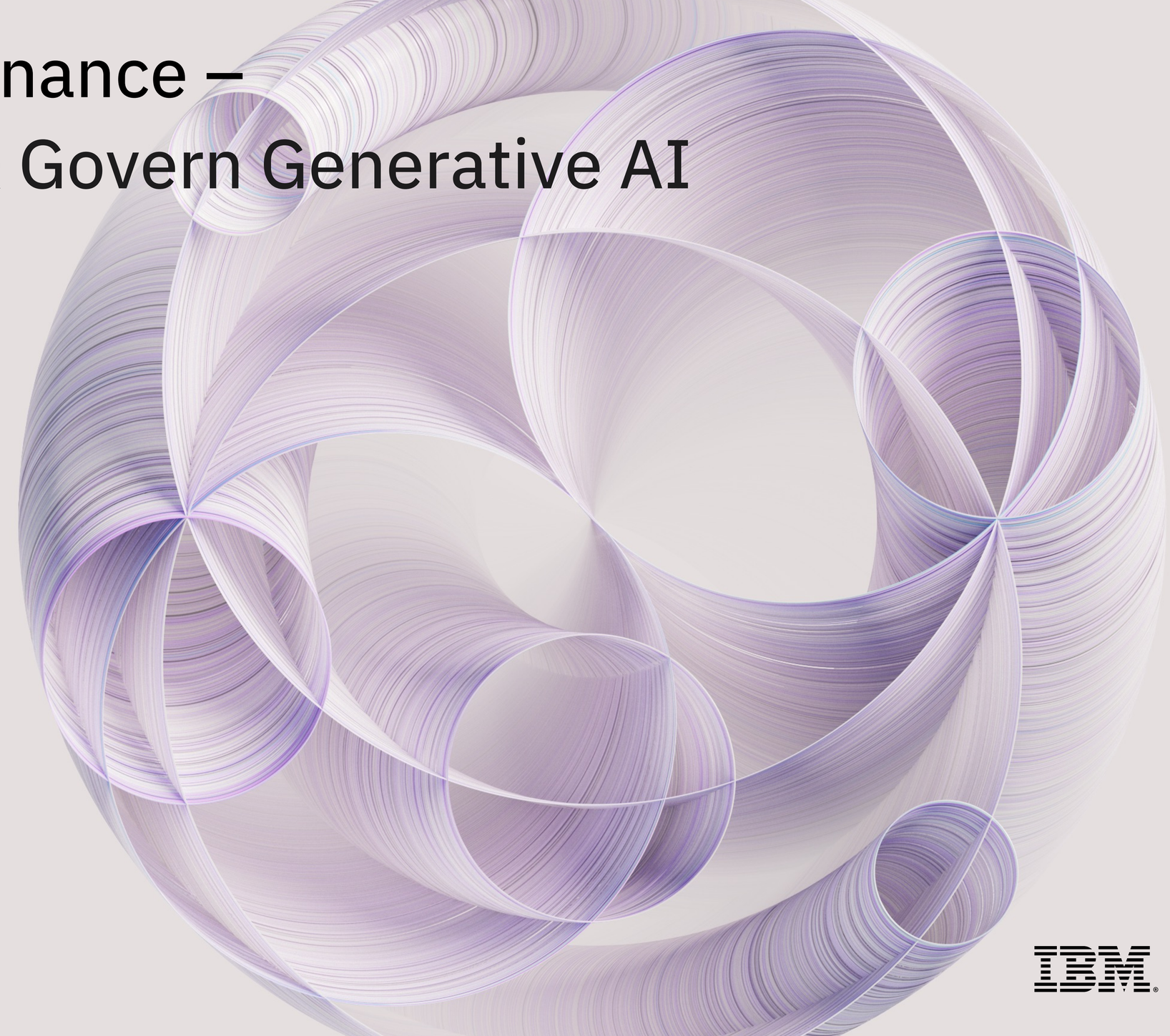
# IBM watsonx.governance – Scale, Accelerate & Govern Generative AI



