority.

85% of execs say generative AI will be interacting directly with customers in the next two years.

Generative AI is closing the gap between IT and the business—finally.

Generative AI opens the door to reinventing employee experience.

AI and data governance are board-level concerns.

Generative AI is about people and skills, not just technology.

As AI gets smarter, employees can do higher value work.

# Elements of AI risk

Accountability

Accuracy

Fairness

Veracity

Transparency

Drift

Trusted data

Energy consumption

Explainability

Adversarial Robustness

IP/PII leakage

...

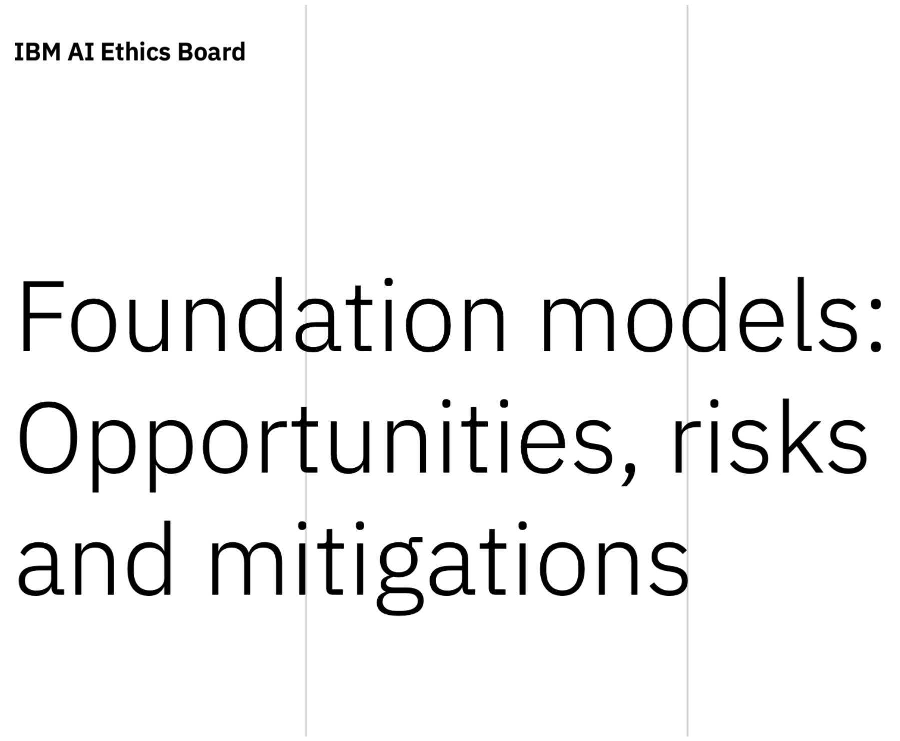
Regulatory Risk

Reputational Risk

Operational Risk

# IBM AI Ethics Board – White Paper on risk

https://www.ibm.com/downloads/cas/E5KE5KRZ

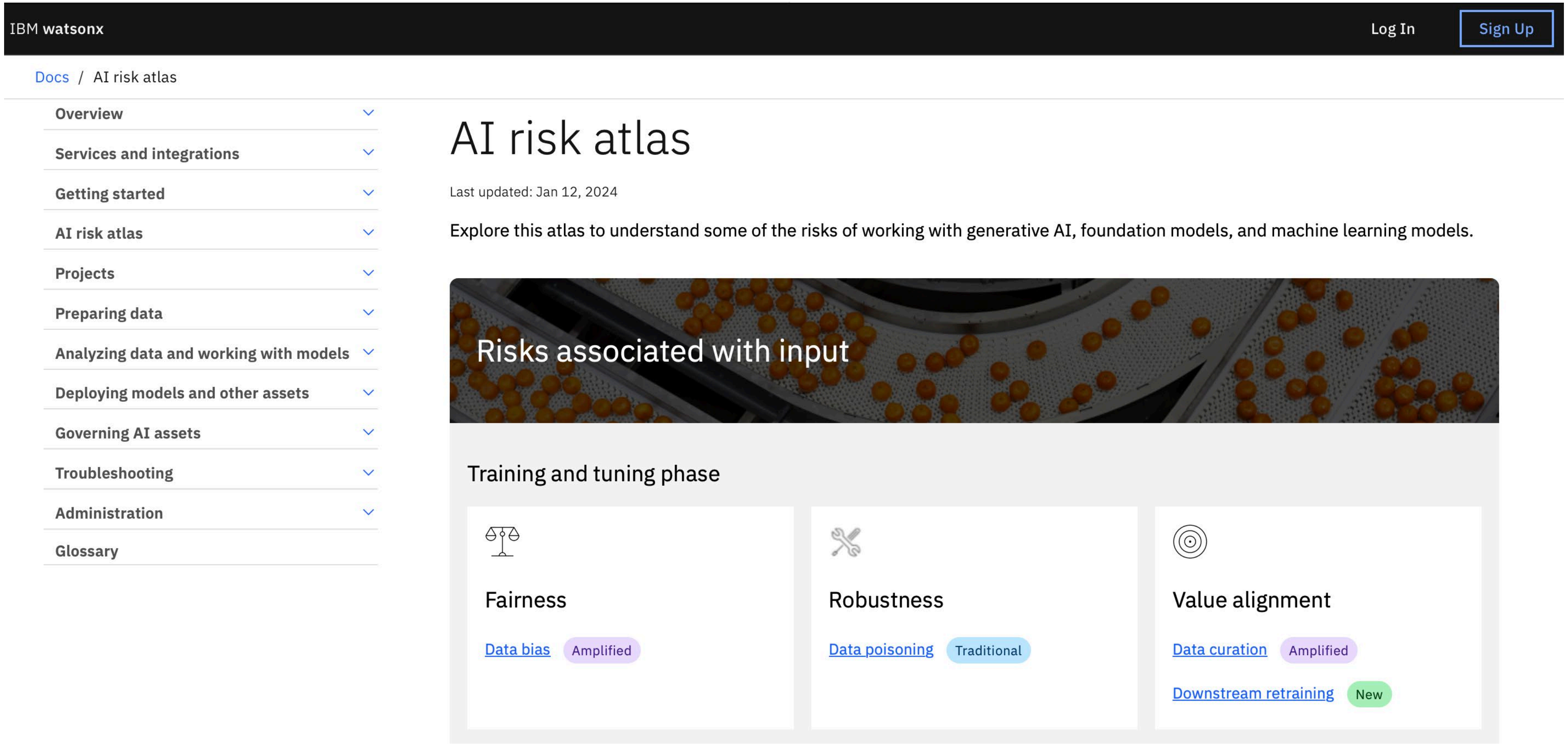


IBM AI Ethics Board

Foundation models: Opportunities, risks and mitigations

# IBM AI Risk Atlas

https://dataplatform.cloud.ibm.com/docs/content/wsj/ai-risk-atlas/ai-risk-atlas.html?context=wx&audience=wdp



IBM **watsonx** Log In Sign Up

Docs / AI risk atlas

- Overview
- Services and integrations
- Getting started
- AI risk atlas
- Projects
