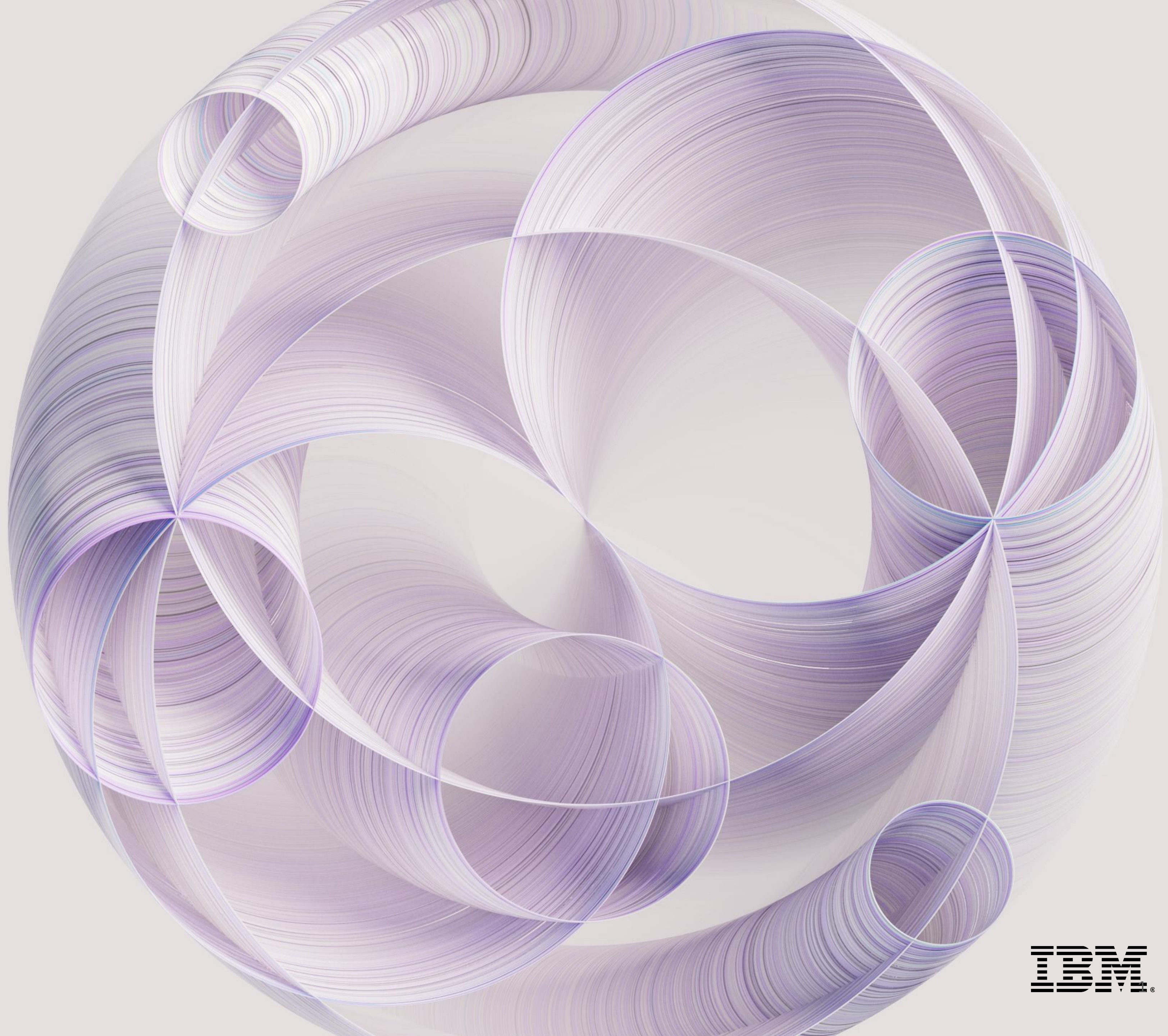


# AI development & AI governance

An end-to-end perspective

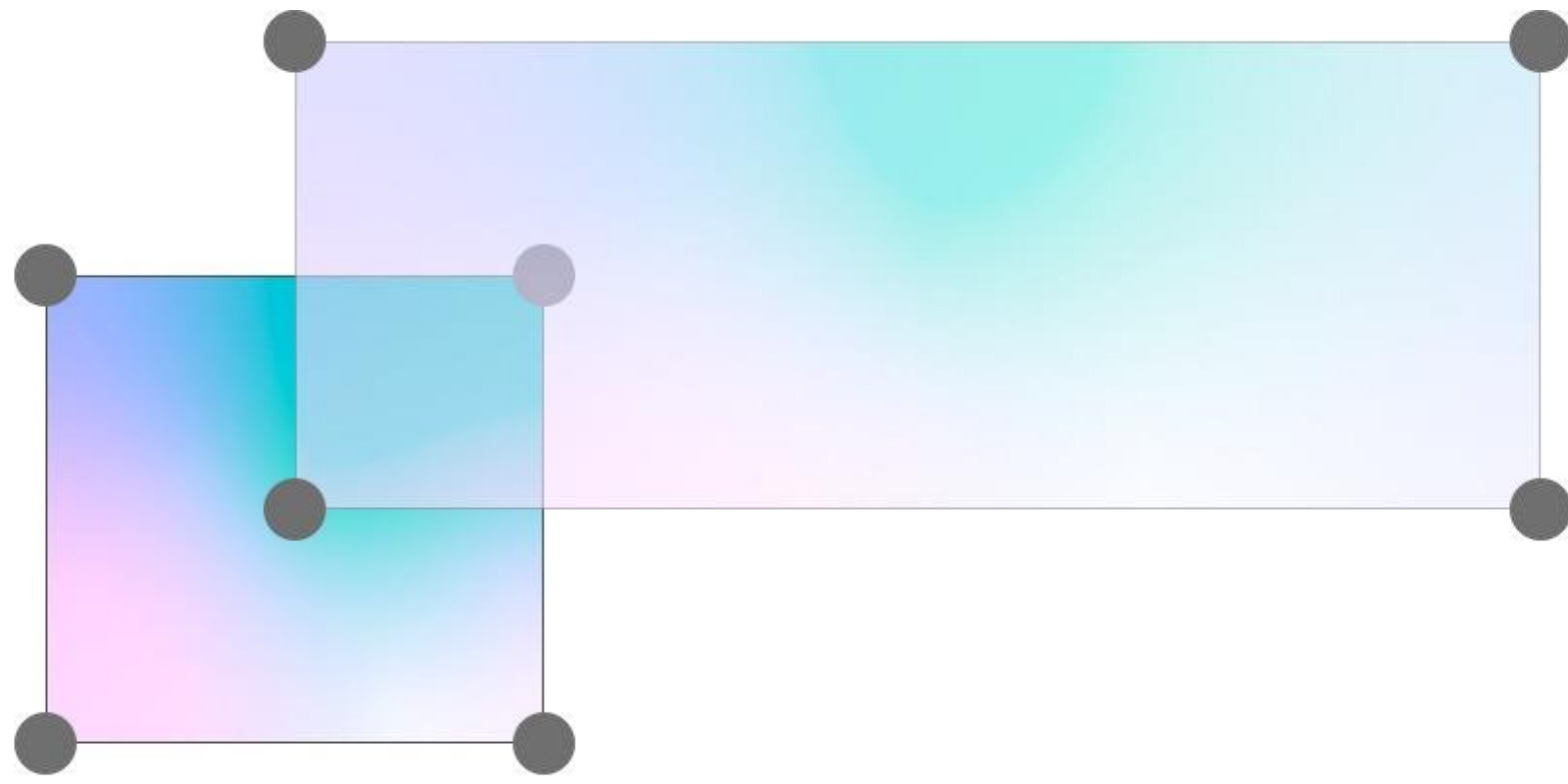


IBM Technology



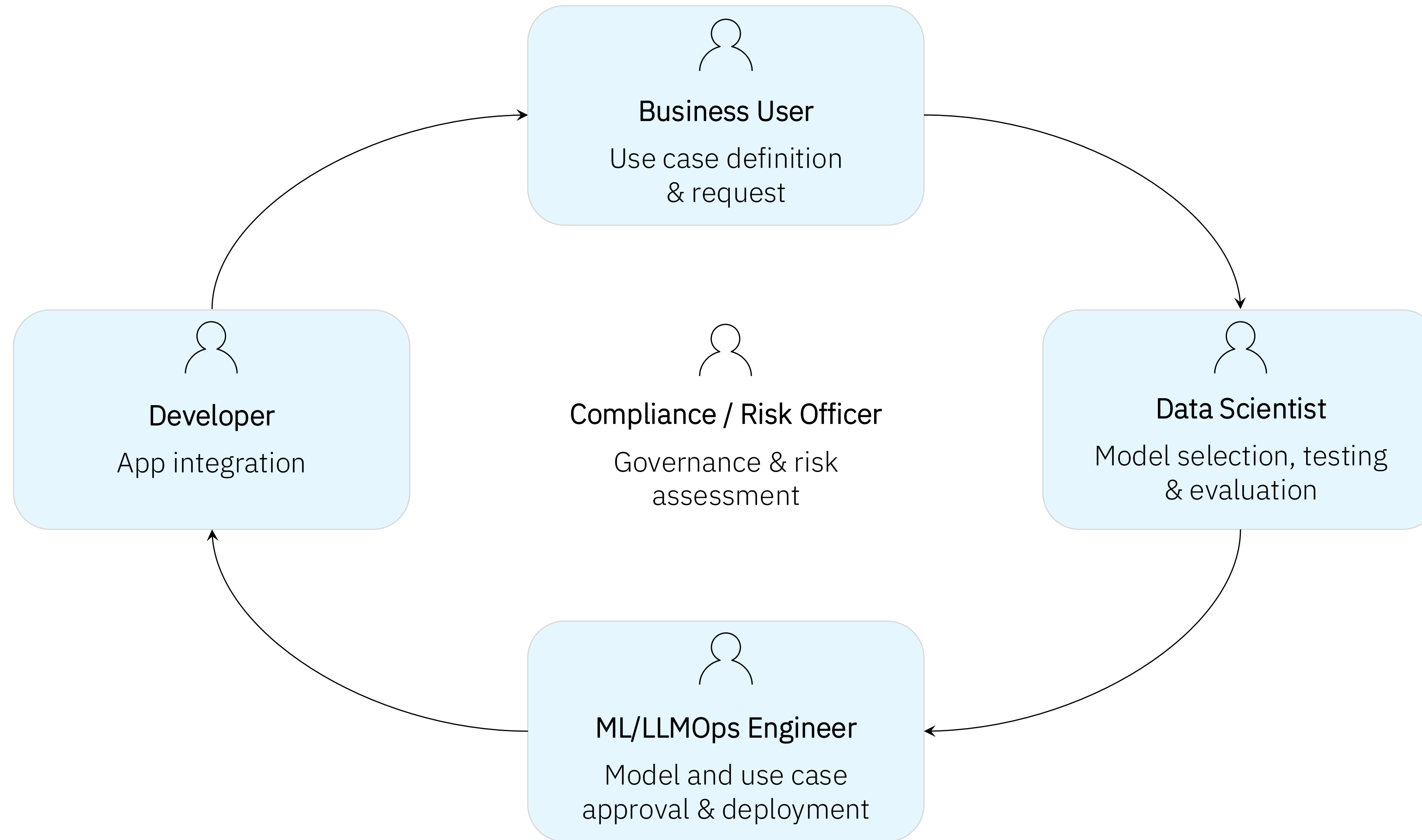
# AI development & AI governance

## An end-to-end perspective



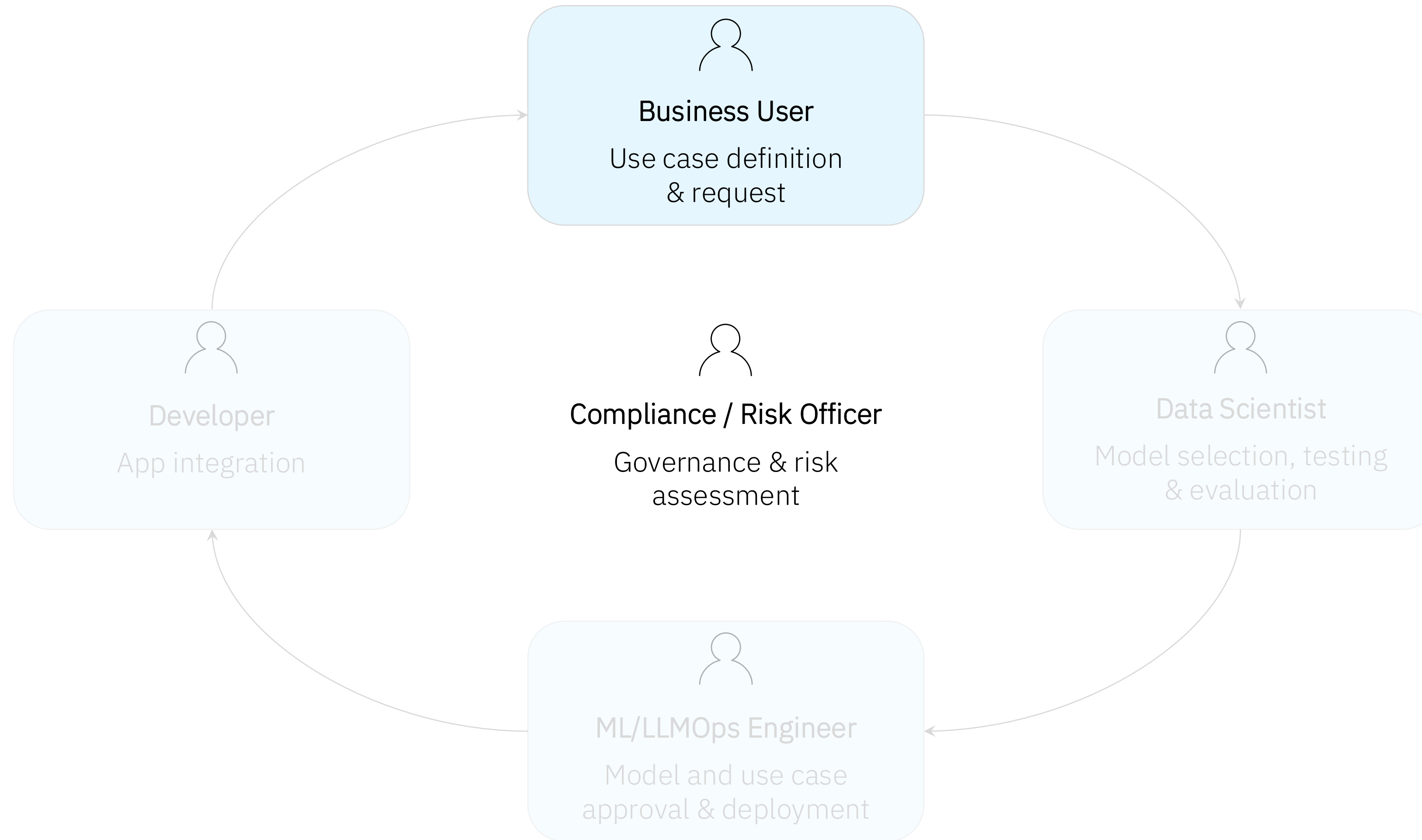
# AI development & AI governance

## Summarization of car insurance claims




# AI development & AI governance

## Summarization of car insurance claims



# AI development & AI governance

Business user would like to derive a brief summary, key information and possible next steps from a claim

  
**Business User**  
Use case definition  
& request

On November 5th, 2023, a fire broke out in the insured's garage, resulting in significant damage to the insured vehicle, a Honda Civic. The insured immediately contacted the fire department, and the fire was extinguished, but not before the vehicle sustained extensive fire damage. The fire damaged the exterior paint, melted parts of the body, and caused smoke and soot damage to the interior. The insured promptly reported the incident to their insurance company and is filing a claim for the repairs. The insurance company has arranged for an assessment of the damages by a qualified auto repair specialist. The insured is providing any necessary documentation, including photographs of the damaged vehicle and a description of the fire incident. The insurance company will cover the cost of repairs or replacement, subject to the terms and conditions of the policy.

Key information

Make & Model	Honda Civic
Location	Not Found
Date	November 5th, 2023
Time of incident	Not Found

Summary

Overview	A fire broke out in the insured's garage, resulting in significant damage to the insured vehicle, a Honda Civic
----------	---

Next steps

Steps to Remediation	<ul style="list-style-type: none"><li><input type="checkbox"/> Verify the insured's policy coverage to ensure that it includes coverage for fire damage and the necessary repairs.</li><li><input type="checkbox"/> Review the provided information regarding the fire incident, including the date of occurrence, location (in the insured's garage), and a description of the damages.</li><li><input type="checkbox"/> Request the insured to provide the police report documenting the fire incident. The police report will serve as crucial evidence and help establish the validity of the claim.</li><li><input type="checkbox"/> Engage a reputable auto repair shop to assess the damages and provide an estimate for the necessary repairs and replacement parts. Consider obtaining multiple estimates to ensure accuracy and fairness.</li><li><input type="checkbox"/> Carefully review all supporting documents, including the police report and the repair shop's estimate. Verify the estimated cost of repairs and validate that the damages align with the incident described by the insured.</li><li><input type="checkbox"/> Maintain regular communication with the insured, providing updates on the claim process and addressing any questions or concerns they may have. Keep them informed about the progress and</li></ul>
----------------------	---

# AI development

## Business user can request an AI use case

  
**Business User**  
Use case definition  
& request

Welcome back, Andreas

Train, deploy, validate,  
and govern AI models  
responsibly.

[Customize my journey](#)

Open in: insurance-demo-v4

[...]

Chat and build prompts with foundation models

Start chatting...

AI

[Open Prompt Lab](#)

Tune a foundation model with  
labeled data

with Tuning Studio

Request or track models in AI  
use cases

with AI governance

Collapse ^

### Jump back in


Recently visited pages

insurance-demo-v4 /  
[Prompt Lab](#)

Collapse Discover section ^

### Discover

#### Resource hub


 Foundation models

#### Featured



# AI development

## Business user can request an AI use case

  
**Business User**  
Use case definition  
& request

### Define a use case

- Risk assessment
- EU AI Act applicability assessment
- Purpose
- Supporting documentation
- Mitigations
- Status, e.g., „Ready for use case approval“



The screenshot shows a user interface with a card titled "Request or track models in AI use cases" featuring a scale icon and the text "with AI governance". A line connects this card to the "Define a use case" list on the left.