K.Nijesh
Lead AI Engineer-watsonx
IBM APAC

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

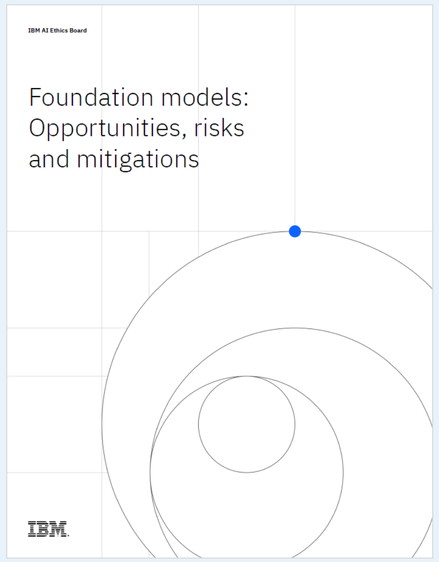
...but how enterprises adopt and execute will define whether they unlock value at scale

# Agenda for watsonx.governance hands-on Workshop

| Time | Topic |
|---|---|
| 0930-1000 | Environment Clinic |
| 1000- 1030 | Intro to watsonx.governance – Concepts and value proposition? |
| 1030-1045 | Lab 101- Project and Spaces Setup |
| 1045-1100 | Break |
| 1100-1115 | Prompt Evaluation Demo |
| 1115-1200 | Lab102 – Assisted prompt Evaluation |
| 1200-1300 | Lunch |
| 1300-1315 | Model Monitoring with watsonx.governance |
| 1315-1415 | Lab 103 - Model Monitoring |
| 1415-1430 | Break |
| 1430-1445 | Automated Model Monitoring Setup |
| 1445-1545 | Lab 104- Automated Model Monitoring |
| 1545- 1600 | Q& A |

# The Risks are amplified

# Foundation Model Risks




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

## Risk Associated with Input

Training and Fine-tuning Phase

- Bias
- Data poisoning attack
- Increased chances of undesirable outputs (inaccurate, inappropriate, etc.) getting incorporated into the training data
- Legal restrictions on data
- Copyright and other IP issues
- Inclusion of PI and SPI
- Data transparency challenges

Inference Phase

- Disclosure of PI/SPI/Copyright/other IP information as a part of prompt
- Adversarial attacks like evasion, prompt injection, prompt leaking, and jail breaking

## Risk Associated with Output

- Bias in generated content
- Performance disparity
- Copyright infringement
- Value alignment issues (e.g., Hallucination)
- Misuse
- Exposing PI and SPI in the output
- Explainability challenges
- Traceability challenges

## Challenges

- Transparency challenges
- Challenge around assigning responsibility
- IP issues
- Human exploitation
- Impact on jobs
- Environmental Impact
- Diversity and Inclusion
- Human agency
- Impact on education

# Evaluate your generative AI platform against these questions

## How was it trained?

- Garbage in -> garbage out
- An enterprise should not use a foundation model trained with a Wikipedia crawl
- The training material must be huge and comprehensive, but must also be curated

## Can it detect & minimize bias & hallucination?

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

## Is it transparent?

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

## Does it support regulatory compliance?

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

## Is it safe?

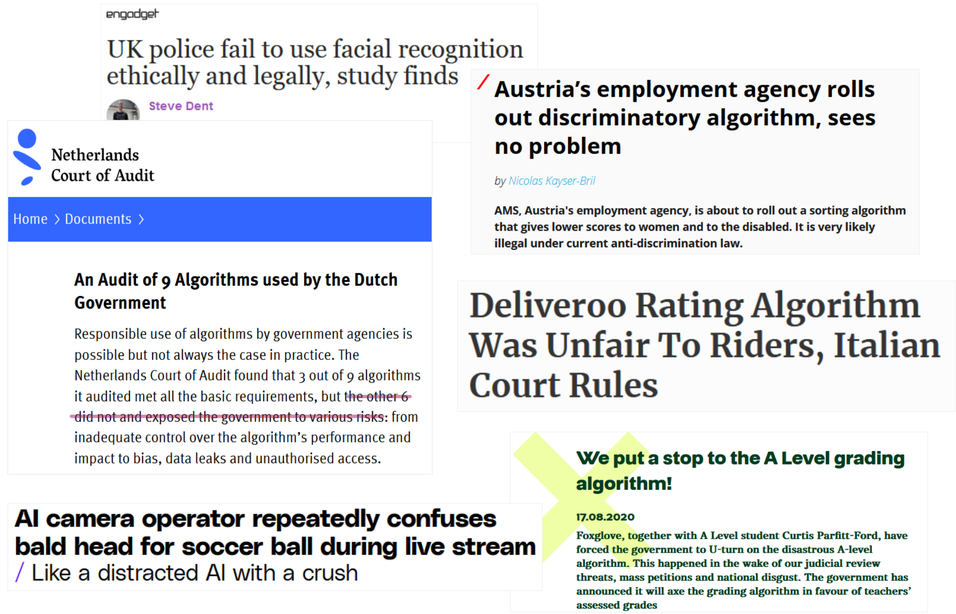
- Who has control over the model, input data, and output data?
- Can you ensure that confidential information is not given out?
- How is it monitored?
- What safety features and guardrails are in place?