- Preparing data
- Analyzing data and working with models
- Deploying models and other assets
- Governing AI assets
- Troubleshooting
- Administration
- Glossary

## AI risk atlas

Last updated: Jan 12, 2024

Explore this atlas to understand some of the risks of working with generative AI, foundation models, and machine learning models.

### Risks associated with input

#### Training and tuning phase

**Fairness**

Data bias Amplified

**Robustness**

Data poisoning Traditional

**Value alignment**

Data curation Amplified

Downstream retraining New

# General Purpose AI Systems (Generative AI)

Foundation model
- OpenAI GPT
- IBM Granite
- Meta Llama 2
- Any of the 300.000+ FMs on Hugging Face

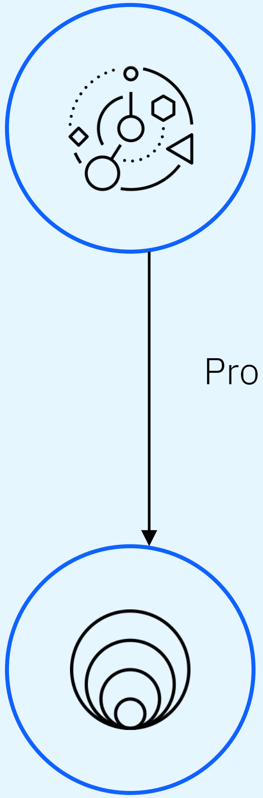
# General Purpose AI Systems (Generative AI)

Generative AI application
- OpenAI ChatGPT
- Microsoft Copilot
- Purpose made for use case

Foundation model
- OpenAI GPT
- IBM Granite
- Meta Llama 2
- Any of the 300.000+ FMs on Hugging Face

# General Purpose AI Systems (Generative AI)

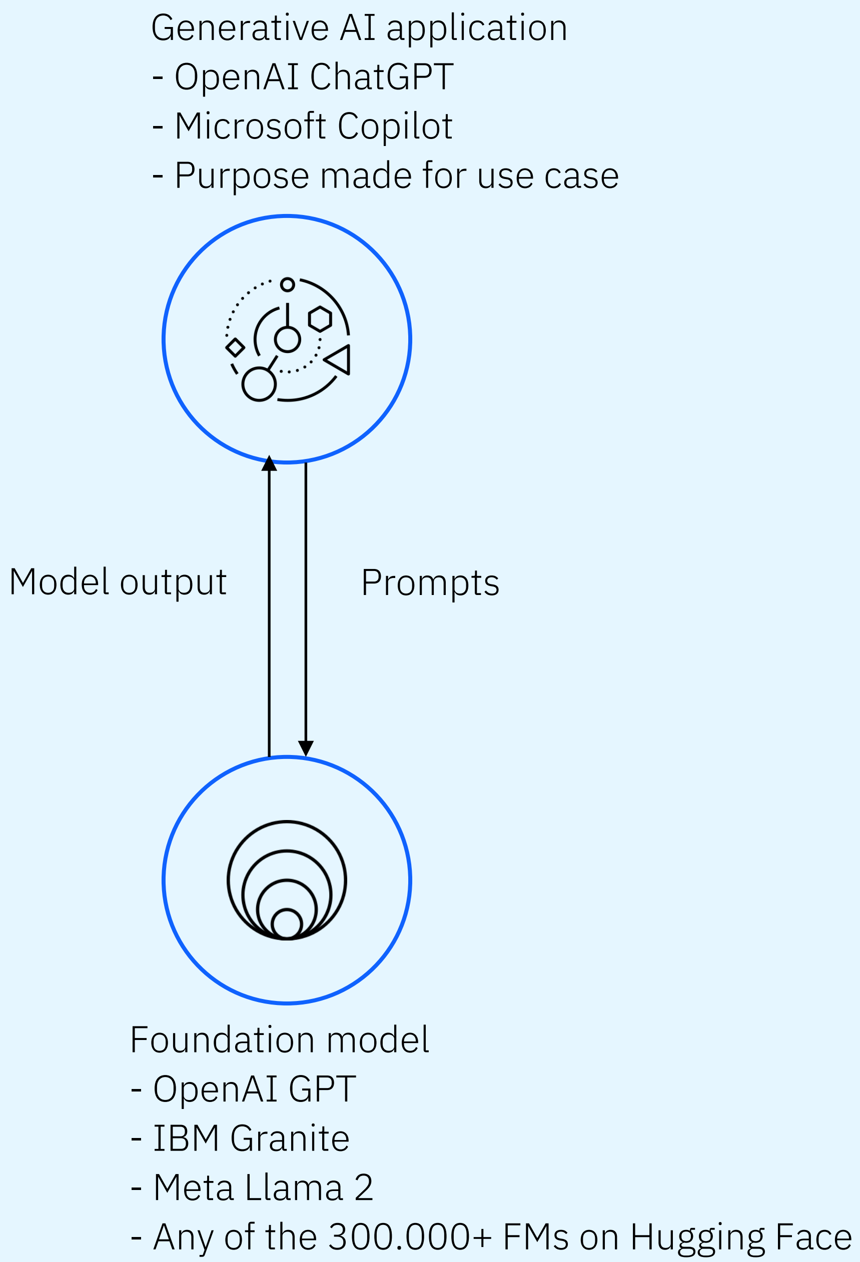
Generative AI application
- OpenAI ChatGPT
- Microsoft Copilot
- Purpose made for use case