Discover

Resource hub

Foundation models

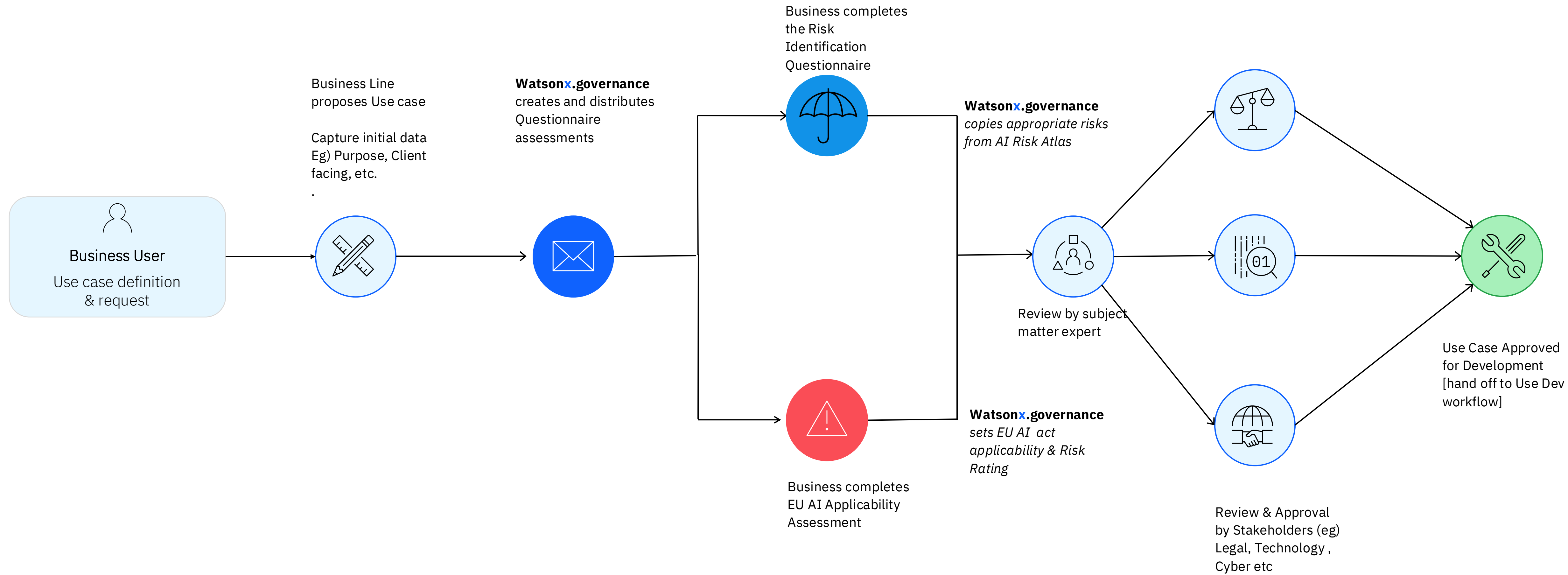
Featured



Collapse Discover section

# AI governance

## Define a use case process

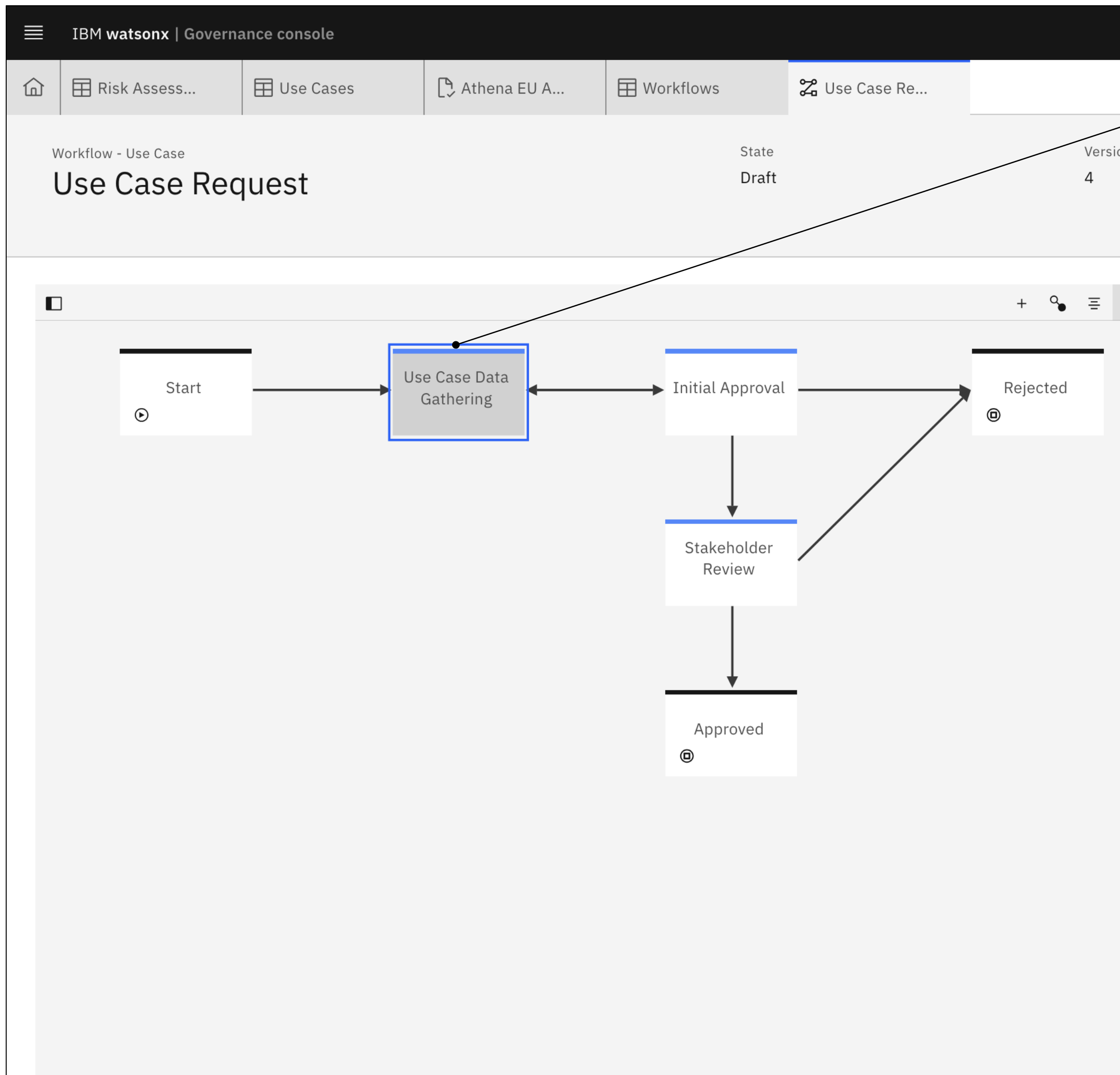


# AI governance

## Define a use case process

### Workflow Management

- Watsonx.governance provides enhanced and customizable workflows which will ensure that the right people get involved in the use case request & approval process at the right time.



Fields [Add Field +](#)

Name	Hidden	Read Only	Key Fields
Technical Owner	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Stakeholder Departments	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Sections [Add Section +](#)

Name	Hidden	Collapsed
No results		

Guidance Title

Use Case Data Gathering


[Edit](#)

Guidance Text

Please capture all relevant information to this AI use case proposal an then submit using the Action button

# AI governance

## New use case request

  
**Business User**  
Use case definition  
& request

IBM watsonx | Governance console

Workflows Use Case Re... + \*New Use Ca... Use Cases Insurance Cl...

### Use Case

## New Use Case

\*Modified Required\*

**General** ⓘ

\*Name\*  \*Owner\*

\*Purpose\*

\*Description\*

\*Use Case Type\*

**Business Entities**

**Primary Business Entity\*** Other Business Entities

### New Use Case Request

- Watsonx.governance provides templates which will guide the use case owner through the process of gathering all relevant information of a use case

use case should be created whenever there [more](#)


1 item requires attention.

All Key Items (6) ▾

- ✓ Name \*
- ✓ Owner \*
- ✓ Purpose
- ✓ Description \*
- ✓ Use Case Type
- ⚠ Primary Business Entity \*

# AI governance

## New use case request

  
**Business User**  
Use case definition  
& request

IBM watsonx | Governance console

Customer attrition prediction ☆ ^

Task Activity Admin

\*Modified Required\*

### Use Case Details

Uses Foundation Models \* Externally Facing Target Implementation Date

Proposed Solution Additional Details

### Risk ⓘ

Risk Level Risk Identification Completion Date Risk Assessment Completion Date

Risk Identification Assessments

Search Add New

Name	Description	Progress (%)	Tags
No results			

Risks

### New Use Case Request

- The use case owner enriches the new use case with additional information.

Criticality Medium

### Tags

No tags have been added yet.

### Use Case Data Gathering ⓘ

Please capture all relevant information to this AI use case proposal and then submit using the Action button

Select an action to validate

1 item requires attention.

All Key Items (6)

- ✓ Purpose
- Risk Level
- ✓ Use Case Type
- ✓ Stakeholder Departments

# AI Risks

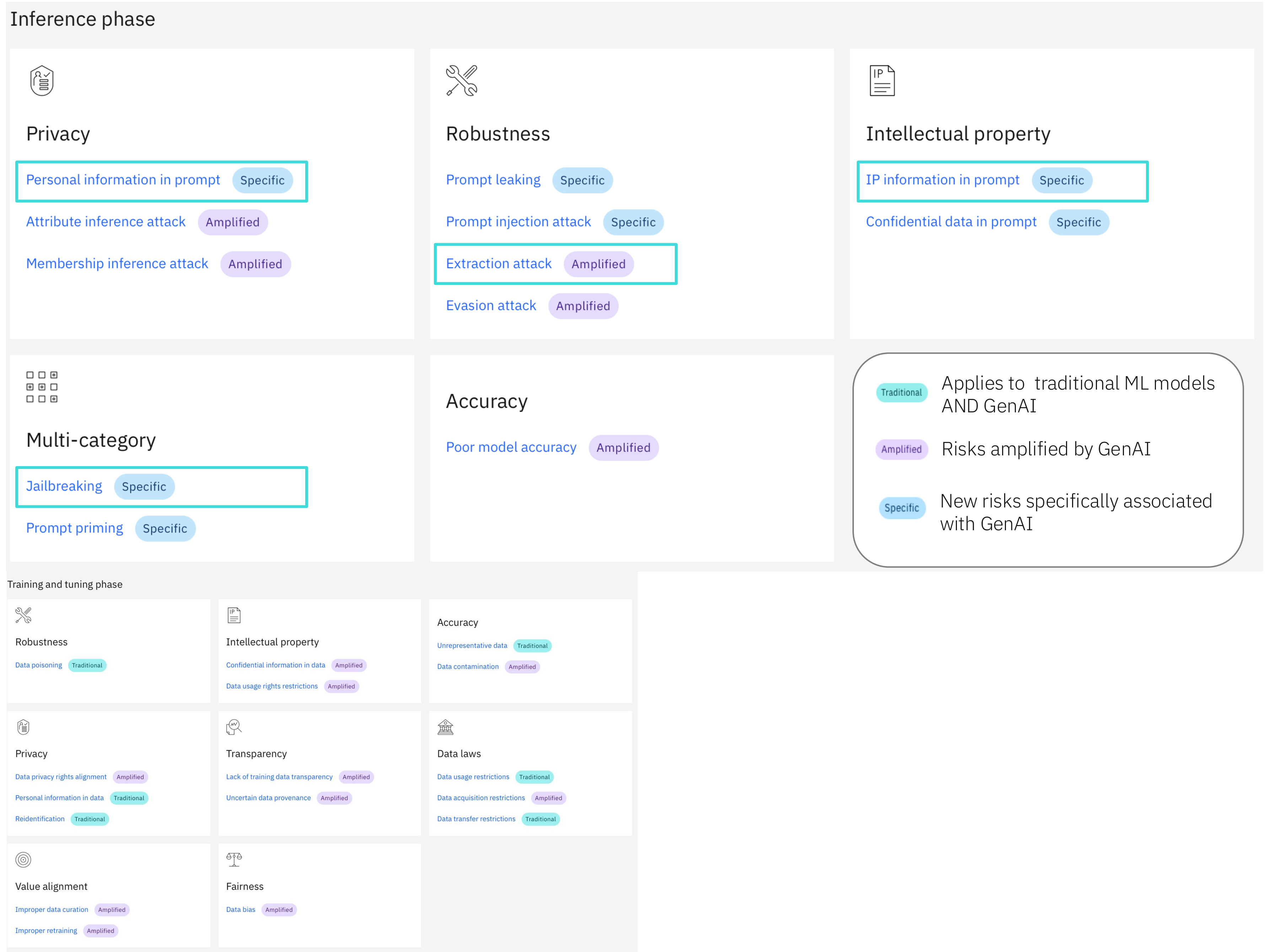
## Risk Associated with Input

### Inference Phase

- Privacy
- Robustness
- Intellectual property
- Multi-category
- Accuracy

### Training and Fine-tuning Phase

- Privacy
- Robustness
- Intellectual property
- Data laws
- Value alignment
- Transparency
- Accuracy



**Traditional** Applies to traditional ML models AND GenAI

**Amplified** Risks amplified by GenAI

**Specific** New risks specifically associated with GenAI

# AI Risks

## Risk Associated with Output

Value alignment

Explainability challenges

Privacy

Misuse

Fairness

Robustness

IP

Harmful code generation



### Value alignment

Toxic output Specific

Harmful output Specific

Incomplete advice Specific

Over- or Under-reliance Amplified



### Explainability

Unreliable source attribution Specific

Unexplainable output Amplified

Inaccessible training data Amplified

Untraceable attribution Amplified



### Privacy

Exposing personal information Amplified



### Misuse

Nonconsensual use Amplified

Non-disclosure Specific

Improper usage Amplified

Dangerous use Specific

Spreading disinformation Specific

Spreading toxicity Specific



### Fairness

Decision bias Traditional

Output bias Specific



### Robustness

Hallucination Specific



### Intellectual property

Revealing confidential information Amplified

Copyright infringement Specific



### Harmful code generation

Harmful code generation Specific

Traditional

Applies to traditional ML models AND GenAI

Amplified


Risks amplified by GenAI

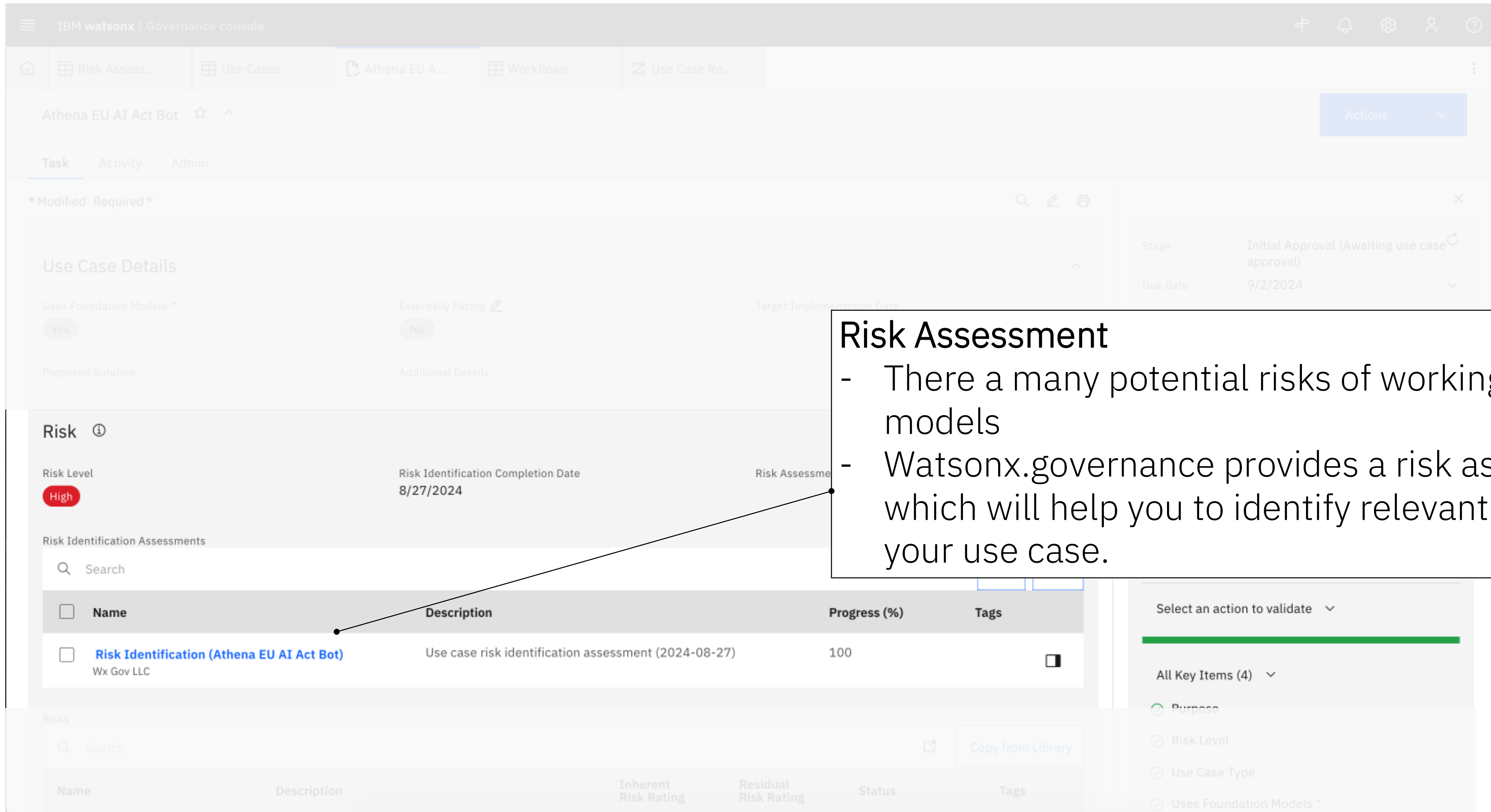
Specific

New risks specifically associated with GenAI

# AI governance

## Define a use case – Risk assessment

  
**Business User**  
Use case definition  
& request




The screenshot displays the IBM Watsonx Governance console interface. At the top, the breadcrumb navigation shows 'IBM watsonx | Governance console'. Below this, there are navigation tabs for 'Risk Assess...', 'Use Cases', 'Athena EU A...', 'Workflows', and 'Use Case Re...'. The main content area is titled 'Athena EU AI Act Bot' and includes an 'Actions' dropdown menu. Underneath, there are tabs for 'Task', 'Activity', and 'Admin'. A search bar with the text '\*Modified Required\*' is visible. The 'Use Case Details' section shows various attributes: 'Uses Foundation Models' (Yes), 'Externally Facing' (No), and 'Target Implementation Date'. A 'Risk' section is highlighted, showing a 'Risk Level' of 'High' and a 'Risk Identification Completion Date' of '8/27/2024'. Below this, there is a table of 'Risk Identification Assessments' with columns for 'Name', 'Description', 'Progress (%)', and 'Tags'. One assessment is listed: 'Risk Identification (Athena EU AI Act Bot)' with a description 'Use case risk identification assessment (2024-08-27)' and a progress of 100%. At the bottom, there is a 'Risks' section with a search bar and a table with columns for 'Name', 'Description', 'Inherent Risk Rating', 'Residual Risk Rating', 'Status', and 'Tags'. A 'Copy from Library' button is also present.