## Can it be customized?

- Hybrid and multicloud?
- Can the model be fine-tuned with your data?
- How can you update, and extend the model to make it more suitable for your use cases?
- How to integrate with applications?

# AI needs governance – the process of directing, monitoring and managing the AI activities of an organization

# For most companies **Regulation** and **Reputation** are the driving factors behind investments in AI Governance


Nearly all (97%) respondents believe that regulation will impact them to some extent and 95% believe that at least part of their business will be affected by the EU regulations specifically.

25% have yet to establish any meaningful Responsible AI capabilities.

*Accenture - From AI compliance to competitive advantage, 2022*

"Fewer than 20% of executives strongly agree that their organizations' practices and actions on AI ethics match (or exceed) their stated principles and values."

*IBM and Oxford Economics – AI ethics in action, 2021*

# AI needs governance – the process of directing, monitoring and managing the AI activities of an organization

Accountability

Accuracy

Fairness

Truthfulness

Transparency

Drift

Trusted data

Energy consumption

Explainability

Adversarial Robustness

IP/PII leakage

...

Regulatory Risk

Reputational Risk

Operational Risk

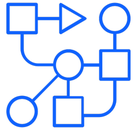
# Governance necessities

## Monitor and evaluate

- Monitor predictive models for fairness, accuracy, and drift
- Monitor generative models for PII and HAP, with additional monitors coming soon
- Explain model predictions and output

## Track facts and metrics

- Automatically gather model metrics and metadata
- Provide model information in a fully-managed, searchable catalog
- Track models throughout the entire lifecycle



## Manage lifecycle and risk

- Fully customize model approval workflows, from initial request to production deployment
- Track risk for all models across the enterprise
- Configure dashboards and reporting for model performance

# IBM watsonx.governance

a powerful toolkit built to direct, manage and monitor the AI activities of an organization

Build enduring consumer trust with your brand

Boost productivity and accelerate business outcomes

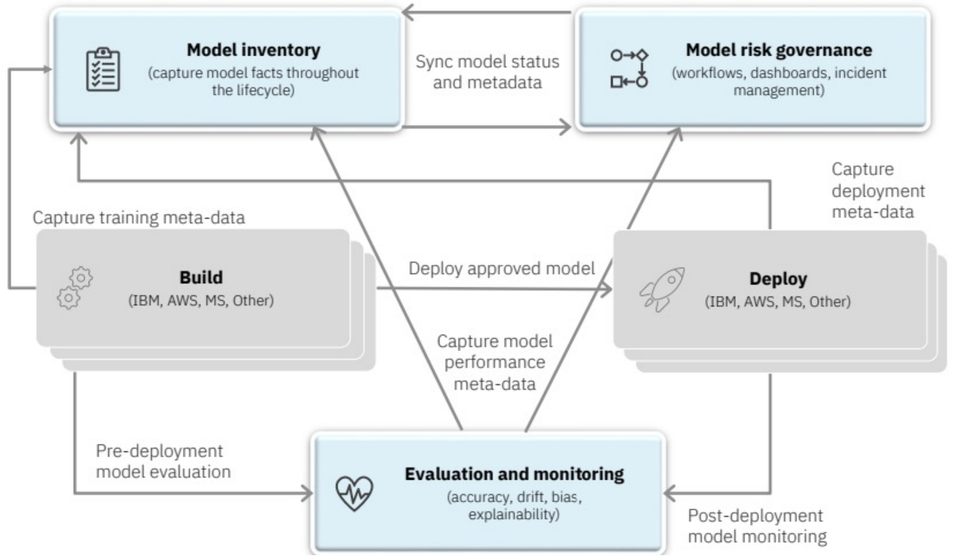
Mitigate risk and minimize cost of compliance

# Scale and Accelerate the impact of AI

## watsonx.ai

A next generation enterprise studio for AI builders to train, validate, tune and deploy generative AI, foundation models, and machine learning capabilities

- Foundation Model Library with IBM and open-source models
- Prompt Lab to experiment with foundation models and build prompts for various use cases and tasks
- Tuning Studio to tune your foundation models with labeled data
- Data Science and MLOps to build machine learning models automatically with model training, development and visual modeling




## watsonx.governance

Accelerate responsible, transparent and explainable AI workflows across the entire lifecycle.