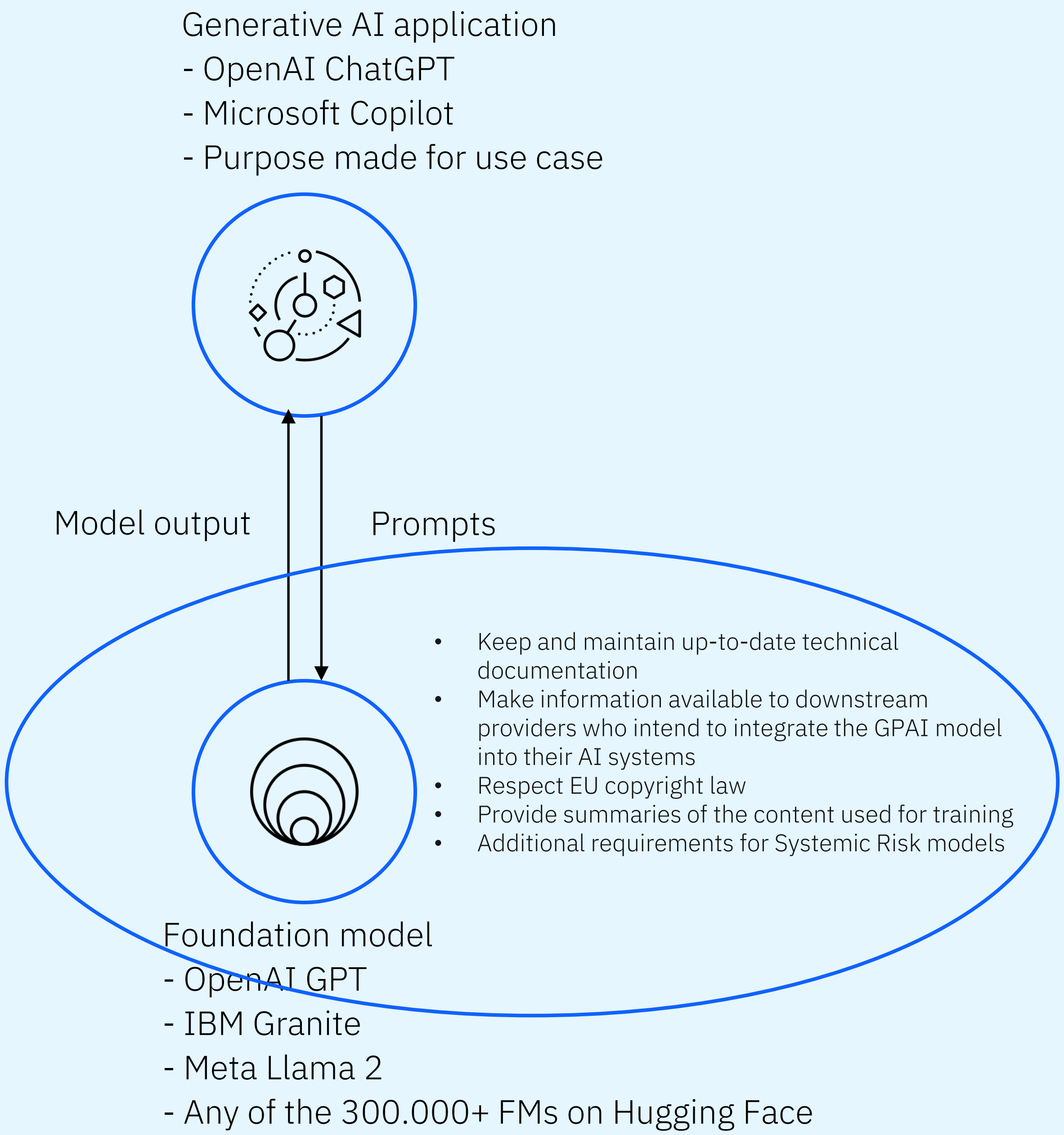
Prompts

Foundation model
- OpenAI GPT
- IBM Granite
- Meta Llama 2
- Any of the 300.000+ FMs on Hugging Face