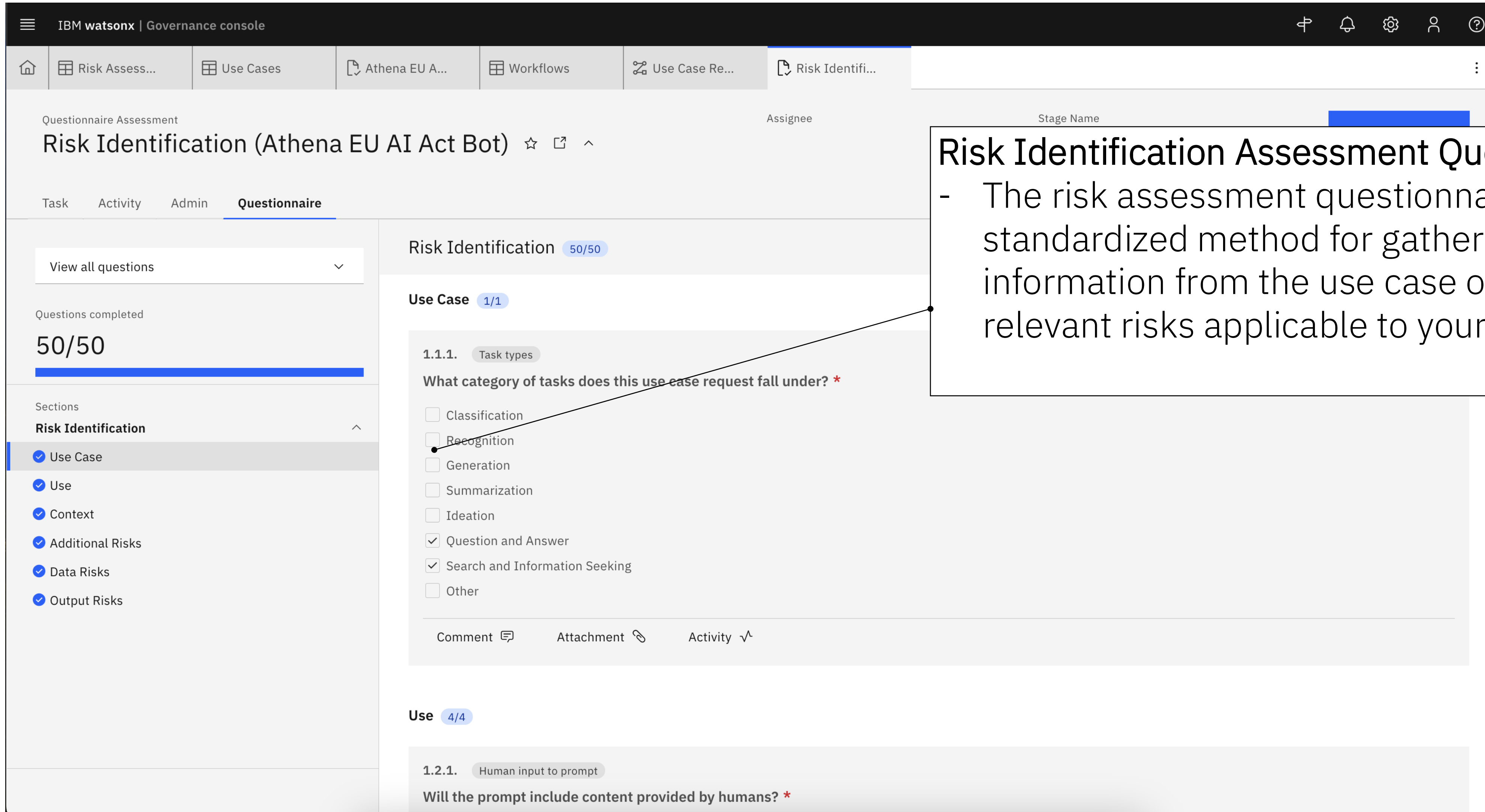
### Risk Assessment

- There are many potential risks of working with AI models
- Watsonx.governance provides a risk assessment which will help you to identify relevant risks based on your use case.

# AI governance

## Define a use case – Risk assessment

  
**Business User**  
Use case definition  
& request




The screenshot displays the IBM Watsonx Governance console interface. The top navigation bar includes the IBM Watsonx logo and the text 'Governance console'. Below this, a breadcrumb trail shows 'Risk Assess...' > 'Use Cases' > 'Athena EU A...' > 'Workflows' > 'Use Case Re...' > 'Risk Identifi...'. The main content area is titled 'Questionnaire Assessment' and 'Risk Identification (Athena EU AI Act Bot)'. It features a 'Questionnaire' tab and a progress indicator showing '50/50' questions completed. A sidebar on the left lists sections under 'Risk Identification', including 'Use Case', 'Use', 'Context', 'Additional Risks', 'Data Risks', and 'Output Risks'. The main content area shows a 'Use Case' section with a task type 'Task types' and a question: 'What category of tasks does this use case request fall under? \*'. The question has several radio button options: 'Classification', 'Recognition', 'Generation', 'Summarization', 'Ideation', 'Question and Answer', 'Search and Information Seeking', and 'Other'. The 'Question and Answer' and 'Search and Information Seeking' options are checked. Below the question, there are icons for 'Comment', 'Attachment', and 'Activity'. A 'Use' section is partially visible at the bottom, with a task type 'Human input to prompt' and a question: 'Will the prompt include content provided by humans? \*'.

**Risk Identification Assessment Questionnaire**

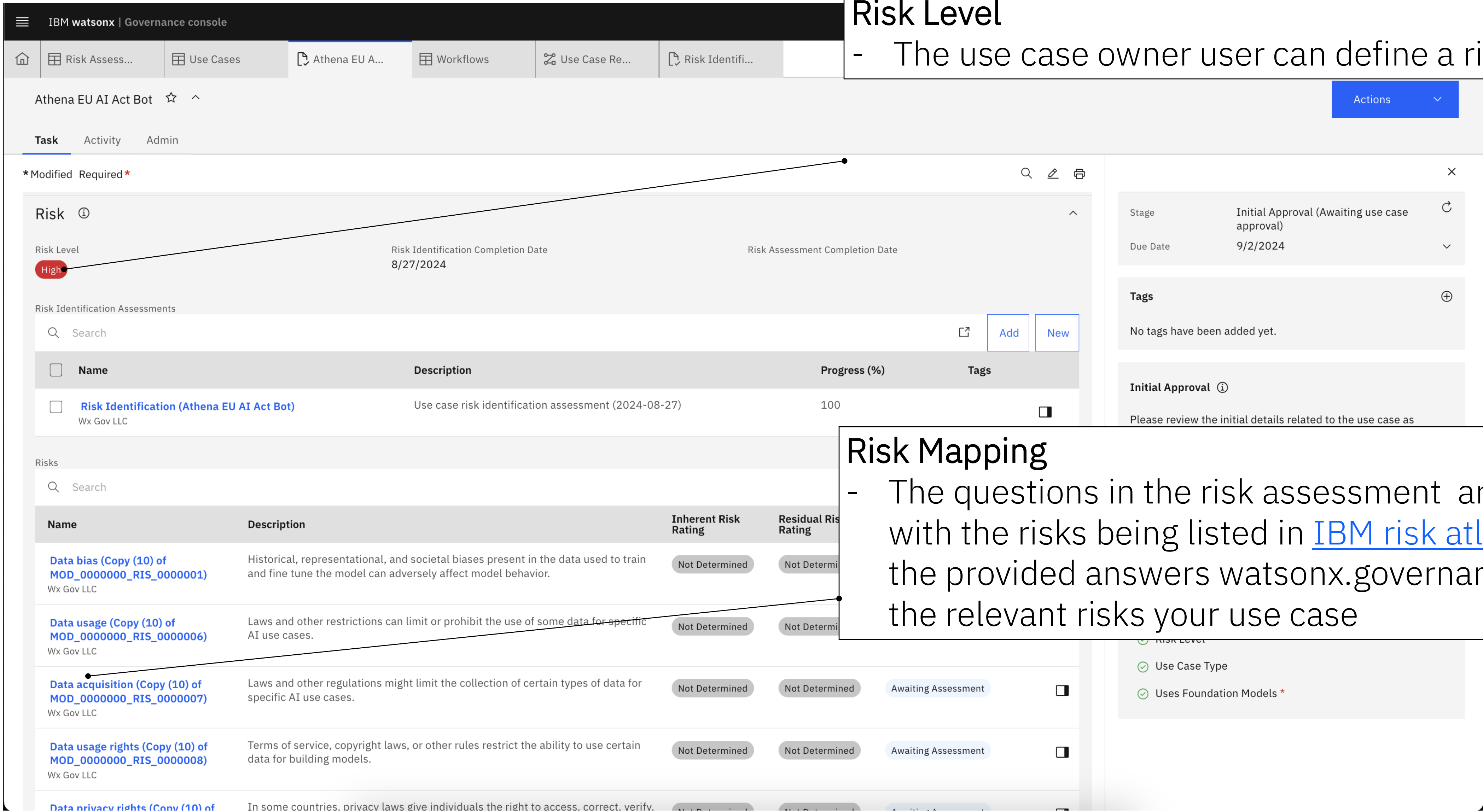
- The risk assessment questionnaire is a structured, standardized method for gathering all pertinent information from the use case owner to identify the relevant risks applicable to your use case.

# AI governance

## Define a use case – Risk assessment

  
**Business User**  
Use case definition & request

**Risk Level**  
- The use case owner user can define a risk level



The screenshot shows the IBM Watsonx Governance console interface. At the top, there's a navigation bar with 'IBM watsonx | Governance console' and several menu items: 'Risk Assess...', 'Use Cases', 'Athena EU A...', 'Workflows', 'Use Case Re...', and 'Risk Identifi...'. Below this, the specific use case 'Athena EU AI Act Bot' is selected, with a 'Task' tab active. A 'Risk' section is highlighted, showing 'Risk Level' set to 'High', 'Risk Identification Completion Date' as '8/27/2024', and 'Risk Assessment Completion Date'. Below this is a table of 'Risk Identification Assessments' with one entry: 'Risk Identification (Athena EU AI Act Bot)' with a progress of 100%. A 'Risks' section follows, listing several risks with their descriptions and ratings. A 'Risk Mapping' callout box points to the 'Data acquisition' risk entry. On the right side, there are panels for 'Initial Approval' (showing 'Initial Approval (Awaiting use case approval)' and 'Due Date' '9/2/2024') and 'Tags' (showing 'No tags have been added yet.').

Name	Description	Inherent Risk Rating	Residual Risk Rating	Progress (%)	Tags
Data bias (Copy (10) of MOD_0000000_RIS_0000001) Wx Gov LLC	Historical, representational, and societal biases present in the data used to train and fine tune the model can adversely affect model behavior.	Not Determined	Not Determined	100	
Data usage (Copy (10) of MOD_0000000_RIS_0000006) Wx Gov LLC	Laws and other restrictions can limit or prohibit the use of some data for specific AI use cases.	Not Determined	Not Determined		
Data acquisition (Copy (10) of MOD_0000000_RIS_0000007) Wx Gov LLC	Laws and other regulations might limit the collection of certain types of data for specific AI use cases.	Not Determined	Not Determined	Awaiting Assessment	
Data usage rights (Copy (10) of MOD_0000000_RIS_0000008) Wx Gov LLC	Terms of service, copyright laws, or other rules restrict the ability to use certain data for building models.	Not Determined	Not Determined	Awaiting Assessment	
Data privacy rights (Copy (10) of MOD_0000000_RIS_0000009) Wx Gov LLC	In some countries, privacy laws give individuals the right to access, correct, verify, and delete their personal data.	Not Determined	Not Determined	Awaiting Assessment	

**Risk Mapping**  
- The questions in the risk assessment are mapped with the risks being listed in [IBM risk atlas](#). Based on the provided answers watsonx.governance will attach the relevant risks your use case

# EU Artificial Intelligence Act



01

Use high-quality training, validation and testing data.

02

Establish documentation and design logging features.

03

Ensure appropriate certain degree of transparency.

04

Ensure human oversight (measures built into the system and/or to be implemented by users).

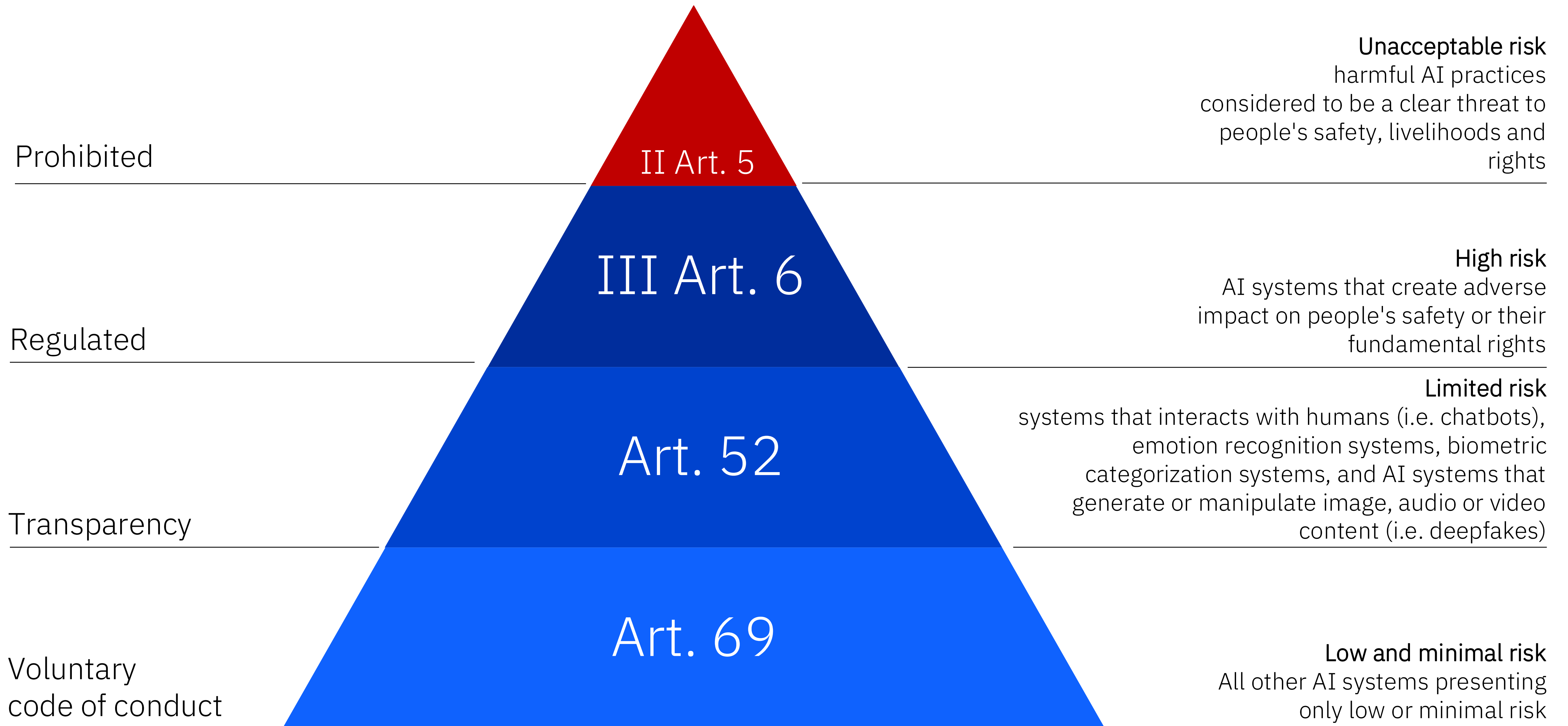
05

Ensure robustness, accuracy, and cybersecurity.

06

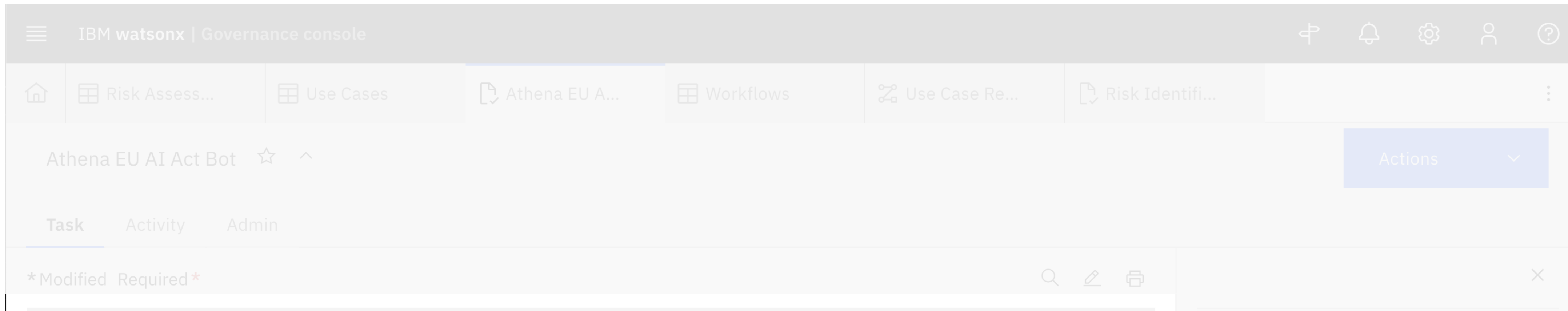
Implement a risk management system

# A risk-based approach to regulations



# AI governance

## Define a use case – Risk assessment



### Regulatory Information

EU AI Risk Category

Applicability Assessment Completion Date

### Applicability Assessments

Mandates

Search

<input type="checkbox"/>	Name	Description	Pr (%
<input checked="" type="checkbox"/>	<a href="#">Applicability Assessment (Athena EU AI Act Bot)</a> Wx Gov LLC	Use case applicability assessment (2024-08-27)	10


### EU AI Act applicability assessment

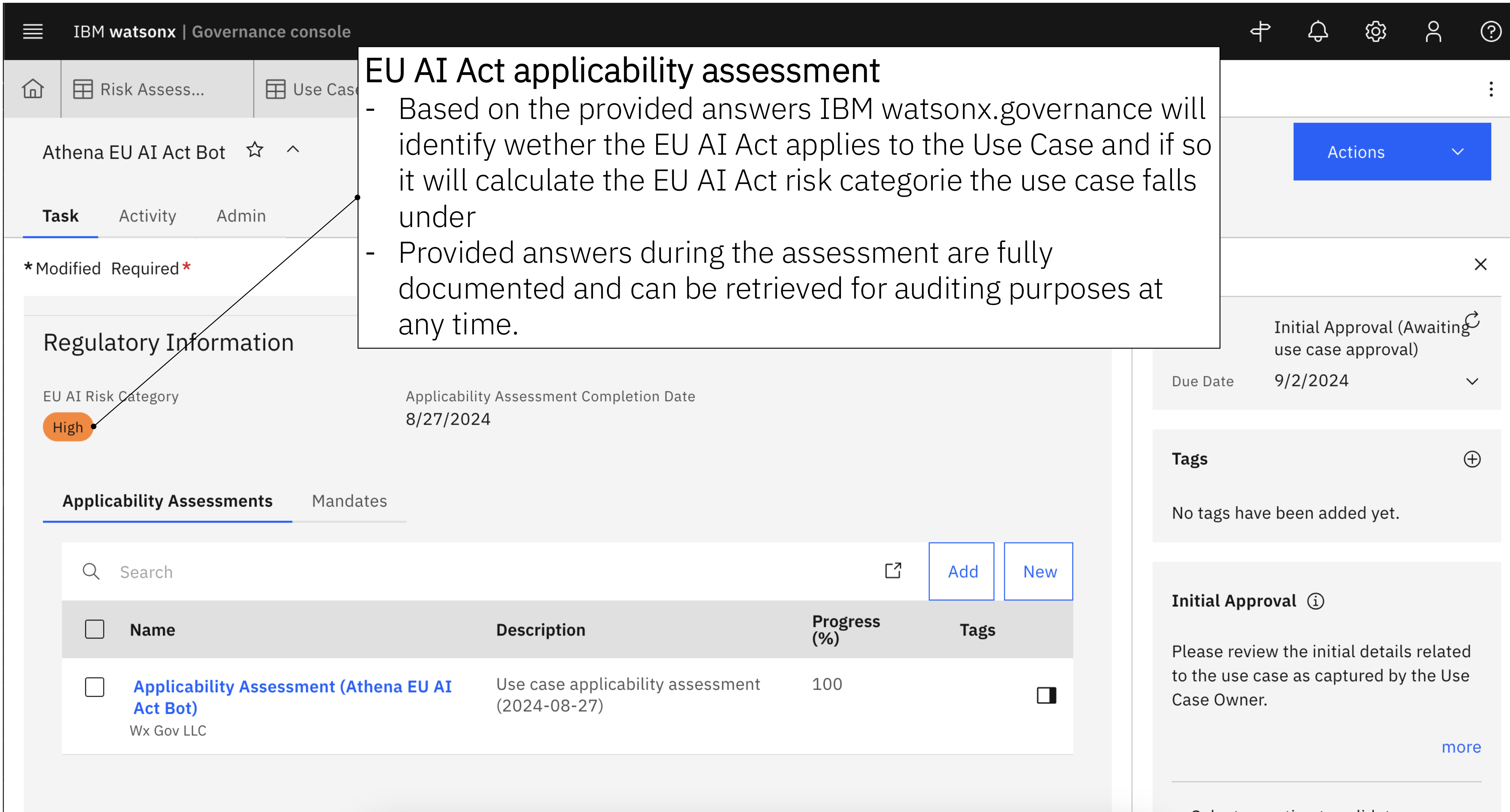
watsonx.governance provides an EU AI Act applicability assessment which comes out of the box with the solution. Future changes to the EU AI Act will be incorporated into questionnaire.