- Automate and consolidate tools, applications, and platforms
- Govern ML models, including those from 3rd parties and generative models (now in tech preview, GA in December 2023)
- Manage risk and protect reputation setting tolerances to proactively detect bias and drift
- Capture metadata and document lineage throughout the model lifecycle
- Improve adherence to AI regulations such as the proposed EU AI Act, internal policies and industry standards
- Improve collaboration & communication with customizable dashboards & reports

Build – train – validate – tune – deploy ⟷ manage – monitor – retrain – document facts

**watsonx.governance**

Accelerate responsible, transparent and explainable AI

*One unified, integrated AI Governance platform to govern Generative AI and Predictive ML*

## Lifecycle Governance

Govern across the AI lifecycle. Automate and consolidate tools, applications and platforms. Capture metadata at each stage and support models built and deployed in 3rd party tools.

## Risk Management

Manage risk and protect reputation by automating workflows to ensure quality and better detect bias and drift.

## Regulatory Compliance

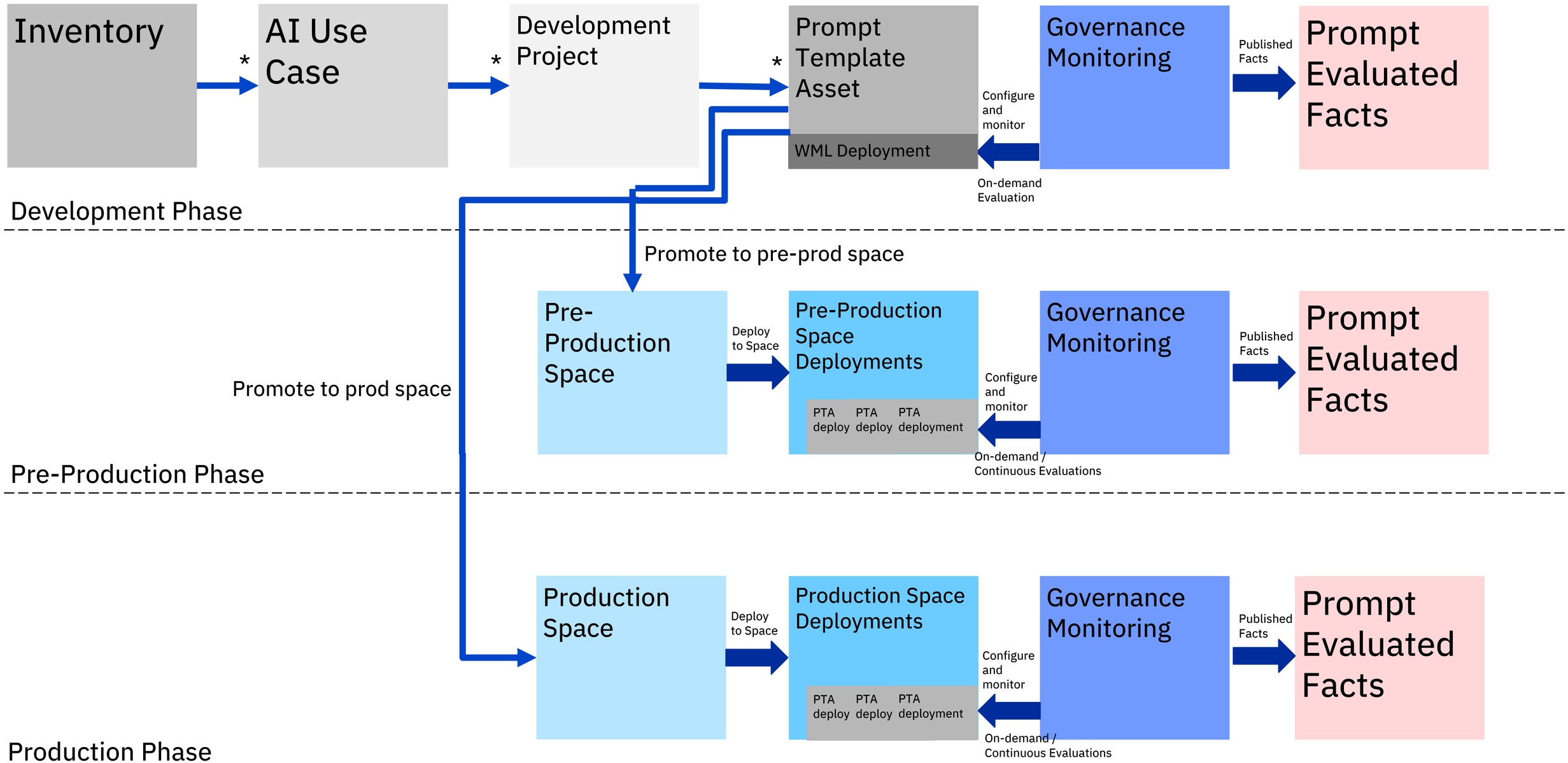
Adhere to regulatory compliance by translating growing regulations into enforceable policies.