# General Purpose AI Systems (Generative AI)



# General Purpose AI Systems (Generative AI)



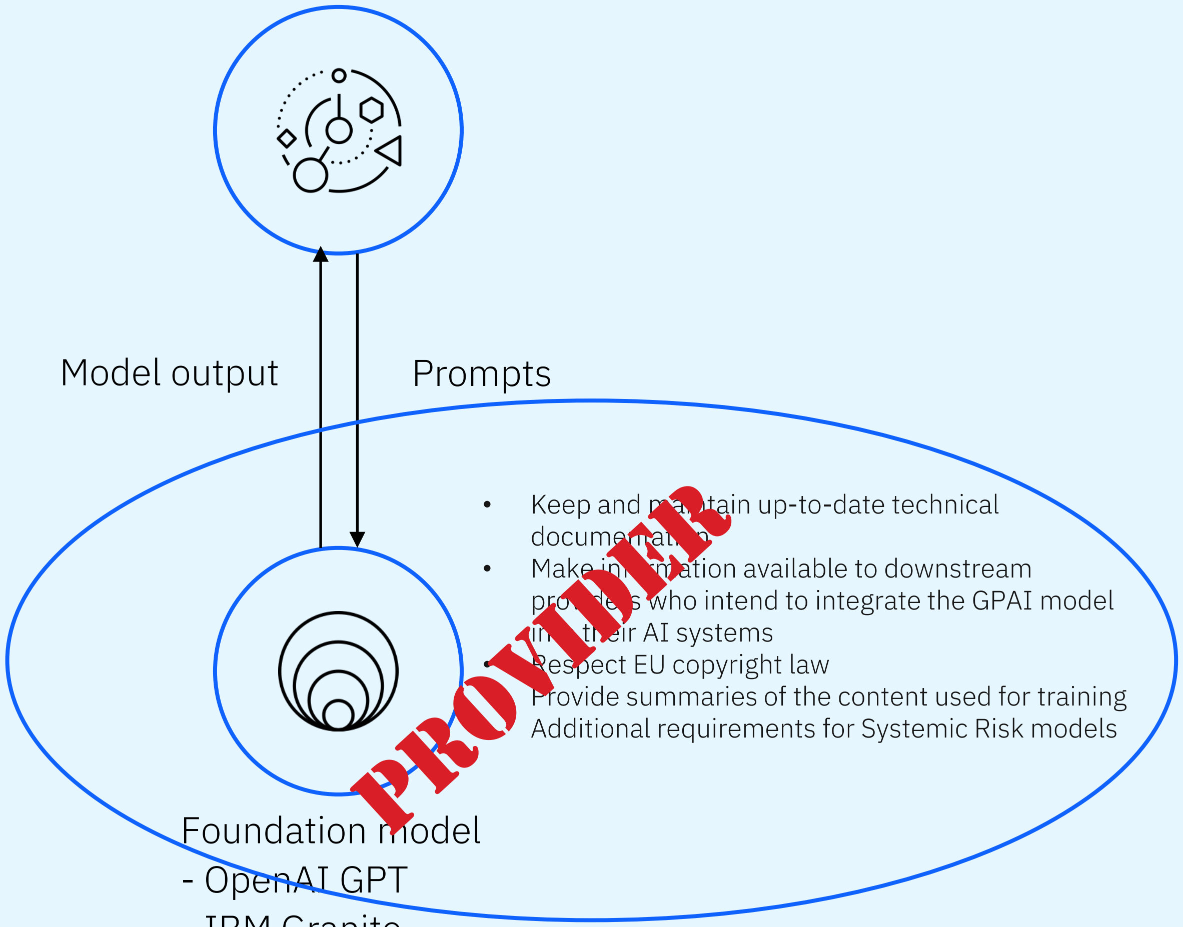
Generative AI application
- OpenAI ChatGPT
- Microsoft Copilot
- Purpose made for use case

- Keep and maintain up-to-date technical documentation
- Make information available to downstream providers who intend to integrate the GPAI model into their AI systems
- Respect EU copyright law
- Provide summaries of the content used for training
- Additional requirements for Systemic Risk models

Foundation model
- OpenAI GPT
- IBM Granite
- Meta Llama 2
- Any of the 300.000+ FMs on Hugging Face

# General Purpose AI Systems (Generative AI)

Generative AI application
- OpenAI ChatGPT
- Microsoft Copilot
- Purpose made for use case