The Applicability Assessment is based on the EU AI Act Applicability Assessment questionnaire template. This assessment enables clients to assess their AI uses cases using a simple questionnaire that aids in determining whether a use case is in scope for the EU AI Act and which risk category the use case aligns to (Prohibited, High, Limited, Minimal).

# AI governance

## Define a use case – EU AI Act applicability assessment

  
**Business User**  
Use case definition  
& request



The screenshot displays the IBM Watsonx Governance console interface. At the top, the header shows 'IBM watsonx | Governance console' and navigation icons. Below the header, there are tabs for 'Risk Assess...' and 'Use Cas...'. The main content area is titled 'Athena EU AI Act Bot' and includes tabs for 'Task', 'Activity', and 'Admin'. A notification '\*Modified Required\*' is visible. The 'Regulatory Information' section shows 'EU AI Risk Category' as 'High' and 'Applicability Assessment Completion Date' as '8/27/2024'. Below this, there are tabs for 'Applicability Assessments' and 'Mandates'. A table lists the assessment details:

Name	Description	Progress (%)	Tags
<input type="checkbox"/> <b>Applicability Assessment (Athena EU AI Act Bot)</b> Wx Gov LLC	Use case applicability assessment (2024-08-27)	100	<input type="checkbox"/>

On the right side, there is an 'Initial Approval' section with a status of 'Awaiting use case approval' and a 'Due Date' of '9/2/2024'. Below this, there is a 'Tags' section with the message 'No tags have been added yet.' and an 'Initial Approval' section with a message: 'Please review the initial details related to the use case as captured by the Use Case Owner.' and a 'more' link.

**EU AI Act applicability assessment**

- Based on the provided answers IBM watsonx.governance will identify whether the EU AI Act applies to the Use Case and if so it will calculate the EU AI Act risk categories the use case falls under
- Provided answers during the assessment are fully documented and can be retrieved for auditing purposes at any time.

# AI governance

Compliance / Risk Officer checks the status of model compliance, regulatory requirements, metrics and use cases



IBM watsonx | Governance Console

Welcome, schneider!  
Last successful login 5/15/2024, 2:29 PM

Dashboard My Tasks (1) Subscription Tasks (0) Oversight Tasks (23)

### My Tasks

1

Overdue (1)  
Due soon (0)  
Due in 2+ weeks (0)

Top 5 by due date  
5/20/2024 [Auto Policy Risk AS150524](#)

### Regulatory Data

My Regulatory Tasks: 0  
[New Regulatory Task](#)

My Regulatory Events: 0  
[New Regulatory Event](#)

My Open Regulatory Changes: 0

Completed Evaluations: 0

### Models by Provider

63

- IBM
- Amazon
- Microsoft
- Databricks
- Hugging Face
- OpenAI
- Dataiku
- DataRobot
- Google
- Other

### Metric Breach Status

98

- Red
- Yellow
- Green
- Not Determined
- (No Value)

### Model Compliance

63

Compliance Status	Count
Compliant	42
Non-compliant	10
(No Value)	11

### Use Case by Lifecycle Phase

46

- Proposed
- Awaiting Approval
- Approved
- Under Review
- Rejected
- Decommissioned

### My Favorites

By time added

- [1-Insurance Claim - Agent Assist](#) Enable agents to process and respond to claims faster by using genAI to: 1. Summarize...

### Useful Links

- [Responsible AI Institute](#)
- [Model Risk Management - Comptroller's Handbook \(OCC\)](#)
- [EU Draft AI Regulation](#)
- [SR 11-7 Information](#)
- [E-23 Information](#)
- [NYC Local Law 144](#)
- [AI Bill of Rights](#)

### Use Case Summary


Use Case Risk Breakdown: 46

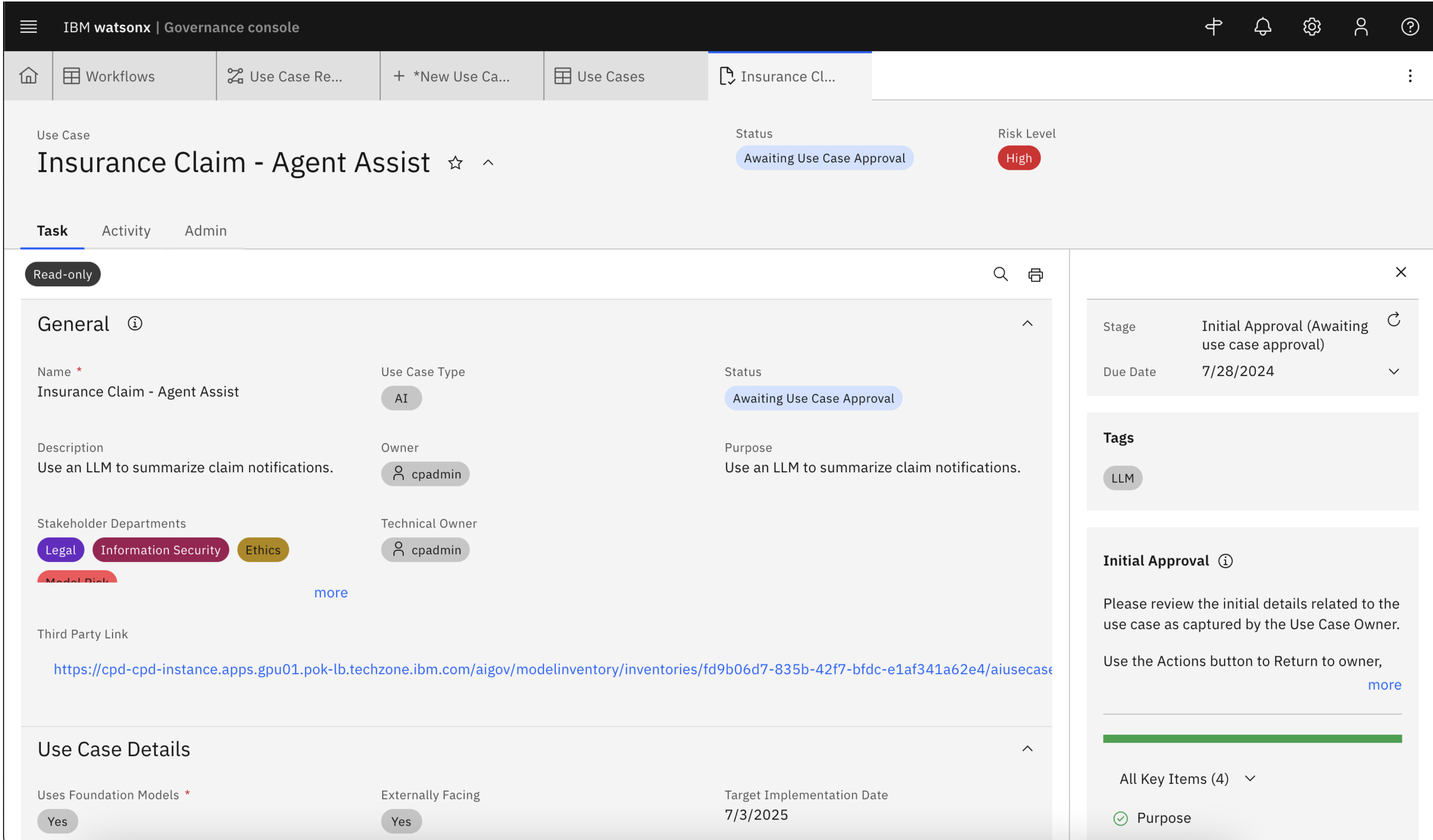
Use Case By Status: 46

High Risk Models: 6

# AI governance

## Define a use case – Documented use case

  
**Business User**  
Use case definition & request



The screenshot displays the IBM Watsonx Governance console interface. At the top, the navigation bar includes 'IBM watsonx | Governance console' and various utility icons. Below this, a breadcrumb trail shows 'Workflows', 'Use Case Re...', '\*New Use Ca...', 'Use Cases', and 'Insurance Cl...'. The main content area is titled 'Use Case' and features the name 'Insurance Claim - Agent Assist' with a star icon and an upward arrow. To the right, the 'Status' is 'Awaiting Use Case Approval' and the 'Risk Level' is 'High'. Below the title, there are tabs for 'Task', 'Activity', and 'Admin'. A 'Read-only' badge is visible in the top left of the details panel. The details panel is divided into 'General' and 'Use Case Details' sections. The 'General' section includes fields for Name, Use Case Type, Status, Description, Owner, Purpose, Stakeholder Departments, and Technical Owner. The 'Use Case Details' section includes fields for Uses Foundation Models, Externally Facing, and Target Implementation Date. A right-hand sidebar contains a 'Stage' section with 'Initial Approval (Awaiting use case approval)' and a 'Due Date' of '7/28/2024'. Below this is a 'Tags' section with an 'LLM' tag. The 'Initial Approval' section contains a message: 'Please review the initial details related to the use case as captured by the Use Case Owner. Use the Actions button to Return to owner, more'. At the bottom of the sidebar, there is a section for 'All Key Items (4)' with a dropdown arrow, and a single item 'Purpose' is listed with a green checkmark.

IBM watsonx | Governance console

Workflows Use Case Re... \*New Use Ca... Use Cases Insurance Cl...

Use Case

Insurance Claim - Agent Assist ☆ ^

Status: Awaiting Use Case Approval Risk Level: High

Task Activity Admin

Read-only

**General** ⓘ

Name \* Insurance Claim - Agent Assist Use Case Type AI Status Awaiting Use Case Approval

Description Use an LLM to summarize claim notifications. Owner cadmin Purpose Use an LLM to summarize claim notifications.

Stakeholder Departments Legal Information Security Ethics Model Risk more

Technical Owner cadmin

Third Party Link <https://cpd-cpd-instance.apps.gpu01.pok-lb.techzone.ibm.com/aigov/modelinventory/inventories/fd9b06d7-835b-42f7-bfdc-e1af341a62e4/aiusecase>

**Use Case Details** ^

Uses Foundation Models \* Yes Externally Facing Yes Target Implementation Date 7/3/2025

Stage Initial Approval (Awaiting use case approval) Due Date 7/28/2024

**Tags** LLM

**Initial Approval** ⓘ

Please review the initial details related to the use case as captured by the Use Case Owner. Use the Actions button to Return to owner, more

All Key Items (4) v

✓ Purpose

# AI governance

Compliance / Risk Officer checks the status of model compliance, regulatory requirements, metrics and use cases



**AI Risk Governance**

- Use case & AI asset inventory
- Workflows | Risk assessments
- Dashboards | Issue management

**My Tasks**  
1  
Overdue (1)  
Due soon (0)  
Due in 2+ weeks (0)  
Top 5 by due date  
5/20/2024 Auto Policy Risk AS150524

**Regulatory Data**  
My Regulatory Tasks: 0  
New Regulatory Task  
My Regulatory Events: 0  
New Regulatory Event  
My Open Regulatory Changes: 0  
Completed Evaluations: 0

**Models by Provider**  
63  
IBM, Amazon, Microsoft, Google, Databricks, Hugging Face, OpenAI, Dataiku, DataRobot

**Model Compliance**  
63  
Compliant, Non-compliant, (No Value)

**Use Case by Lifecycle Phase**  
46  
Proposed, Awaiting Approval, Approved, Under Review, Rejected, Decommissioned

**My Favorites**  
By time added  
1-Insurance Claim - Agent Assist  
Enable agents to process and respond to claims faster by using genAI to: 1. Summarize...

**Useful Links**  
Responsible AI Institute  
Model Risk Management - Comptroller's Handbook (OCC)  
EU Draft AI Regulation  
SR 11-7 Information  
E-23 Information  
NYC Local Law 144  
AI Bill of Rights

**Use Case Summary**  
Use Case Risk Breakdown: 46  
Use Case By Status: 46  
High Risk Models: 6

# AI governance

Compliance / Risk Officer checks the status of model compliance, regulatory requirements, metrics and use cases



**Compliance / Risk Officer**  
Governance & risk assessment

IBM watsonx | Governance Console

Welcome, schneider!  
Last successful login 5/15/2024, 2:29 PM

Dashboard My Tasks (1) Subscription Tasks (0) Oversight Tasks (23)

### My Tasks

1

Overdue (1)  
Due soon (0)  
Due in 2+ weeks (0)

Top 5 by due date

5/20/2024	Auto Policy Risk AS150524
-----------	---------------------------

### Regulatory Data

My Regulatory Tasks: 0

My Regulatory Events: 0

My Open Regulatory Changes: 0

Completed Evaluations: 0

### Models by Provider

1-Insurance Claim - Agent Assist

Enable agents to process and respond to claims faster by using genAI to: 1. Summarize...

### Metric Breach Status

### Model Compliance

63

### Use Case by Lifecycle Phase

46

- Proposed
- Awaiting Approval
- Approved
- Under Review
- Rejected
- Decommissioned

### Useful Links

- Responsible AI Institute
- Model Risk Management - Comptroller's Handbook (OCC)
- EU Draft AI Regulation
- SR 11-7 Information
- E-23 Information
- NYC Local Law 144
- AI Bill of Rights

### Use Case Summary

Use Case Risk Breakdown: 46

Use Case By Status: 46

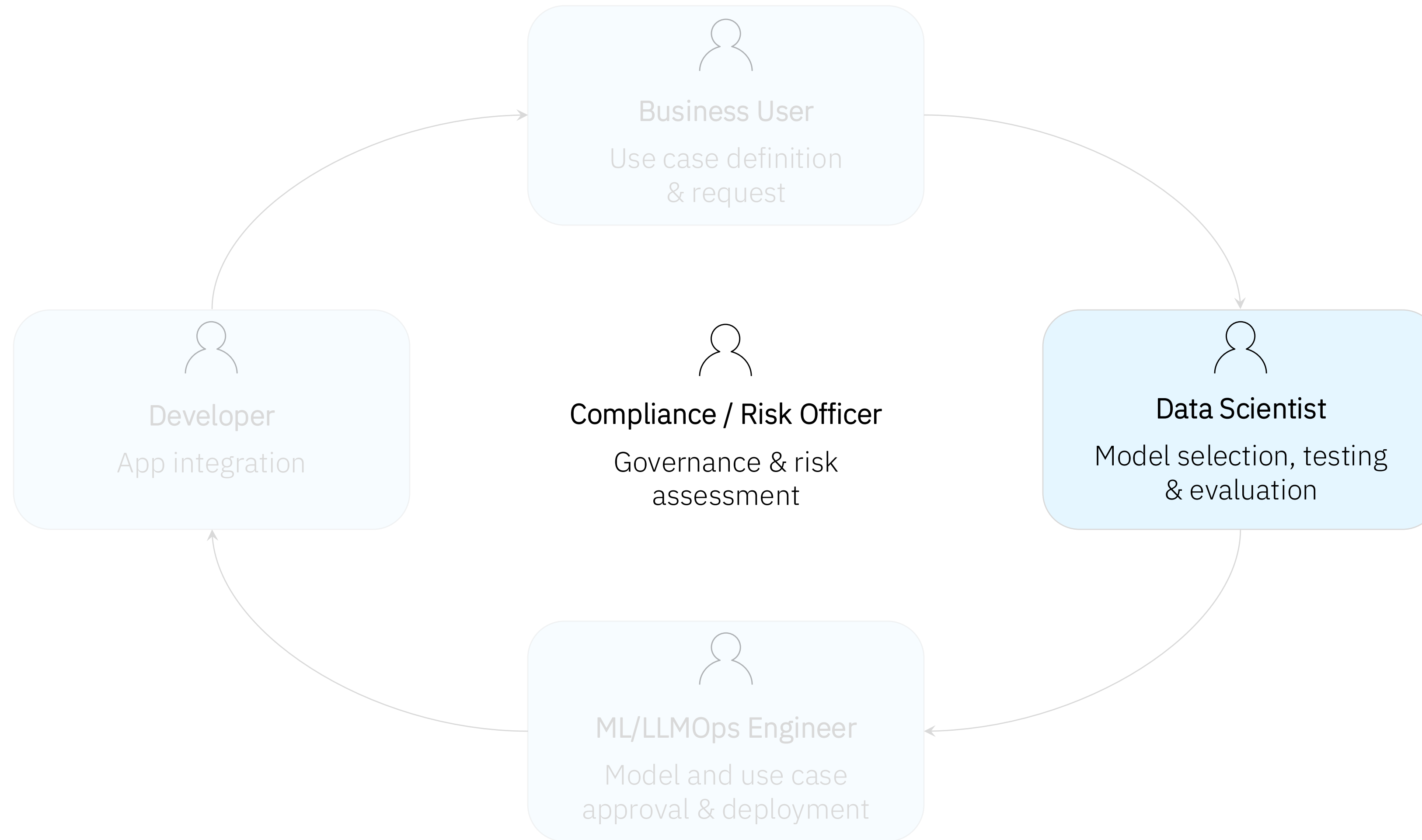
High Risk Models: 6

### Use case by lifecycle phase

- Awaiting approval


# AI development & AI governance

## Summarization of car insurance claims



# AI development

Data Scientist experiments with different prompts and foundation models to evaluate if the desired information can be derived from the claim

  
**Data Scientist**  
Model selection, testing & evaluation

Welcome back, Andreas

Train, deploy, validate, and govern AI models responsibly.