**Comprehensive**
Govern the end-to-end AI lifecycle with metadata capture at each stage

**Open**
Support governance of models built and deployed in 3rd party tools.

**Automatic metadata**
and data transformation/lineage capture though Python notebooks.

# IBM watsonx.governance - Prompt Evaluation Lifecycle



# Factsheets monitors LLM Lifecycle

## Prompt Factsheets to improve the transparency of AI Lifecycle for LLMs

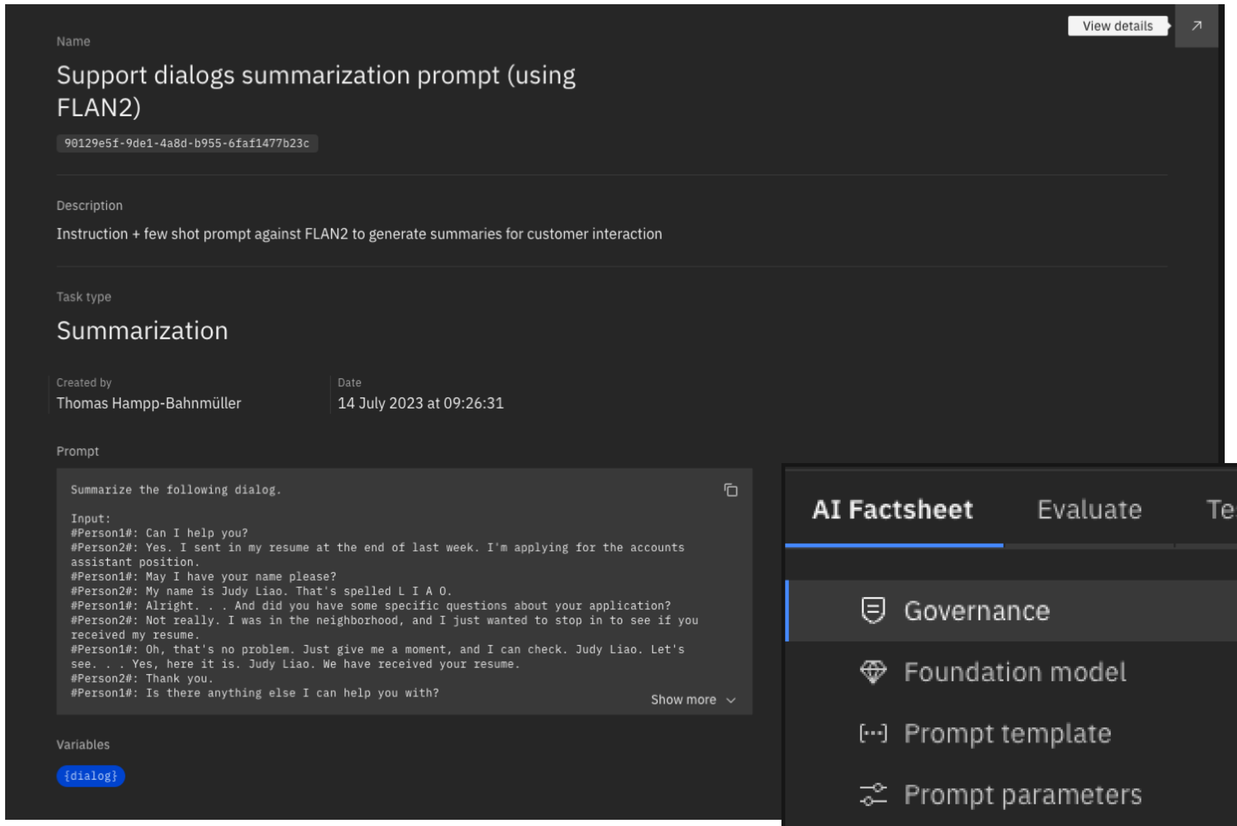
- Automatically capture foundational models and prompt template facts from the prompt lab in watsonx.ai.
- Review Prompt Factsheets details from project, spaces, or prompt lab in watsonx.ai