Model output

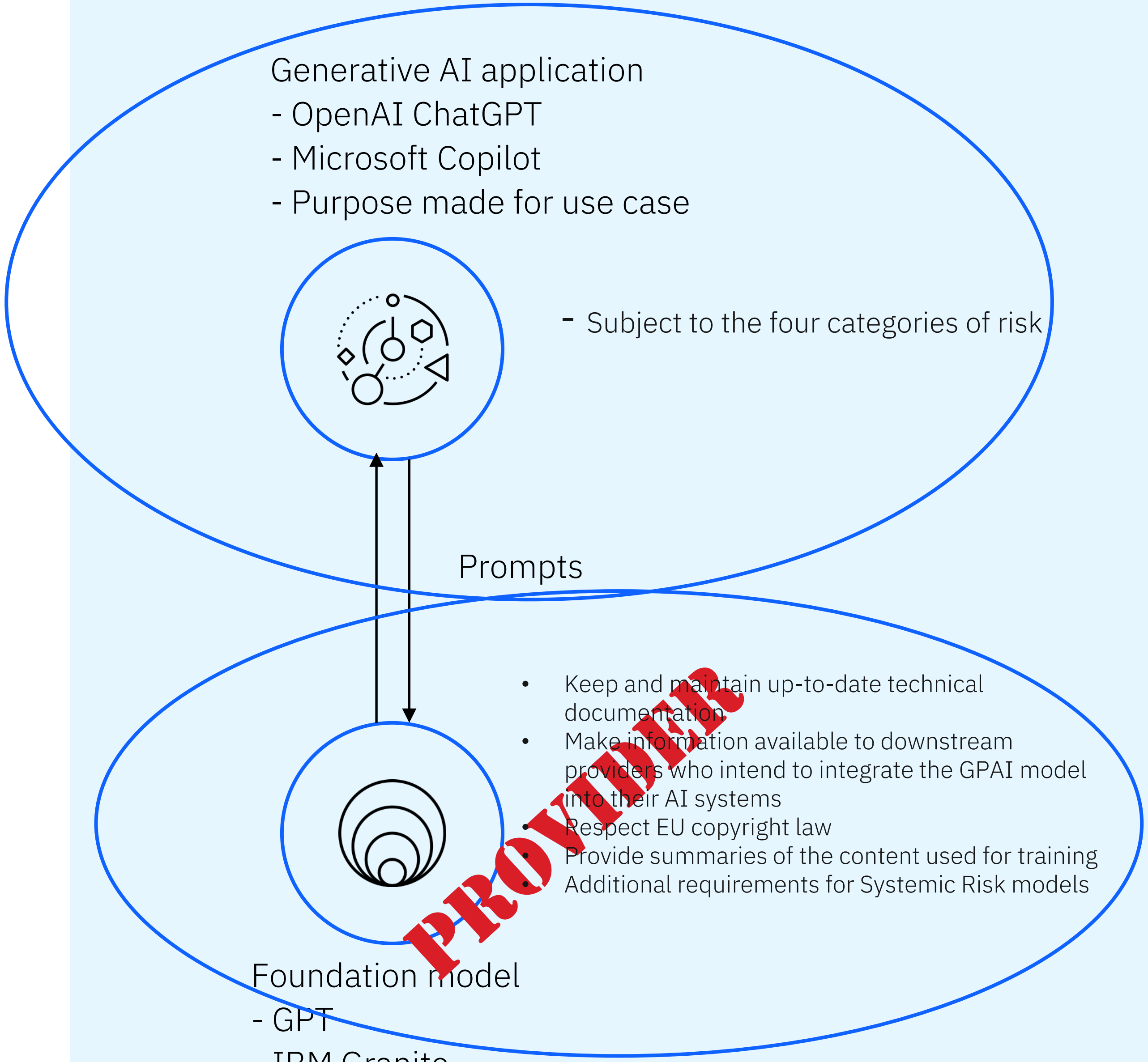
Prompts

- Keep and maintain up-to-date technical documentation
- Make information available to downstream providers who intend to integrate the GPAI model into their AI systems
- Respect EU copyright law
- Provide summaries of the content used for training
- Additional requirements for Systemic Risk models

PROVIDER

Foundation model
- OpenAI GPT
- IBM Granite
- Meta Llama 2
- Any of the 300.000+ FMs on Hugging Face

# General Purpose AI Systems (Generative AI)



Generative AI application
- OpenAI ChatGPT
- Microsoft Copilot
- Purpose made for use case

- Subject to the four categories of risk

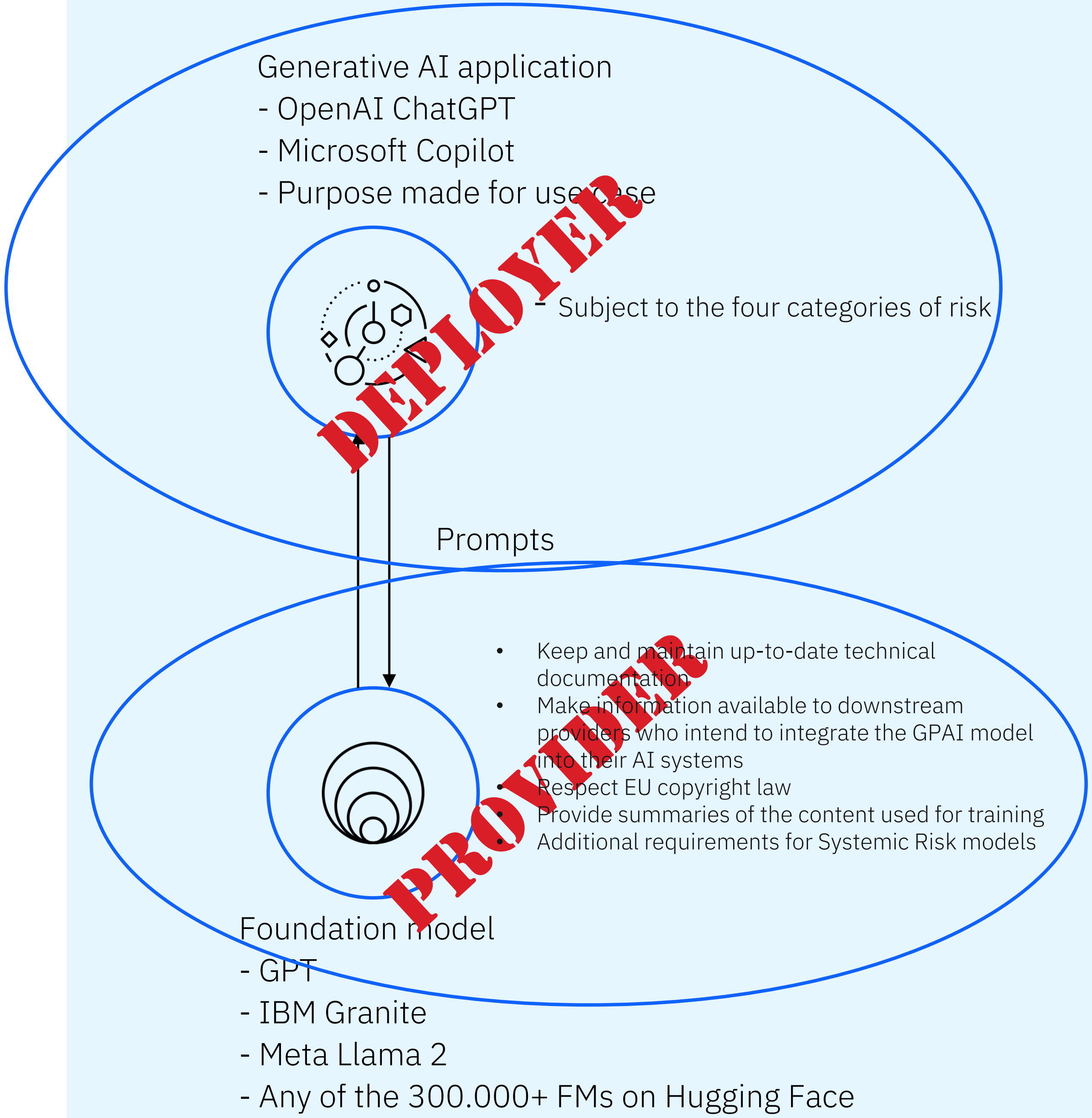
Prompts

- Keep and maintain up-to-date technical documentation
- Make information available to downstream providers who intend to integrate the GPAI model into their AI systems
- Respect EU copyright law
- Provide summaries of the content used for training
- Additional requirements for Systemic Risk models