[Customize my journey](#)

Open in: insurance-demo-v4

[...]

Chat and build prompts with foundation models

Start chatting...

[Open Prompt Lab](#)

Tune a foundation model with labeled data

with Tuning Studio

Request or track models in AI use cases

with AI governance

Collapse ^

## Jump back in

Recently visited pages

insurance-demo-v4 / [Prompt Lab](#)


Collapse Discover section ^

## Discover

**Resource hub**


Foundation models

**Featured**



# AI development

Data Scientist experiments with different prompts and foundation models to evaluate if the desired information can be derived from the claim

  
**Data Scientist**  
Model selection, testing & evaluation

Welcome back, Andreas

Train, deploy, validate, and govern AI models responsibly.

[Customize my journey](#)

Open in: insurance-demo-v4


[...]

**Chat and build prompts with foundation models**

Start chatting... ▶

[Open Prompt Lab](#)

  
Tune a foundation model with labeled data  
  
with Tuning Studio

  
Request or track models in AI use cases  
  
with AI governance

Collapse ^

Jump back in  
Recently visited pages

insurance-demo-v4 /  
Prompt Lab

Collapse Discover section ^

Discover

### Resource hub


 Foundation models

### Featured



# AI development

Data Scientist experiments with different prompts and foundation models to evaluate if the desired information can be derived from the claim

  
**Data Scientist**  
Model selection, testing & evaluation

IBM watsonx 2613574 - itz-marketsaas-... Dallas AS

Projects / insurance-demo-v4 / insurance-claim-summarization-... [..]

Chat Structured Freeform AI Model: granite-13b-chat-v2 {#} TXT </> X

**Set up** ^

Instruction (optional) ⓘ

<|system|> You are Granite Chat, an AI language model developed by IBM. You are a cautious assistant. You carefully follow instructions. You are helpful and harmless and you follow ethical ...

Examples (optional) ⓘ

Input:	Output:
Enter your example input here.	Enter your desired output.

Add example +

**Try** ^

Test your prompt ⓘ

Input:	Output:
[Document] {input} [End] Summarize the following insurance claim input. Focus on the car and the damage. Make the ...	The insured's 2004 Honda Civic was stolen on November 1st, 2023, and crashed into a tree, resulting in significant front-end damage. The insured has filed a claim for the damages, and a police report and witness statements have been provided as evidence of the theft and accident.

Stop reason: End of sequence token encountered  
Tokens: 205 input + 60 generated = 265 out of 8192  
Time: 2.2 seconds

Clear output ↻ **Generate** →

**Model parameters**

Decoding

Greedy  Sampling ⓘ

Repetition penalty

1  2 1

**Stopping criteria** ⓘ

Stop sequences


Min tokens Max tokens

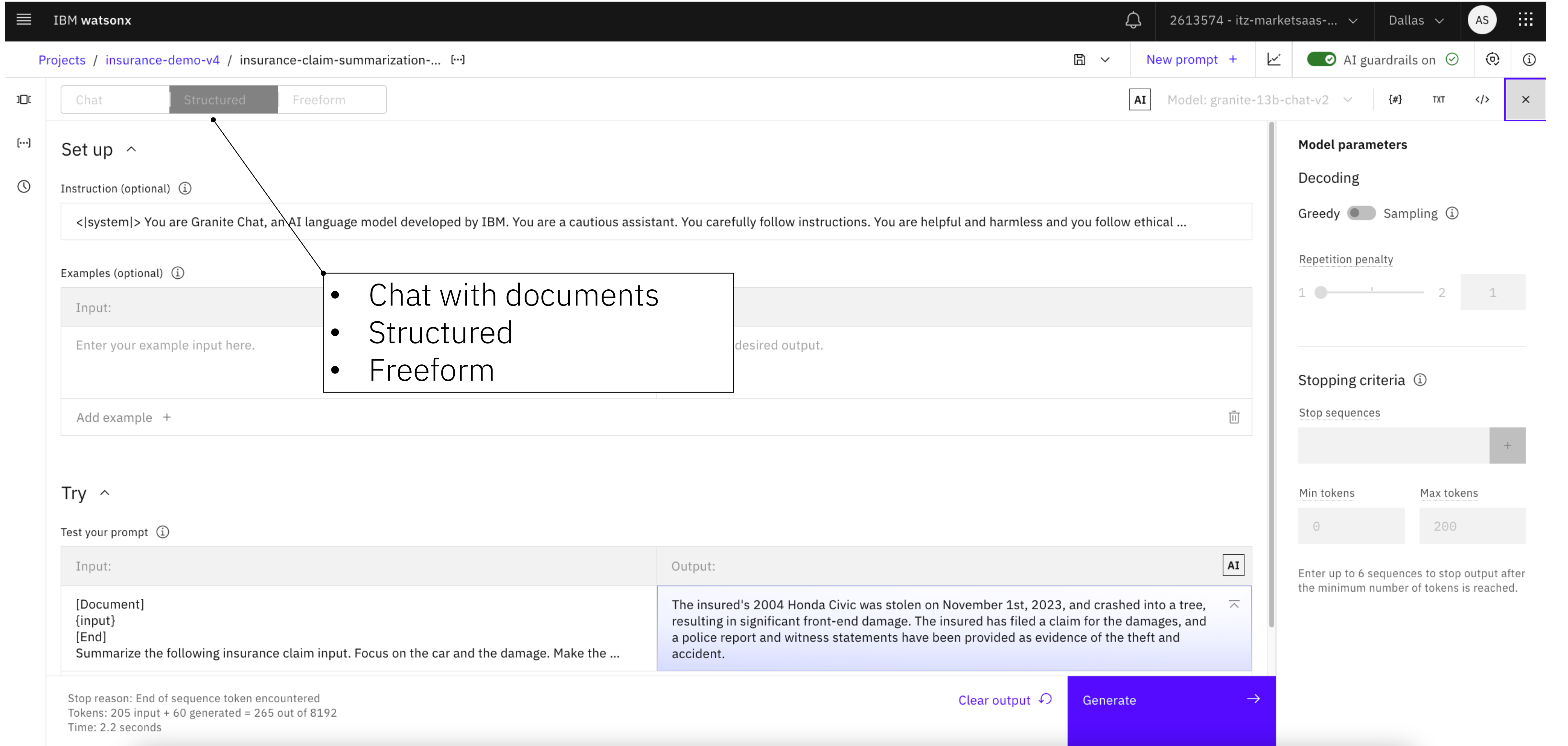
0 200

Enter up to 6 sequences to stop output after the minimum number of tokens is reached.

# AI development

Data Scientist experiments with different prompts and foundation models to evaluate if the desired information can be derived from the claim

  
**Data Scientist**  
Model selection, testing & evaluation



The screenshot shows the IBM watsonx AI interface. At the top, the user is logged in as 'Data Scientist' with the role 'Model selection, testing & evaluation'. The interface is for a project named 'insurance-demo-v4' with a sub-project 'insurance-claim-summarization-...'. The 'Structured' tab is selected, and the model is 'granite-13b-chat-v2'. The 'AI guardrails' are turned on. The 'Set up' section shows an instruction: '<|system|> You are Granite Chat, an AI language model developed by IBM. You are a cautious assistant. You carefully follow instructions. You are helpful and harmless and you follow ethical ...'. Below this, there are 'Examples (optional)' with an 'Input' field containing '[Document] {input} [End] Summarize the following insurance claim input. Focus on the car and the damage. Make the ...' and an 'Output' field containing 'The insured's 2004 Honda Civic was stolen on November 1st, 2023, and crashed into a tree, resulting in significant front-end damage. The insured has filed a claim for the damages, and a police report and witness statements have been provided as evidence of the theft and accident.' The 'Model parameters' section on the right shows 'Decoding' options: 'Greedy' is selected, 'Repetition penalty' is set to 1, 'Min tokens' is 0, and 'Max tokens' is 200. The 'Stopping criteria' section shows 'Stop sequences' and a note: 'Enter up to 6 sequences to stop output after the minimum number of tokens is reached.' At the bottom, there is a 'Generate' button and a 'Clear output' button. The status bar at the bottom left shows: 'Stop reason: End of sequence token encountered', 'Tokens: 205 input + 60 generated = 265 out of 8192', and 'Time: 2.2 seconds'.

IBM watsonx

2613574 - itz-marketsaas-... Dallas AS

Projects / insurance-demo-v4 / insurance-claim-summarization-...

Chat Structured Freeform

AI Model: granite-13b-chat-v2 {#} TXT </> X

Set up ^

Instruction (optional) ⓘ

<|system|> You are Granite Chat, an AI language model developed by IBM. You are a cautious assistant. You carefully follow instructions. You are helpful and harmless and you follow ethical ...

Examples (optional) ⓘ

Input:

Enter your example input here.

- Chat with documents
- Structured
- Freeform

desired output.

Add example +

Try ^

Test your prompt ⓘ

Input:

[Document] {input} [End] Summarize the following insurance claim input. Focus on the car and the damage. Make the ...

Output:

AI

The insured's 2004 Honda Civic was stolen on November 1st, 2023, and crashed into a tree, resulting in significant front-end damage. The insured has filed a claim for the damages, and a police report and witness statements have been provided as evidence of the theft and accident.

Model parameters

Decoding

Greedy  Sampling ⓘ

Repetition penalty

1 2 1

Stopping criteria ⓘ

Stop sequences

Min tokens Max tokens

0 200

Enter up to 6 sequences to stop output after the minimum number of tokens is reached.


Stop reason: End of sequence token encountered  
Tokens: 205 input + 60 generated = 265 out of 8192  
Time: 2.2 seconds

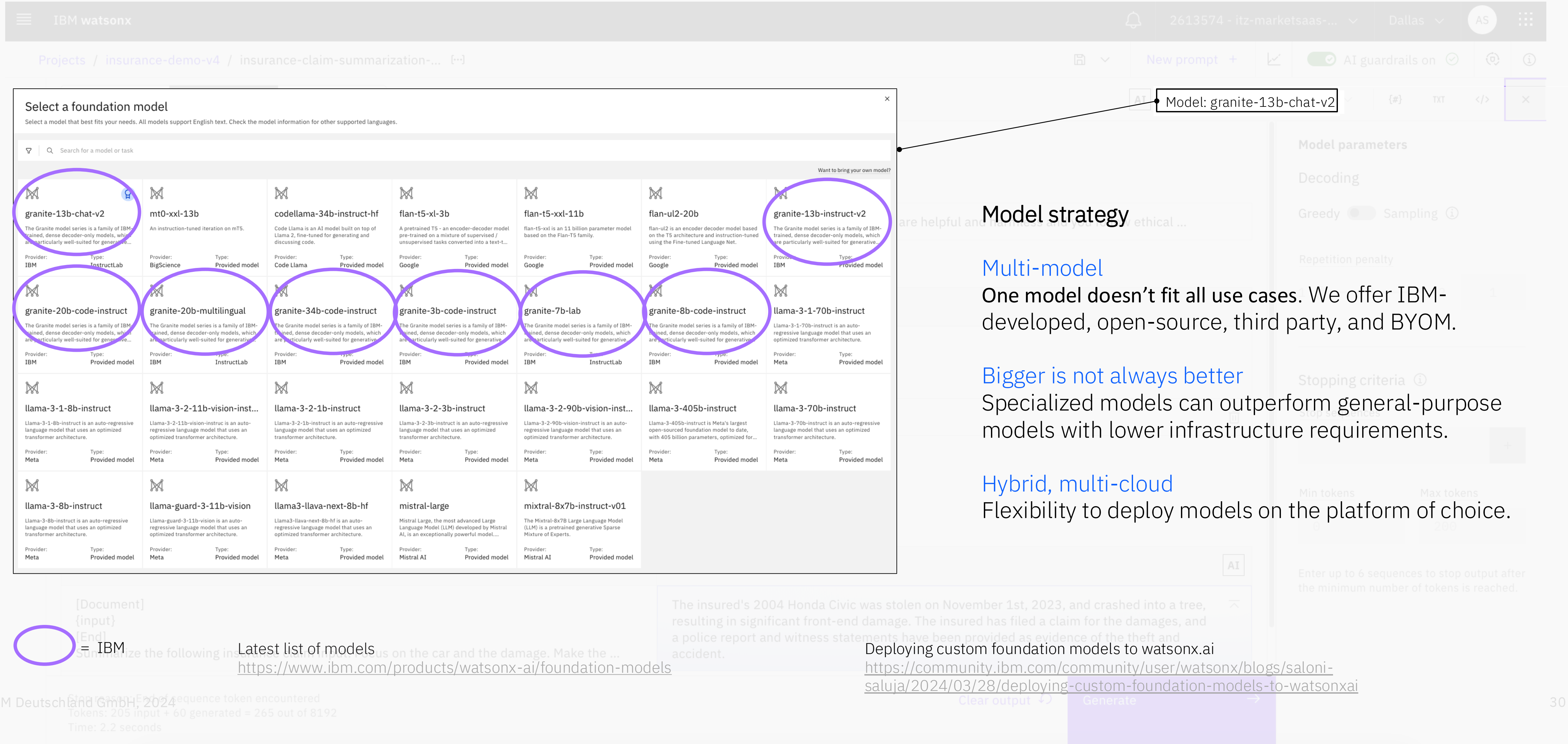
Clear output ↻

Generate →

# AI development

Data Scientist experiments with different prompts and foundation models to evaluate if the desired information can be derived from the claim

  
**Data Scientist**  
Model selection, testing & evaluation



The screenshot shows the IBM watsonx interface. At the top, there's a navigation bar with 'IBM watsonx', a search bar, and user information. Below that, a breadcrumb trail shows 'Projects / insurance-demo-v4 / insurance-claim-summarization-...'. The main area is a 'Select a foundation model' dialog box. This dialog has a search bar and a grid of model cards. Several cards are circled in purple, including 'granite-13b-chat-v2', 'granite-20b-code-instruct', 'granite-20b-multilingual', 'granite-34b-code-instruct', 'granite-3b-code-instruct', 'granite-7b-lab', 'granite-8b-code-instruct', and 'granite-13b-instruct-v2'. A callout box points to the 'granite-13b-chat-v2' card with the text 'Model: granite-13b-chat-v2'. To the right of the dialog, there's a 'Model parameters' section with options for 'Decoding' (Greedy/Sampling), 'Repetition penalty', and 'Stopping criteria'. Below the dialog, a chat window shows a prompt: '[Document] {input} [End] Summarize the following information on the car and the damage. Make the ...'. The chat output shows a summary of a car accident claim. At the bottom, there's a status bar with 'IBM Deutschland GmbH, 2024', 'Tokens: 205 input + 60 generated = 265 out of 8192', and 'Time: 2.2 seconds'. A 'Generate' button is visible at the bottom right.

## Model strategy

### Multi-model


One model doesn't fit all use cases. We offer IBM-developed, open-source, third party, and BYOM.

### Bigger is not always better

Specialized models can outperform general-purpose models with lower infrastructure requirements.

### Hybrid, multi-cloud

Flexibility to deploy models on the platform of choice.

 = IBM


Latest list of models  
<https://www.ibm.com/products/watsonx-ai/foundation-models>


Deploying custom foundation models to watsonx.ai  
<https://community.ibm.com/community/user/watsonx/blogs/saloni-saluja/2024/03/28/deploying-custom-foundation-models-to-watsonxai>

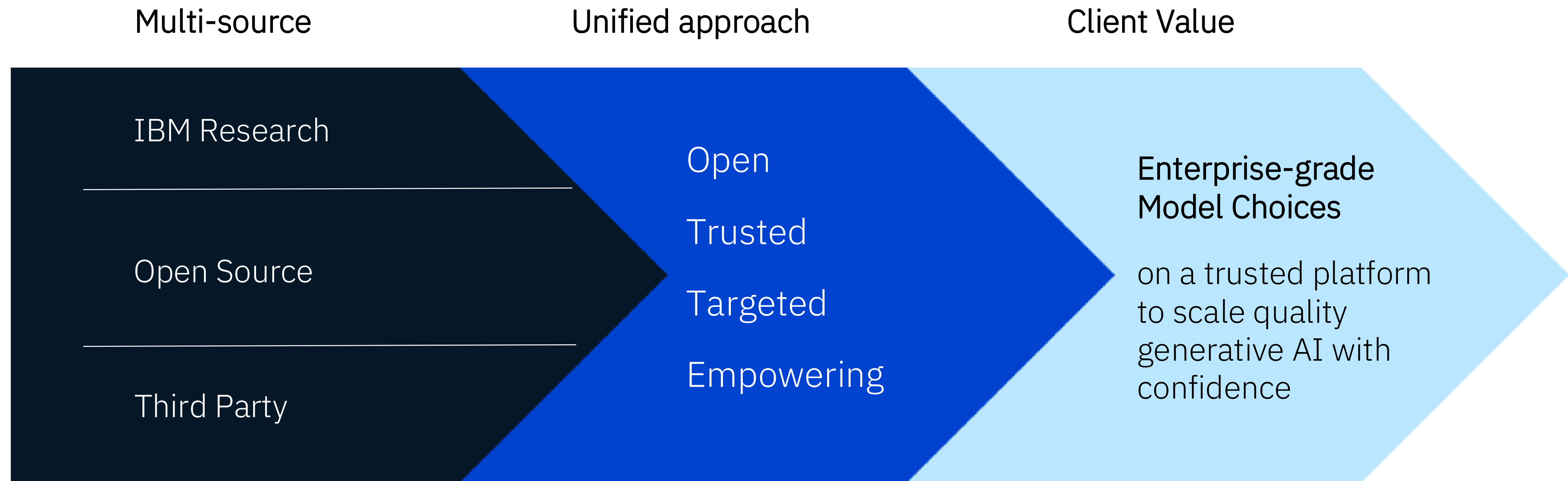
# AI development: Foundation models

## Open, multi-model approach to delivering enterprise-grade foundation models

IBM manages to provide enterprise clients wide selection of enterprise-grade model choices while maintaining quality of generation.