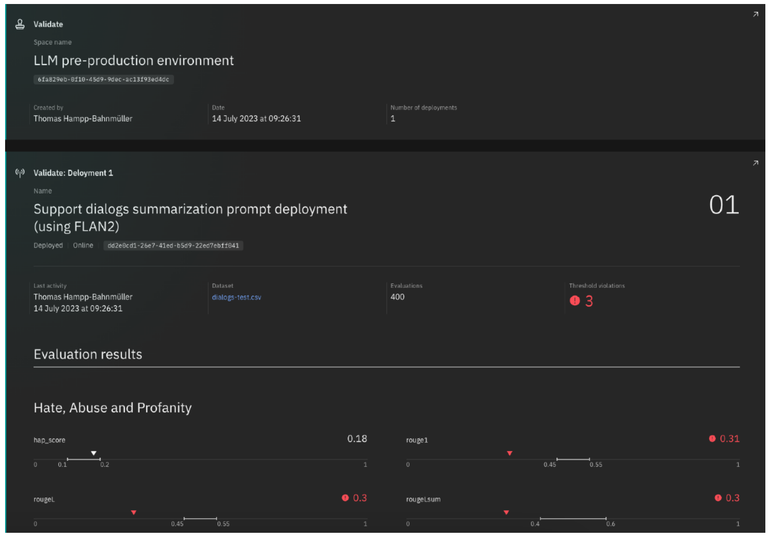
## End-to-end LLM prompt governance

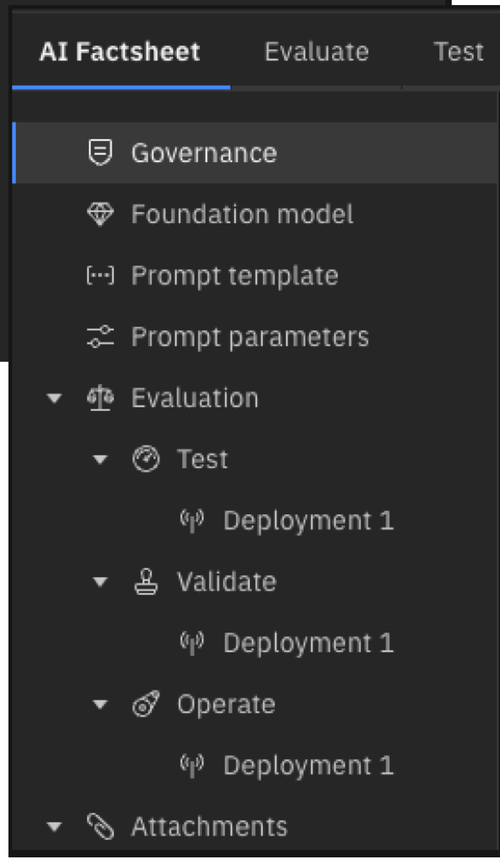
- Submit prompt templates for governance from projects, or from prompt lab.
- Track AI use cases, approaches, and versions associated with prompt templates with the help of prompt factsheets.
- Integrated with monitoring and risk management capabilities within watsonx.governance to provide end-to-end governance across the AI lifecycle.

## Factsheet 2.0

- Easy access to all AI facts within one factsheet with all facts from project/spaces/deployment.
- Quickly analyze details about AI assets using visualized data and enhanced in-page navigation.



# Metrics for evaluating Large Language Models

- Users will be able to 'Evaluate' their Prompt template for LLMs in watsonx.ai through the Prompt Lab within their Project itself during pre-production, as well as for continuous monitoring in production
- Users can evaluate their prompts for various metrics for use cases like Text Summarization, Text Classification, Language Translation, Content generation and Q&A.