PROVIDER

Foundation model
- GPT
- IBM Granite
- Meta Llama 2
- Any of the 300.000+ FMs on Hugging Face

# General Purpose AI Systems (Generative AI)



Generative AI application
- OpenAI ChatGPT
- Microsoft Copilot
- Purpose made for use case

- Subject to the four categories of risk

Prompts

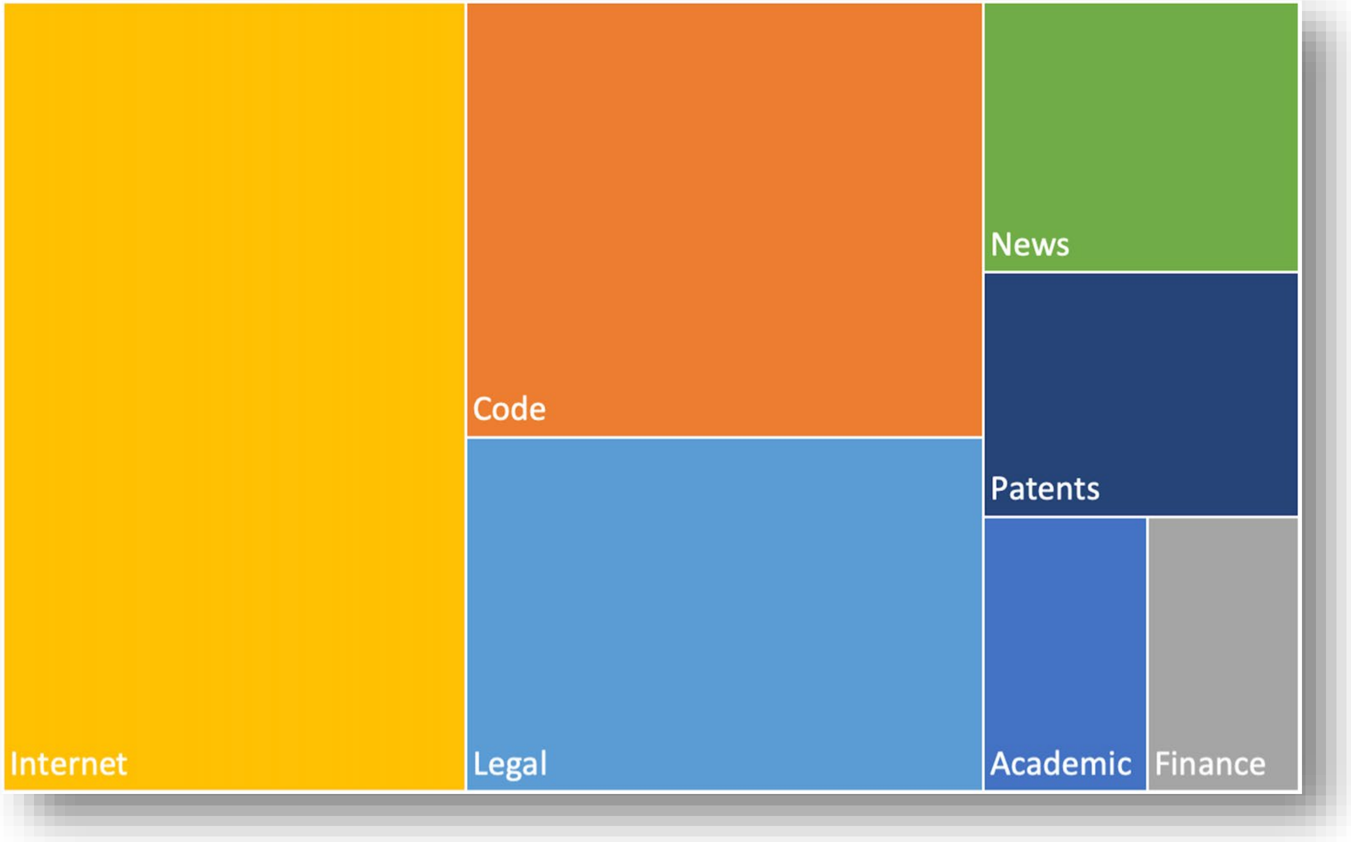
- Keep and maintain up-to-date technical documentation
- Make information available to downstream providers who intend to integrate the GPAI model into their AI systems
- Respect EU copyright law
- Provide summaries of the content used for training
- Additional requirements for Systemic Risk models

Foundation model
- GPT
- IBM Granite
- Meta Llama 2
- Any of the 300.000+ FMs on Hugging Face

# **IBM Data Pile**: 1+ Trillion Tokens used for **watsonx.ai** training

Crawled and downloaded petabytes of raw data, including broad coverage of a range of different kinds of natural language and code.

Data preprocessed to remove hate, profanity, duplicates and filtered for content crawled from questionable websites (URL-blocklisting), resulting in a curated multi-terabyte dataset for model training.



# Indemnification

“IBM provides its standard contractual intellectual property indemnification for IBM-developed foundation models, similar to those it provides for IBM hardware and software products

IBM also offers a peace of mind to clients by

- not requiring them to indemnify IBM for their use of its models
- not capping its IP indemnification liability.