  
**Data Scientist**  
Model selection, testing & evaluation

  
**Compliance / Risk Officer**  
Governance & risk assessment



**watsonx**




**InstructLab**

**Red Hat**

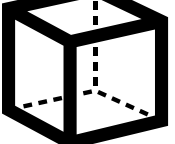

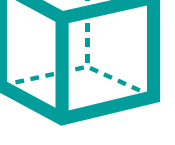
# AI development: Foundation models

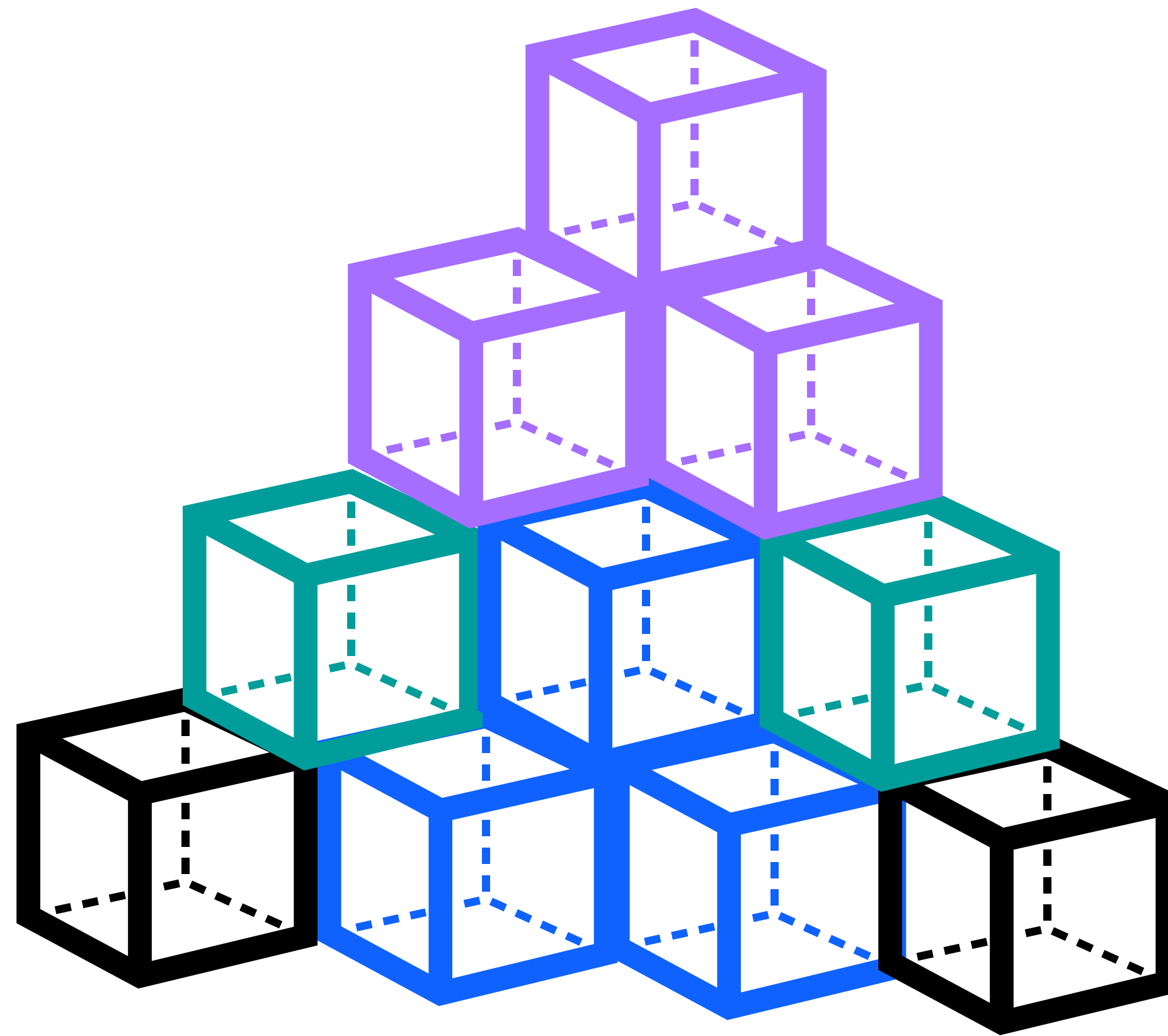
## Approach of selecting third-party / open-source models

  
**Data Scientist**  
Model selection, testing  
& evaluation

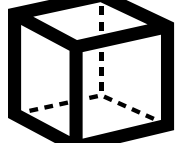

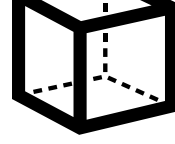

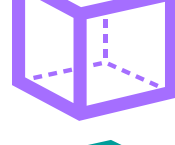
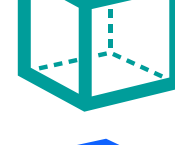
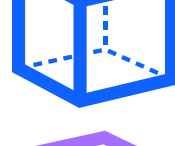
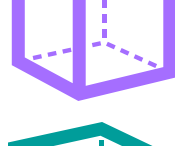

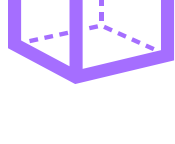
  
**Compliance / Risk Officer**  
Governance & risk  
assessment

### Considerations


-  Model performance
-  Research
-  Ethics
-  Legal and data



### Workflow

-  1 Review technical papers
-  2 Model Information
-  3 Performance Benchmark
-  4 Internal IBM use
-  5 Commercial Applicability
-  6 Licensing
-  7 Reputation
-  8 Use Case Alignment
-  9 Training Data
-  10 Infrastructure

# AI development: Foundation models: IBM ensures integrity of Granite models <sup>1, 2</sup> through comprehensive data management across the training phase to build trusted models

  
**Data Scientist**  
Model selection, testing & evaluation

  
**Compliance / Risk Officer**  
Governance & risk assessment

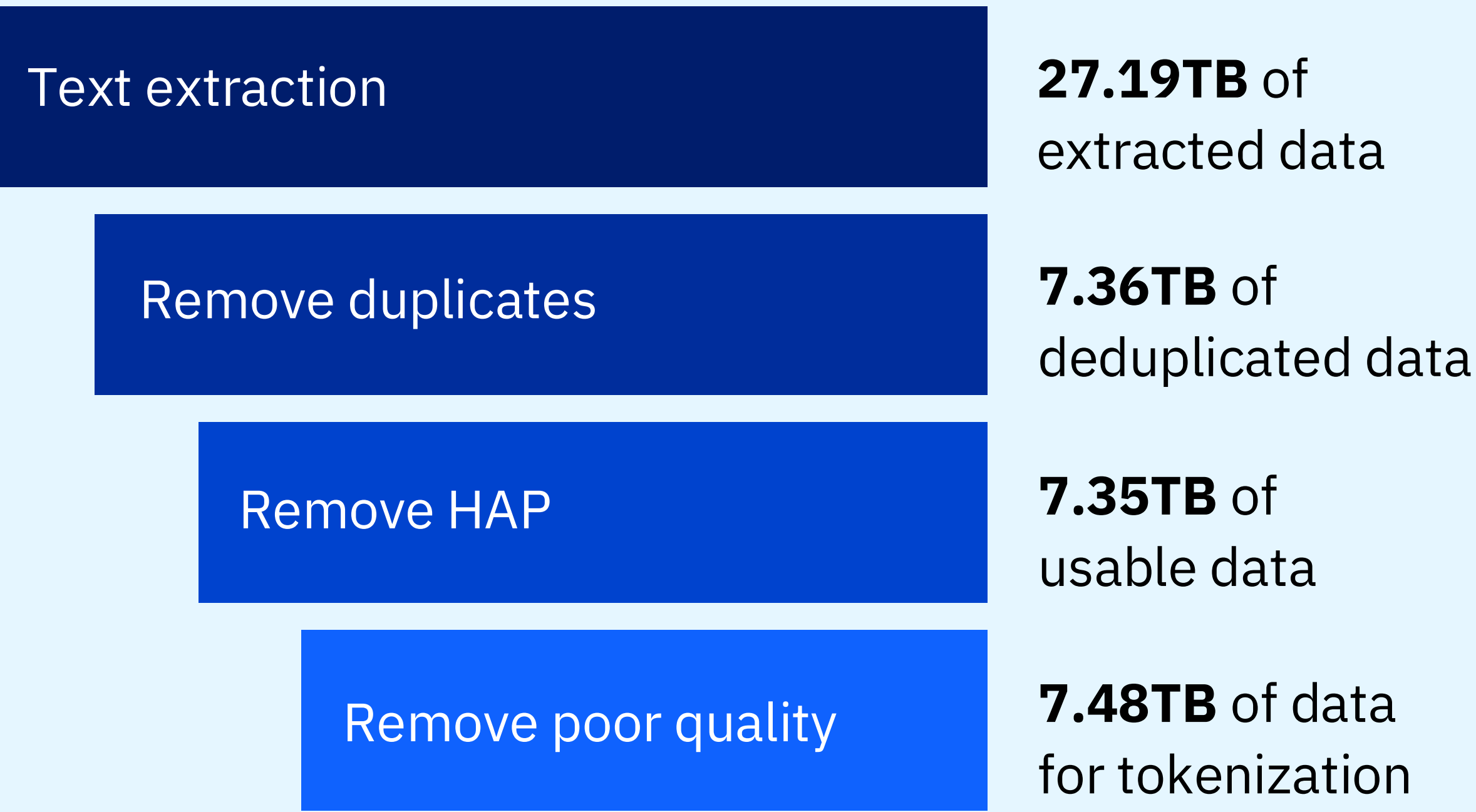
IBM's approach to AI model development is [grounded in core principles of trust and transparency.](#)

You can use them for...

- Summarization
- Insight extraction & classification
- Retrieval-Augmented Generation

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

- Internet
- Academic
- Code
- Legal
- Finance



2.5 trillion tokens used for training IBM Granite base models

TB = Terabytes

<sup>1</sup> Granite Foundation Models ([Research paper](#))

<sup>2</sup> [The Foundation Model Transparency Index](#)

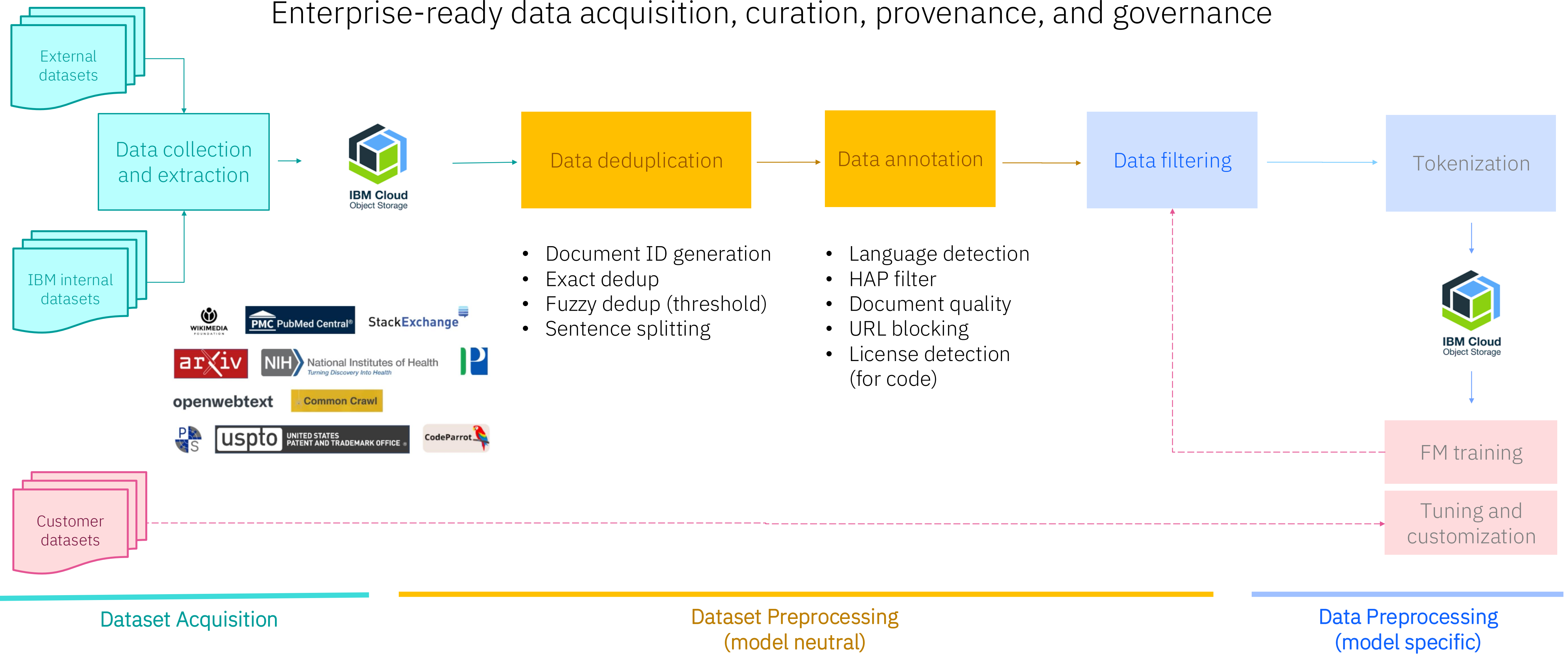
# AI development: Foundation models

## Quality data for Granite models

**Data Scientist**  
Model selection, testing & evaluation

**Compliance / Risk Officer**  
Governance & risk assessment


### Enterprise-ready data acquisition, curation, provenance, and governance



# AI development: Foundation models

Granite models are targeted at specific industry domains like finance and legal

Based on IBM Research internal development tests & evaluations, the Granite-13B models demonstrates superior acumen across different financial tasks



**Data Scientist**  
Model selection, testing & evaluation



**Compliance / Risk Officer**  
Governance & risk assessment

Task	Task Description	Dataset	Dataset Description	N-shot Prompt	Metric
Sentiment Classification	2 classes	Earnings Call Transcripts [36]	Earnings call transcripts, the related stock prices and the sector index in terms of volume	5-shot	Weighted F1
Classification	9 classes	News Headline [37]	The gold commodity news annotated into various dimensions	5-shot	Weighted F1
Named Entity Recognition	4 numerical entities	Credit Risk Assessment (NER) [33]	Eight financial agreements (totalling 54,256 words) from SEC filings were manually annotated for entity types: location, organization person and miscellaneous	20-shot	Entity F-1
	4522 numerical entities	KPI-Edgar [38]	A dataset for Joint Named Entity Recognition and Relation Extraction building on financial reports uploaded to the Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system, where the main objective is to extract Key Performance Indicators (KPIs) from financial documents and link them to their numerical values and other attributes	20-shot	Modified Adjusted F1
	139 numerical entities	FiNER-139 [39]	1.1M sentences annotated with extensive Business Reporting Language (XBRL) tags extracted from annual and quarterly reports of publicly-traded companies in the US, focusing on numeric tokens, with the correct tag depending mostly on context, not the token itself.	10-shot	Entity F1
Question Answering	Document relevance ranking	Opinion-based QA (FiQA) [40]	Text documents from different financial data sources (microblogs, reports, news) for ranking document relevance based on opinionated questions, targeting mined opinions and their respective entities, aspects, sentiment polarity and opinion holder.	5-shot	RR@10
	3 classes	Sentiment Analysis (FiQA SA) [40]	Text instances in the financial domain (microblog message, news statement or headline) for detecting the target aspects which are mentioned in the text (from a pre-defined list of aspect classes) and predict the sentiment score for each of the mentioned targets.	5-shot	Weighted F1
	Ranking	Insurance QA [41]	Questions from real world users and answers with high quality composed by professionals with deep domain knowledge collected from the website Insurance Library <sup>4</sup>	5-shot	RR@5
	Exact value match	Chain of Numeric Reasoning (ConvFinQA) [42]	Multi-turn conversational finance question answering data for exploring the chain of numerical reasoning	1-shot	Accuracy
Summarization	Long documents	Financial text summarization (EDT) [43]	303893 news articles range from March 2020 to May 2021 for abstractive text summarization	5-shot	Rouge-L


# AI development: Foundation models

## IBM stands behind its Granite models with industry-leading IP indemnification

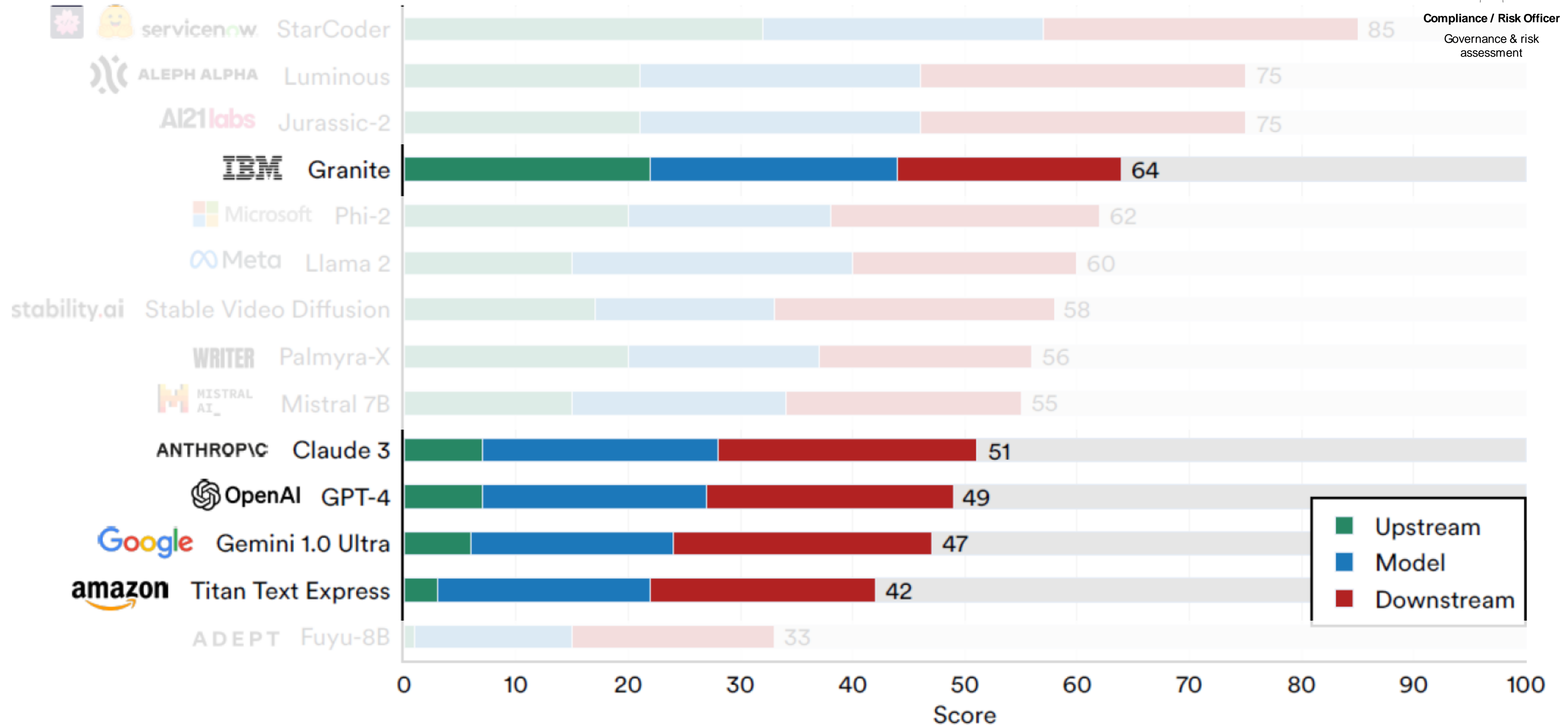
- IBM clients **benefit from** industry-leading **legal protection for Granite models**, backed by decades of experience navigating legal frameworks in tech
- IBM offers clients peace of mind, **eliminating the need for customer indemnification for model usage**
- There is **no cap on IBM's IP indemnification liability**, ensuring robust protection for clients