## Text Summarization Metrics

- ROUGE
- SARI
- Normalized F1, Precision, Recall
- METEOR
- Sentence Similarity - Jaccard Similarity, Cosine Similarity
- BLEU
- Readability, complexity
- HAP Detection on Input and Output Text
- PII Detection on Input and Output Text

## Content Generation Metrics

- ROUGE
- BLEU
- METEOR
- exact_match
- Readability, complexity
- HAP Detection on Input and Output Text
- PII Detection on Input and Output Text

## Q&A Evaluation Metrics

- ROUGE
- BLEU
- METEOR
- exact_match
- HAP Detection on Input and Output Text
- PII Detection on Input and Output Text

## Drift Monitoring (applicable for all Task Types)

- Metadata Drift (applicable for both Input and Output content)
- Context Drift
- Confidence Drift
- Distribution Drift

## Explainability Monitoring

- Attributions Detection using RAG

## Entity Extraction Metrics

- Micro & Macro F1, Precision, Recall
- Rouge
- Exact Match

## Quality/Text Classification Metrics

- Accuracy
- Precision
- Recall
- ROC AUC
- F1 Score
- Matthews Correlation Coefficient
- Label Skew

## Model Health Metrics (applicable for all Task Types)

- Total Scoring Requests
- Number of scoring records
- Input token count
- Output token count
- API Latency
- API Throughput
- Record Latency
- Record Throughput
- User count

# Model Monitoring Platform

- **Foundation Model Monitoring**: Continuously monitoring LLMs for things such as Drift, Fairness, Quality, Model health, Source attribution, etc.
- **Customized Metric Monitoring**: Platform should provide an easy way to define customized metrics which are specific to a use case / department without having to write a lot of code.
- **Alerting**: Capability to generate alerts when metrics cross a threshold.
- **Remediation**: Platform should help remediate the issues that have been reported.

# Lifecycle Governance

- **Customizable Use Case definition**: Platform should be easily extensible so as to capture different requirements specific to each department.
- **Customizable Process Governance**: Platform should make it very easy to automate and govern the process for each division.
- **Customizable Dashboards**: Platform should make it easy to build customizable AI Governance dashboards for different stack holders across departments

# Types of Drift in LLMs

- **Metadata Drift:** Identifies change in structure of Input and Output to the LLM. E.g., number of words, paragraphs, sentences, characters, etc.

  Input structure: LLM approved for summarizing customer emails. Now being used to summarize customer phone calls.
  Output structure: Customer sentiment classification output changes to “Happy” to “Customer is happy”

- **Content Drift**: Change in topic of the data which is sent to input or output of LLM.

  LLM was approved to summarize drug trials for diabetes and heart disease but is now being used for cancer documents.

  Also identifies keywords which are signifying the drift.

- **Confidence Drift**: Identify change in the LLM confidence when predicting a class.

  E.g., LLM was analyzing test report and was classifying it as either “Urgent – Meet Doctor” or “Normal”. If test reports are different than what model was trained on, its confidence will drop. This will be identified using confidence drift.

- **Distribution Drift**: Identifies changes in the distribution of the prediction as well as input to the LLM.

  E.g., LLM was classifying “Malignant” 15% of the times and ”Benign” 85%. It has changed to 25% / 75%.

- **Root Cause Analysis: We will not only identify the drift, but also identify the root cause for it.**

  E.g., The increase in Malignant is predominantly from State=NY. This could point to some wrong data processing pipeline sending wrong data to the LLM.

# LLM Health and Monitoring

- Latency and Throughput
- Number of scoring requests/ API Calls
- Cost / Number of tokens
- Payload sizes
- Number of users

# Explainability for LLMs using RAG

- For LLMs available in watsonx.ai, we will provide LLM explainability via attribution by providing context documentation **used as reference** for a particular answer; the location of this context can then be highlighted as a source of the output.

- After providing context/reference data, Prompt engineers can refine prompts and improve responses by being able to attribute how LLM output is mapped to the context data provided.

- Algorithm used: Protodash explainer (IBM Research) - it finds prototypical explanations from a context dataset that are similar to the dataset you want to explain

- This functionality can be used to explain what parts of the context/reference data contributed to the result response from the LLM. The source would be the context/reference data, and the attribution is the parts of this context/reference data that helped the model give the result