The current watsonx models now under these protections include:

- *Slate* family of encoder-only models
- *Granite* family of a decoder-only models”

# AI governance

The process of creating *policies*, assigning *decision rights* and ensuring *organizational accountability* for risks and investment decisions for the application and use of artificial intelligence techniques [...], addressing the *predictive and generative* nature of AI.

## Why This Is Important

With AI now delivering value in the enterprise, data and analytics leaders observe that **scaling AI without governance is ineffective and dangerous**.

Generative AI and applications, like OpenAI's ChatGPT, make AI governance a necessity, as using pretrained AI models billions of times sharpens risk concerns. The leaders want to *balance* AI's business value and the need for appropriate oversight.

AI draws the *attention of legislators worldwide*, who mandate actions by clarifying AI governance priorities.

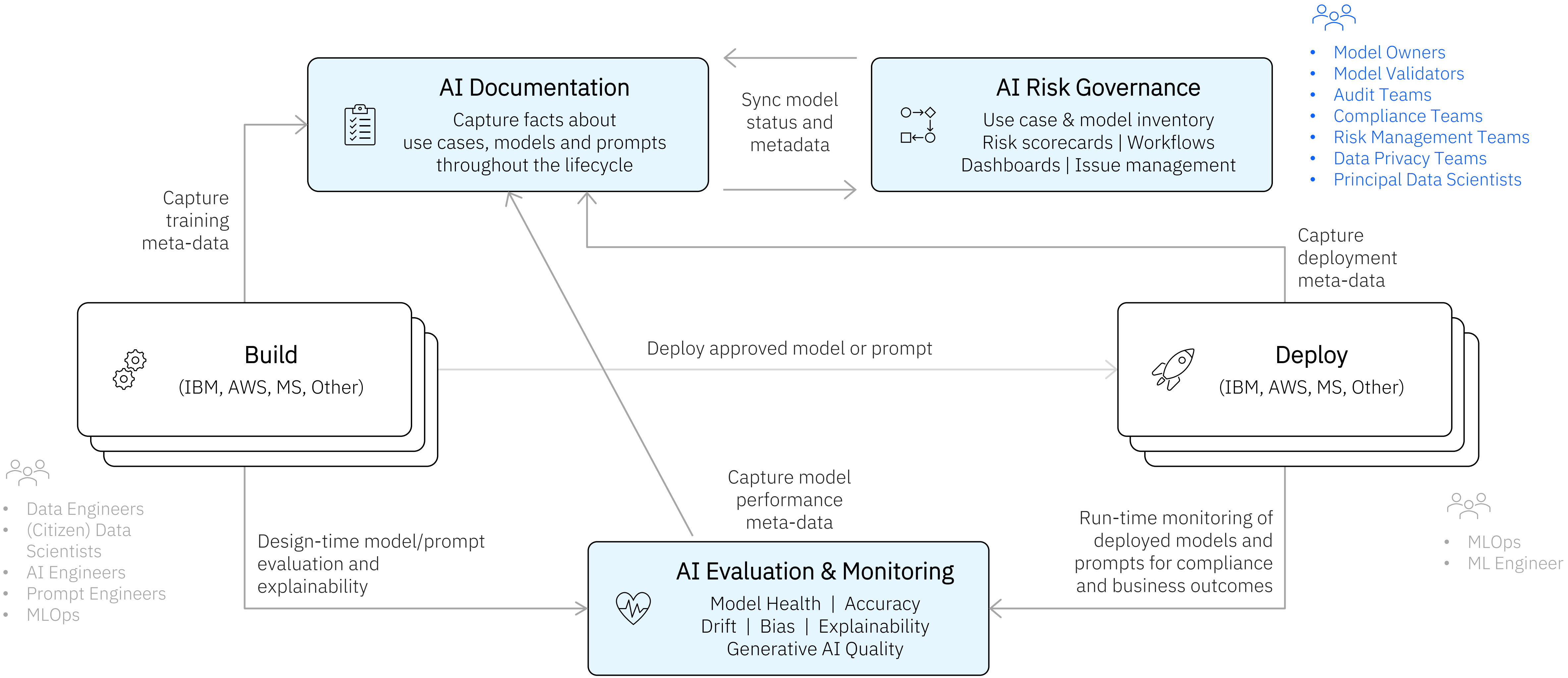
## Business Impact

AI governance, as part of the organizational governance structure, **enacts responsible AI**, and provides *common implementation and adherence mechanisms* across the business ecosystem when it comes to:

- Ethics, fairness, and safety to protect the business and its reputation
- Trust and transparency to support AI adoption via *explainability*, *bias mitigation*, model *governance*, *operationalization*, and *collaboration* norms and capabilities
- Diversity to ensure the right technology and roles for each AI project

Gartner – Hype Cycle for Data and Analytics Governance, 2023

# watsonx.governance



## AI Documentation

Capture facts about
use cases, models and prompts
throughout the lifecycle

Sync model
status and
metadata

## AI Risk Governance

Use case & model inventory
Risk scorecards | Workflows
Dashboards | Issue management

- Model Owners
- Model Validators
- Audit Teams
- Compliance Teams
- Risk Management Teams
- Data Privacy Teams
- Principal Data Scientists

Capture
training
meta-data

Capture
deployment
meta-data

## Build

(IBM, AWS, MS, Other)

Deploy approved model or prompt

## Deploy

(IBM, AWS, MS, Other)

- Data Engineers
- (Citizen) Data Scientists
- AI Engineers
- Prompt Engineers
- MLOps