# Foundation Model Transparency Index Scores by Domain, May 2024

Source: May 2024 Foundation Model Transparency Index

 **Data Scientist**  
Model selection, testing & evaluation


 **Compliance / Risk Officer**  
Governance & risk assessment



 Upstream  
 Model  
 Downstream

# AI development: Foundation models

## Granite model benefits



**Data Scientist**  
Model selection, testing  
& evaluation



**Compliance / Risk Officer**  
Governance & risk  
assessment

### Dataset quality

- Trained on datasets that meet IBM's rigorous data governance, risk, and compliance criteria
- Copyright, licensed, and Hate, Abusive, Profanity (HAP) contents removed

### Trust and integrity

- Created based on IBM's AI Ethics Principles and IBM's Office of Privacy & Responsible Technology
- Indemnified by IBM<sup>1</sup>

### Enterprise-grade

- Built specifically for enterprise
- Governed by rules and safeguards
- Source data is highly curated removing harmful contents

### Scalability

- Size ranges from 3 to 34 billion parameters
- Provides scalability to different use cases

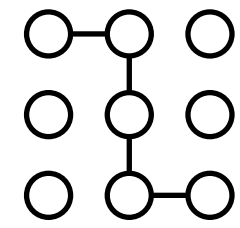
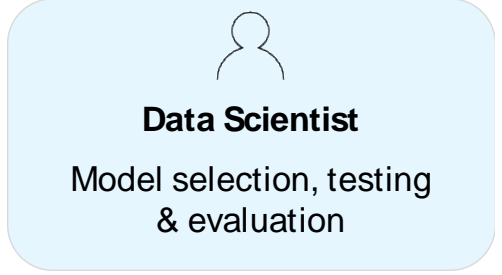
### Customization

- Clients can tune and guide the models with their enterprise data
- Can adapt models easily to multiple downstream tasks

### Wide applicability

- Trained in 116 programming languages
- Applicable to a variety of industries and use cases

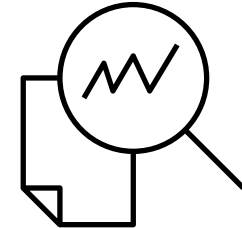
# AI development: Foundation models Granite everywhere



## Open-Source communities (Hugging Face & GitHub)

For generative AI model testing & experimentation directly on your laptop or workstation to collaborate on models, datasets, or applications

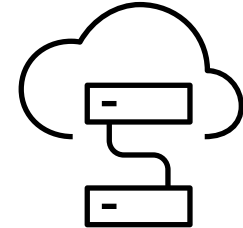
Access new modalities before anyone else, all Granite models released under Apache 2.0 license i.e., code, time series, language, and geospatial



## InstructLab

Open-source, model-agnostic toolkit, enables individuals to contribute knowledge and skills to foundation models, facilitating a new “cheaper” approach to alignment tuning that leverages both human and model evaluations for quality assurance

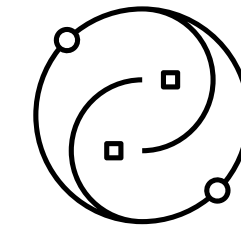
Easily generate synthetic data using a teacher model to train future version of the base model



## Red Hat Enterprise Linux AI (RHEL AI)

Bring together the Granite family of open sourced-licensed LLMs, InstructLab model alignment toolkit, a bootable image of Red Hat Enterprise Linux, and enterprise-grade technical support

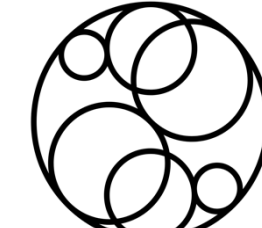
Model IP indemnification for use of Granite open sourced-licensed models provided by Red Hat accessible in RHEL AI



## watsonx.ai

Enterprise-grade AI studio that helps AI developers innovate with all the tools, runtimes, and APIs to deploy AI applications at enterprise scale with lifecycle governance, customizable model choices, and tooling to build AI assistants and agents

IP indemnification for use of Granite models on watsonx.ai, including proprietary Granite language models to watsonx.ai



## watsonx Assistants

Access Granite specialized models that are fine-tuned and embedded directly into AI applications to solve specific use cases i.e., code generation from COBOL to Java, code completion, etc. for a range of supported languages


Granite models cannot be downloaded or manipulated outside the application

# AI development: Foundation models

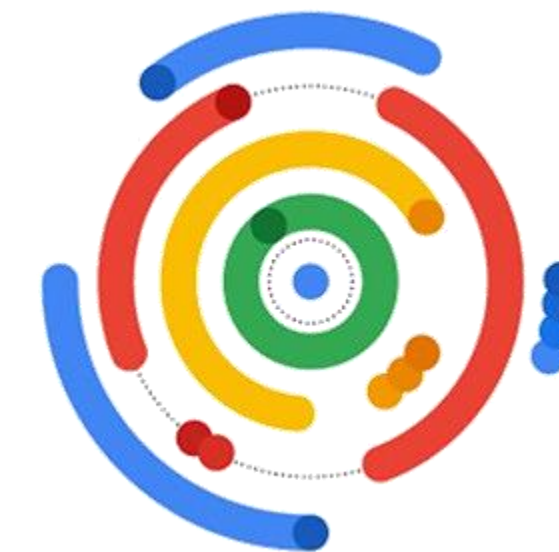
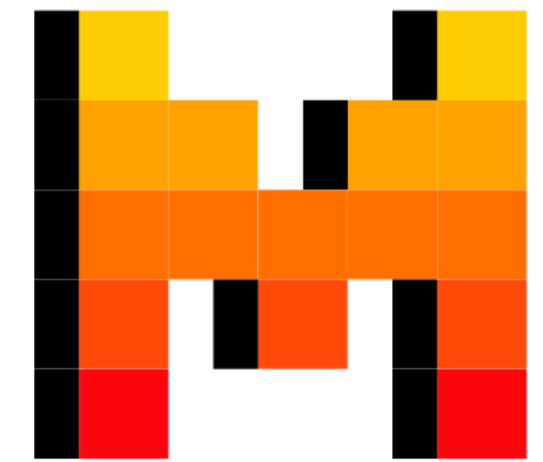
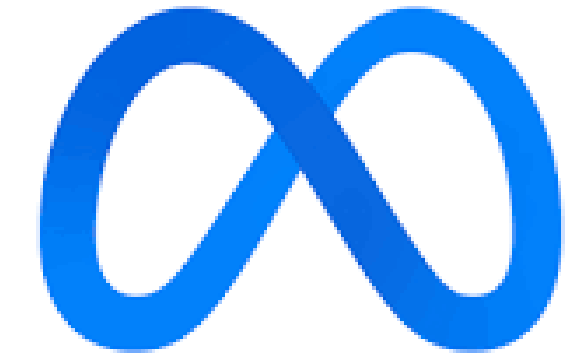
## Open-source and third-party models

IBM partners with Hugging Face and other third-party vendors to make available top open-source models to watsonx.ai

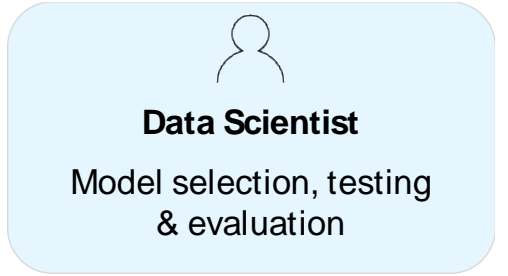
- Many models on the Hugging Face leaderboard
- Focus on variety – not size
- Localized
- 3<sup>rd</sup> party vendors: Meta, Google, Mistral, and more
- Coding support (SDK, Notebooks, API and code samples) for all models
- Model details, availability (some are geo-specific) and sample prompts are provided as references<sup>1</sup>

  
**Data Scientist**  
Model selection, testing  
& evaluation

  
**Compliance / Risk Officer**  
Governance & risk  
assessment



# AI development: Foundation models IBM models and third-party models



Sometimes it's useful to use third party Generative AIs for specific use cases.  
The extensive library of open-source models [opens more possibilities](#).

## IBM Models

- **Foundation models** trained with big data sets of unlabeled data curated by IBM

Smaller models (3b to 34b parameter models):

- Works well on domain specific knowledge
- Easy to fine-tune
- Easier to train and run
- Less expensive
- More efficient

## Third Party Models

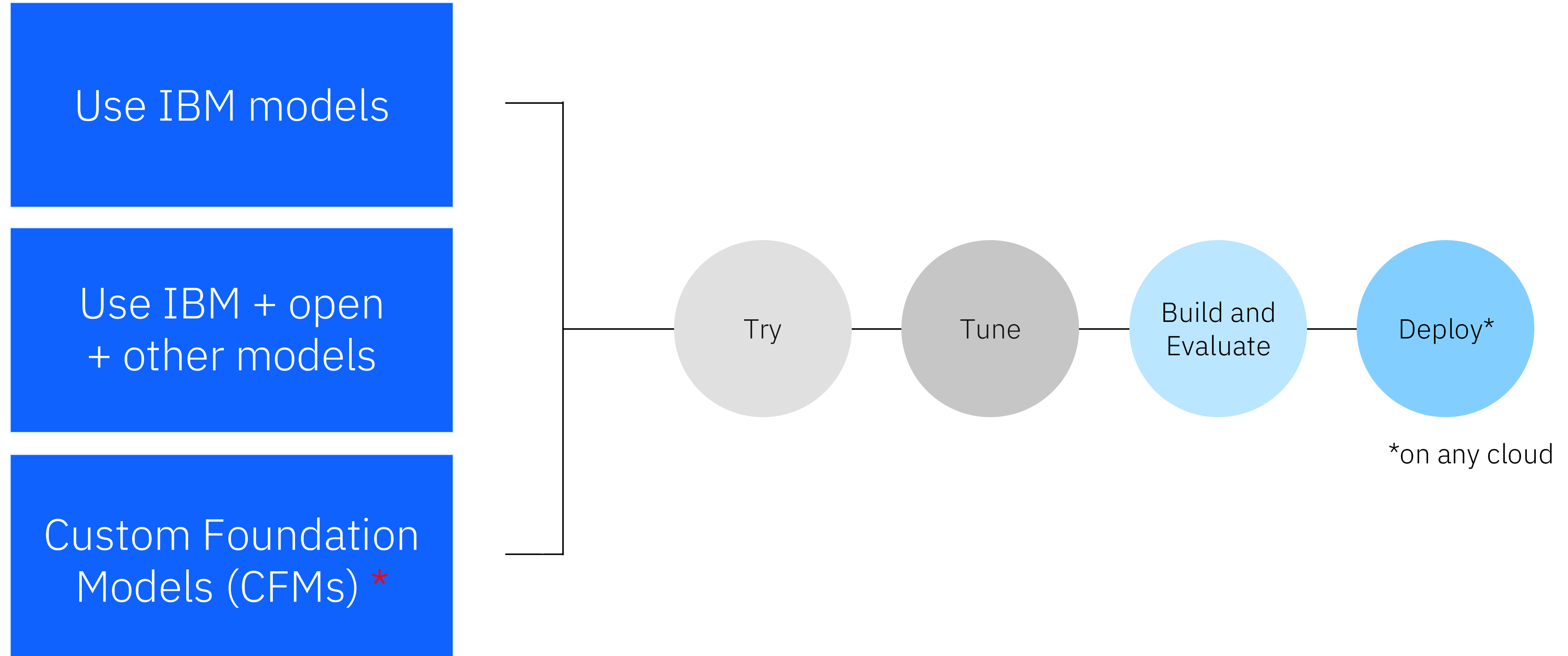
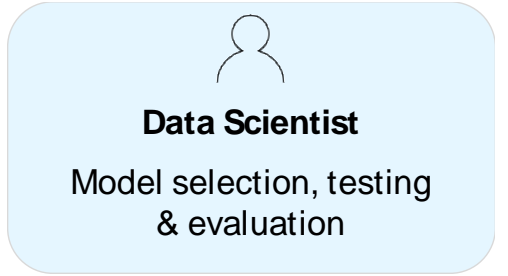
- **Foundation models** trained with big data sets of unlabeled data curated by third parties

Larger models:

- Works well out-of-the-box on various NLP tasks
- Energy intensive to train and run
- More expensive
- Less efficient

# AI development: Foundation models

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



\* Note: Previously known as Bring Your Own Model (BYOM) is now called Custom Foundation Models (CFMs)

# AI development: Foundation models

## Why custom foundation models?

### Clients may have...

- worked with other models for their use cases
- tuned/optimized models for their use cases

### Clients want...

- to leverage watsonx.ai to deploy, manage, and update their deployment

### But also want...

- to use a model not currently available in the watsonx.ai model catalog

### Even more flexibility


- 1000+ more model choices
- Access to more language support
- Bring your own fine-tuned model catered to an industry or business domain

### Wide range of supported architectures

- bloom
- codegen
- falcon
- gpt\_bigcode
- gpt\_neox
- gptj
- llama

# AI development

Data Scientist experiments with different prompts and foundation models to evaluate if the desired information can be derived from the claim

  
**Data Scientist**  
Model selection, testing & evaluation

IBM watsonx 2613574 - itz-marketsaas-... Dallas AS

Projects / insurance-demo-v4 / insurance-claim-summarization-... [..]

Chat Structured Freeform AI Model: granite-13b-chat-v2 {#} TXT </> X

**Set up** ^

Instruction (optional) ⓘ

<|system|> You are Granite Chat, an AI language model developed by IBM. You are a cautious assistant. You carefully follow instructions. You are helpful and harmless and you follow ethical ...

Examples (optional) ⓘ

Input:	Output:
Enter your example input here.	Enter your desired output.

Add example +

**Try** ^

Test your prompt ⓘ

Input:	Output:
[Document] {input} [End] Summarize the following insurance claim input. Focus on the car and the damage. Make the ...	The insured's 2004 Honda Civic was stolen on November 1st, 2023, and crashed into a tree, resulting in significant front-end damage. The insured has filed a claim for the damages, and a police report and witness statements have been provided as evidence of the theft and accident.

Stop reason: End of sequence token encountered  
Tokens: 205 input + 60 generated = 265 out of 8192  
Time: 2.2 seconds

Clear output ↻ **Generate** →

**Model parameters**

Decoding

Greedy  Sampling ⓘ

Repetition penalty

1  2 1

**Stopping criteria** ⓘ

Stop sequences

+

Min tokens  Max tokens


Enter up to 6 sequences to stop output after the minimum number of tokens is reached.

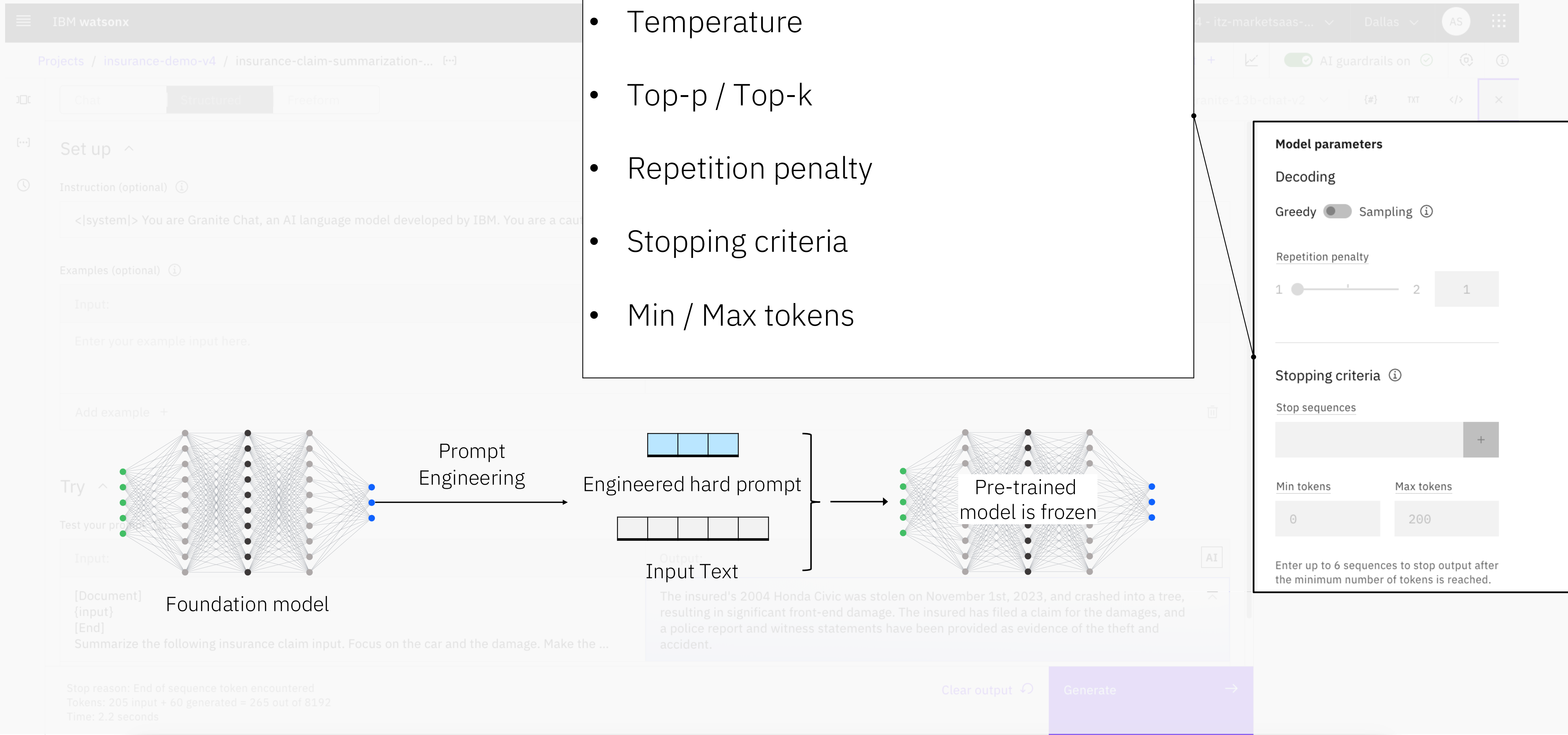
# AI development

Data Scientist experiments with different models to evaluate if the desired information can be extracted from the data.

## Model parameters

- Greedy vs. Sampling
- Temperature
- Top-p / Top-k
- Repetition penalty
- Stopping criteria
- Min / Max tokens

  
**Data Scientist**  
Model selection, testing & evaluation



The screenshot shows the IBM watsonx interface for a project named 'insurance-claim-summarization-...'. The 'Model parameters' panel is open, showing the following settings:


- Decoding:** Greedy (selected), Sampling (unselected)
- Repetition penalty:** Slider set to 1
- Stopping criteria:** Stop sequences (empty), Min tokens (0), Max tokens (200)

A diagram at the bottom illustrates the workflow: a 'Foundation model' receives an 'Input' (a document snippet) and produces an 'Output' (a summarized text). This process is labeled 'Prompt Engineering' and 'Engineered hard prompt'. A note states 'Pre-trained model is frozen'.

Stop reason: End of sequence token encountered  
Tokens: 205 input + 60 generated = 265 out of 8192  
Time: 2.2 seconds

# AI development

Data Scientist experiments with different prompts and foundation models to evaluate if the desired information can be derived from the claim

  
**Data Scientist**  
Model selection, testing & evaluation

Welcome back, Andreas

Train, deploy, validate, and govern AI models responsibly.

[Customize my journey](#)

Open in: insurance-demo-v4

[...]