Capture model
performance
meta-data

Design-time model/prompt
evaluation and
explainability

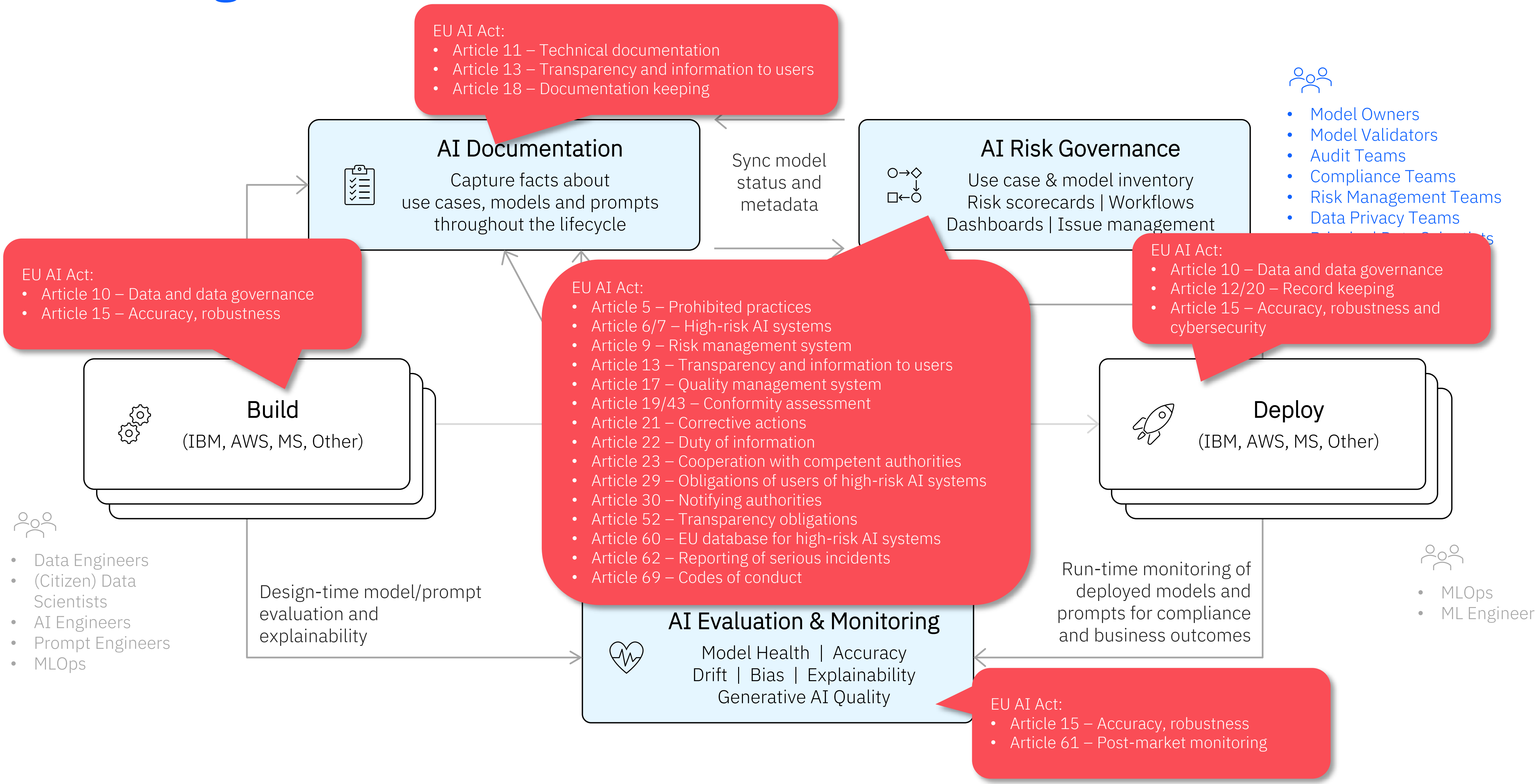
Run-time monitoring of
deployed models and
prompts for compliance
and business outcomes

- MLOps
- ML Engineer

## AI Evaluation & Monitoring

Model Health | Accuracy
Drift | Bias | Explainability
Generative AI Quality

# watsonx.governance



EU AI Act:
- Article 11 – Technical documentation
- Article 13 – Transparency and information to users
- Article 18 – Documentation keeping

## AI Documentation

Capture facts about use cases, models and prompts throughout the lifecycle

Sync model status and metadata

## AI Risk Governance

Use case & model inventory
Risk scorecards | Workflows
Dashboards | Issue management

- Model Owners
- Model Validators
- Audit Teams
- Compliance Teams
- Risk Management Teams
- Data Privacy Teams
- Principal Data Scientists

EU AI Act:
- Article 10 – Data and data governance
- Article 15 – Accuracy, robustness

## Build

(IBM, AWS, MS, Other)

- Data Engineers
- (Citizen) Data Scientists
- AI Engineers
- Prompt Engineers
- MLOps

EU AI Act:
- Article 5 – Prohibited practices
- Article 6/7 – High-risk AI systems
- Article 9 – Risk management system
- Article 13 – Transparency and information to users
- Article 17 – Quality management system
- Article 19/43 – Conformity assessment
- Article 21 – Corrective actions
- Article 22 – Duty of information
- Article 23 – Cooperation with competent authorities
- Article 29 – Obligations of users of high-risk AI systems
- Article 30 – Notifying authorities
- Article 52 – Transparency obligations
- Article 60 – EU database for high-risk AI systems
- Article 62 – Reporting of serious incidents
- Article 69 – Codes of conduct

EU AI Act:
- Article 10 – Data and data governance
- Article 12/20 – Record keeping
- Article 15 – Accuracy, robustness and cybersecurity

## Deploy

(IBM, AWS, MS, Other)

- MLOps
- ML Engineer

Design-time model/prompt evaluation and explainability

Run-time monitoring of deployed models and prompts for compliance and business outcomes

## AI Evaluation & Monitoring

Model Health | Accuracy
Drift | Bias | Explainability
Generative AI Quality

EU AI Act:
- Article 15 – Accuracy, robustness
- Article 61 – Post-market monitoring

IBM