Chat and build prompts with foundation models

Start chatting...

[Open Prompt Lab](#)

Tune a foundation model with labeled data

with Tuning Studio

Request or track models in AI use cases

with AI governance

Collapse ^

## Jump back in

Recently visited pages

insurance-demo-v4 / [Prompt Lab](#)


Collapse Discover section ^

## Discover

**Resource hub**

Foundation models

**Featured**



### Prompt Tuning parameters

#### Batch size

The number of samples to work through at one time

#### Number of epochs

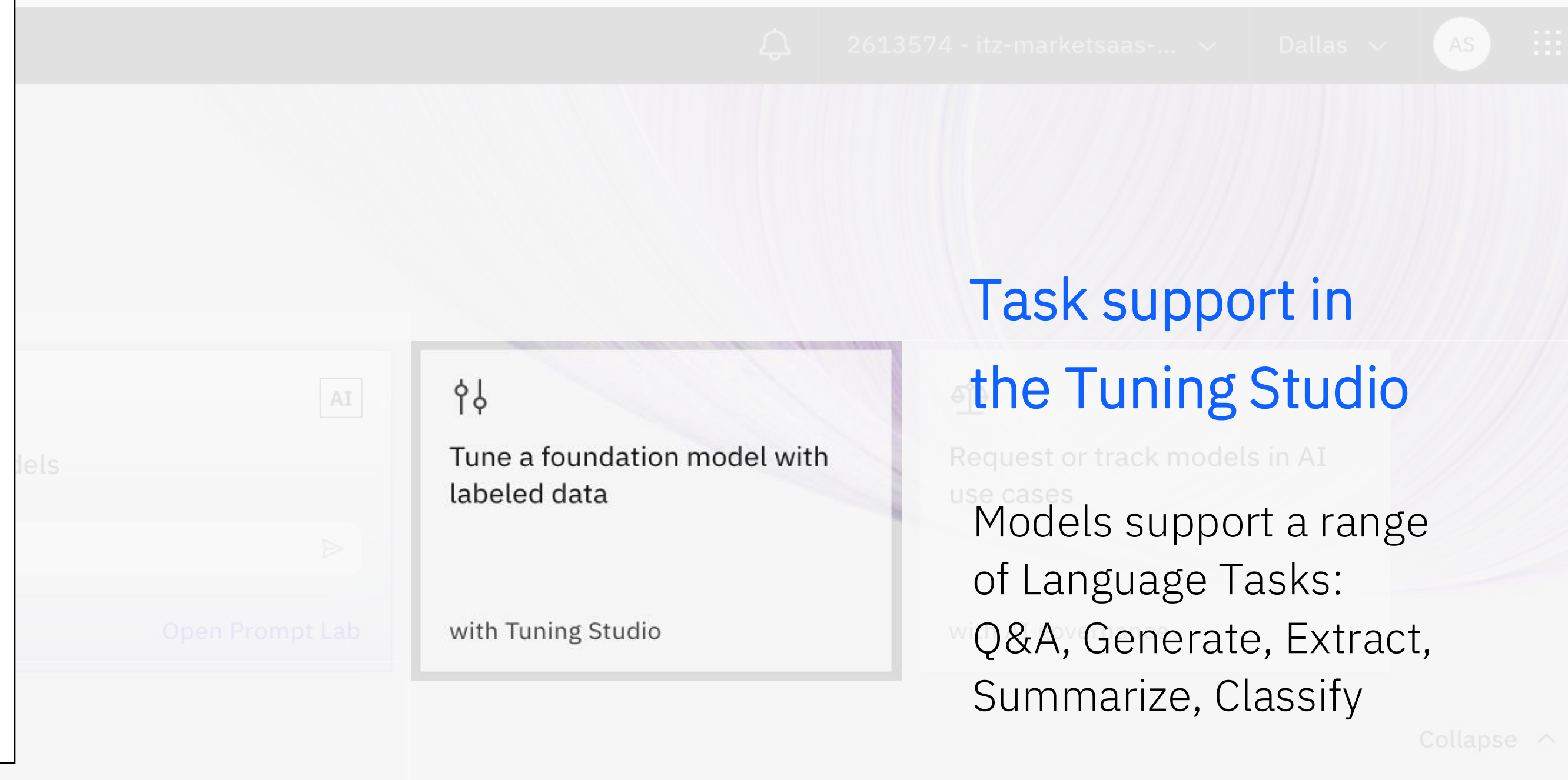
Number of times to cycle through the training data set

#### Learning rate

How fast the neural network will progress towards the optimal “learn” state.

#### Accumulation steps

The combined effect of the number of training steps to accumulate before updating the internal parameters



### Task support in the Tuning Studio

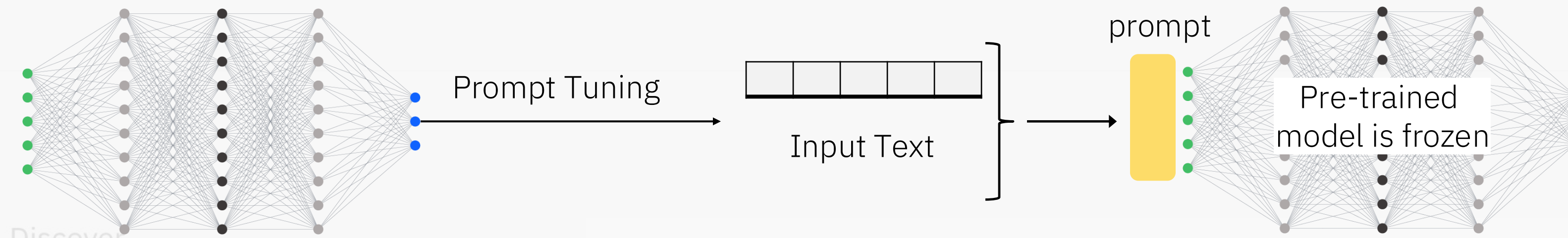
Request or track models in AI use cases  
Models support a range of Language Tasks: Q&A, Generate, Extract, Summarize, Classify

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

Jump back in  
Recently visited pages

insurance-demo-v4 / Prompt Lab



Discover


Resource hub


Foundation models

Featured



# Taxonomy-driven Small Language Model **instruction fine-tuning** based on synthetic data generation: Instruct Lab <sup>1</sup>

  
**Data Scientist**  
Model selection, testing  
& evaluation

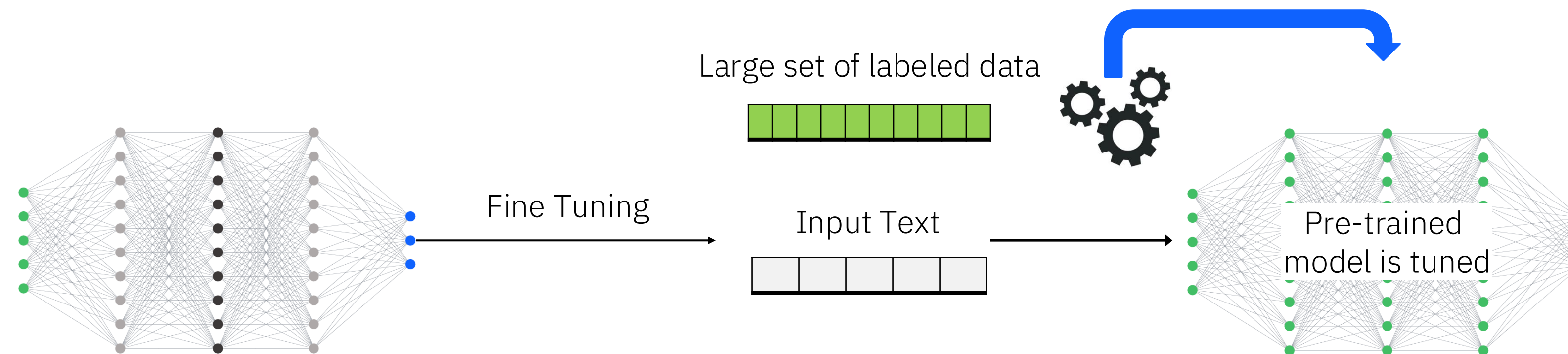
  
**Business User**  
Use case definition  
& request

IBM Instruct Lab intends to offer an intuitive way for a collective community style collaboration of skills and knowledge to Language Models to enable one model resulting in **rapid iterative release of purpose-built models**. The LAB-method is a **multi-phased** framework that allows adding **instruction-following** <sup>2</sup> abilities **without suffering from catastrophic forgetting**. It improves model performance not only on specific tasks, but on **following instructions in general**.

Making LLMs **learn like we humans do**:

with *Knowledge* and *skills*

→ *Adding content* to the taxonomy is a *business user* task




<sup>1</sup> LAB: Large-Scale Alignment for ChatBots

<sup>2</sup> What is instruction tuning?

# AI development

Data Scientist experiments with different prompts and foundation models to evaluate if the desired information can be derived from the claim

  
**Data Scientist**  
Model selection, testing & evaluation

IBM watsonx 2613574 - itz-marketsaas-... Dallas AS

Projects / insurance-demo-v4 / insurance-claim-summarization-... [..]

Chat Structured Freeform AI Model: granite-13b-chat-v2 {#} TXT </> X

**Set up** ^

Instruction (optional) ⓘ

<|system|> You are Granite Chat, an AI language model developed by IBM. You are a cautious assistant. You carefully follow instructions. You are helpful and harmless and you follow ethical ...

Examples (optional) ⓘ

Input:	Output:
Enter your example input here.	Enter your desired output.

Add example +

**Try** ^

Test your prompt ⓘ

Input:	Output:
[Document] {input} [End] Summarize the following insurance claim input. Focus on the car and the damage. Make the ...	The insured's 2004 Honda Civic was stolen on November 1st, 2023, and crashed into a tree, resulting in significant front-end damage. The insured has filed a claim for the damages, and a police report and witness statements have been provided as evidence of the theft and accident.

Stop reason: End of sequence token encountered  
Tokens: 205 input + 60 generated = 265 out of 8192  
Time: 2.2 seconds

Clear output ↻ Generate →

**Model parameters**

Decoding

Greedy  Sampling ⓘ

Repetition penalty

1  2 1

**Stopping criteria** ⓘ

Stop sequences


Min tokens Max tokens


0 200

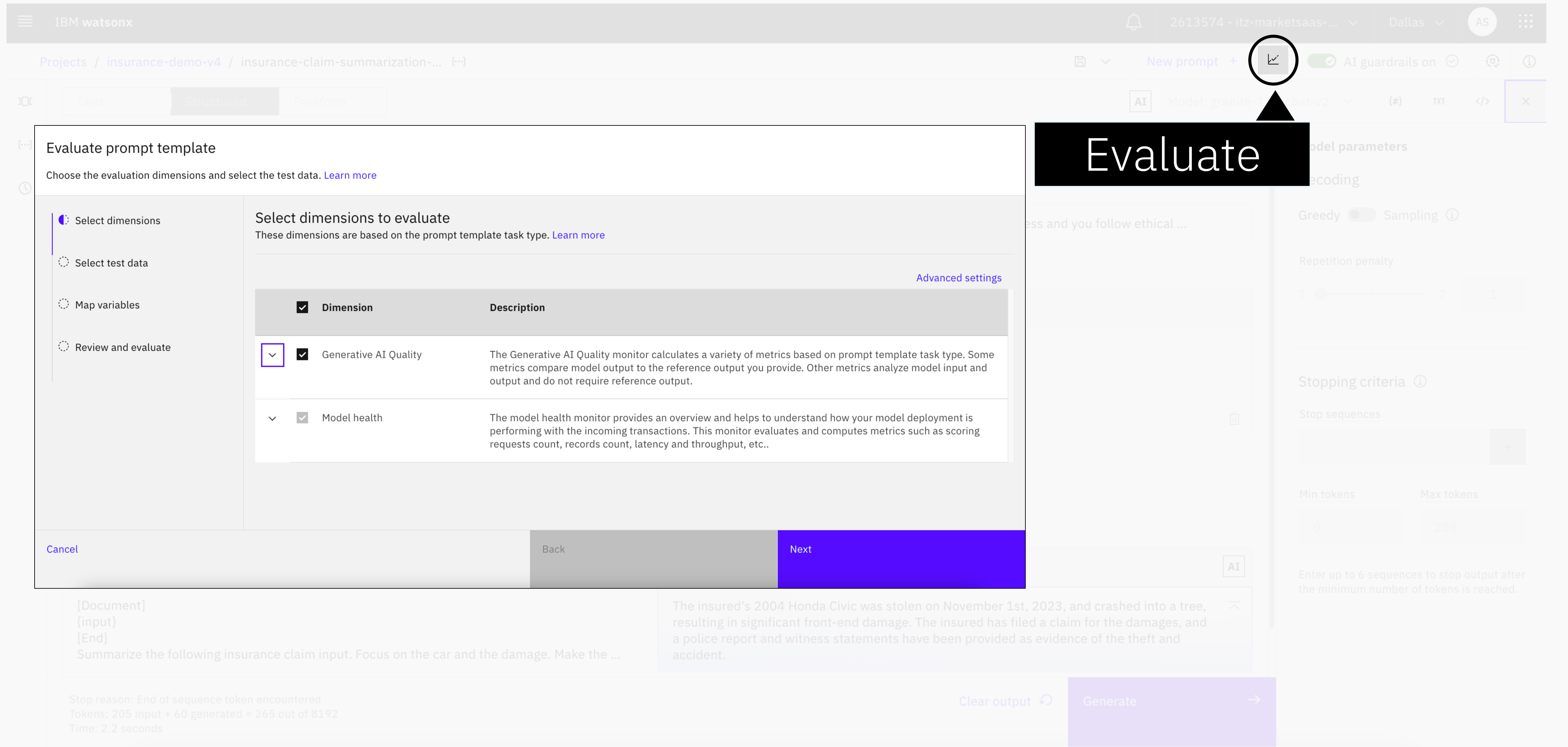
Enter up to 6 sequences to stop output after the minimum number of tokens is reached.

# AI development & AI governance

## Data Scientist evaluates prompt template and creates an AI Factsheet

  
**Business User**  
Use case definition  
& request

  
**Data Scientist**  
Model selection, testing  
& evaluation



IBM watsonx

Projects / insurance-demo-v4 / insurance-claim-summarization-...

Chat | Structured | Freeform

Evaluate prompt template

Choose the evaluation dimensions and select the test data. [Learn more](#)

Select dimensions

Select test data

Map variables

Review and evaluate

Select dimensions to evaluate

These dimensions are based on the prompt template task type. [Learn more](#)

Advanced settings

<input checked="" type="checkbox"/>	Dimension	Description
<input checked="" type="checkbox"/>	Generative AI Quality	The Generative AI Quality monitor calculates a variety of metrics based on prompt template task type. Some metrics compare model output to the reference output you provide. Other metrics analyze model input and output and do not require reference output.
<input checked="" type="checkbox"/>	Model health	The model health monitor provides an overview and helps to understand how your model deployment is performing with the incoming transactions. This monitor evaluates and computes metrics such as scoring requests count, records count, latency and throughput, etc..

Cancel | Back | Next

[Document]  
{input}  
[End]  
Summarize the following insurance claim input. Focus on the car and the damage. Make the ...

The insured's 2004 Honda Civic was stolen on November 1st, 2023, and crashed into a tree, resulting in significant front-end damage. The insured has filed a claim for the damages, and a police report and witness statements have been provided as evidence of the theft and accident.

Stop reason: End of sequence token encountered  
Tokens: 205 input + 60 generated = 265 out of 8192  
Time: 2.2 seconds

Clear output | Generate

# AI governance

## Model evaluation basics: Large Language Models (LLMs)

*Instruction: Read the description below and extract the car make and model, location of the incident like street and date time if there is any mentioned.*

### *Description:*

*A car accident occurred on [Jan 1st, 2023 at 5pm](#) at the intersection of [Woodbridge](#). The insured vehicle, a [Honda Civic](#), was hit by another vehicle that ran a red light. The insured driver, John, was driving within the speed limit and following all traffic rules. The accident resulted in significant damage to the insured vehicle, including a broken bumper and damaged front fender. There were no injuries reported. The insured is filing a claim for the repairs and any necessary medical expenses.*

### Generated output:

*Car Details: Honda Civic  
Location: Woodbridge  
Date: Jan 1st, 2023  
Time of Incident: 5pm*

### Reference output:

*Car Details: Honda Civic  
Location: Woodbridge  
Date: Jan 1st, 2023  
Time of Incident: 5pm*

# AI governance

## Model evaluation basics: Large Language Models (LLMs)

*Instruction: Read the description below and extract the car make and model, location of the incident like street and date time if there is any mentioned.*

### *Description:*

*A car accident occurred on [Jan 1st, 2023 at 5pm](#) at the intersection of [Woodbridge](#). The insured vehicle, a [Honda Civic](#), was hit by another vehicle that ran a red light. The insured driver, John, was driving within the speed limit and following all traffic rules. The accident resulted in significant damage to the insured vehicle, including a broken bumper and damaged front fender. There were no injuries reported. The insured is filing a claim for the repairs and any necessary medical expenses.*

### Generated output:

*Car Details: Honda Civic  
Location: Woodbridge  
Date: Jan 1st, 2023  
Time of Incident: 5pm*

### Reference output:

*Car Details: Honda Civic  
Location: Woodbridge  
Date: Jan 1st, 2023  
Time of Incident: 5pm*

100%

# AI governance

## Model evaluation basics: Large Language Models (LLMs)

*Instruction: Read the description below and extract the car make and model, location of the incident like street and date time if there is any mentioned.*

### *Description:*

*A car accident occurred on [Jan 1st, 2023 at 5pm](#) at the intersection of [Woodbridge](#). The insured vehicle, a [Honda Civic](#), was hit by another vehicle that ran a red light. The insured driver, John, was driving within the speed limit and following all traffic rules. The accident resulted in significant damage to the insured vehicle, including a broken bumper and damaged front fender. There were no injuries reported. The insured is filing a claim for the repairs and any necessary medical expenses.*

### Generated output:

*Car Details: car make and model; location; date and time*

### Expected output:

*Car Details: Honda Civic  
Location: Woodbridge  
Date: Jan 1st, 2023  
Time of Incident: 5pm*

50%

# AI governance

## watsonx.governance metrics for evaluating Large Language Models – Q4'2023

### Text Summarization Metrics

- ROUGE
- SARI
- Normalized F1, Precision, Recall
- METEOR
- Sentence Similarity - Jaccard Similarity
- Sentence Similarity - Cosine Similarity
- BLEU
- Readability, complexity
- Levenshtein distance based Diversity metrics
- HAP Detection on Output Text
- PII Detection on Output Text

### Content Generation Metrics

- ROUGE
- BLEU
- METEOR
- exact\_match
- Readability, complexity
- Levenshtein distance based Diversity metrics
- HAP Detection on Output Text
- PII Detection on Output Text

### Explainability

- Attributions Detection using RAG (API-support)

### Entity Extraction Metrics

- Micro & Macro F1, Precision, Recall

### Q&A Evaluation Metrics

- ROUGE
- BLEU
- METEOR
- exact\_match
- HAP Detection on Output Text
- PII Detection on Output Text

### Quality/Text Classification Metrics

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

### Drift Monitoring

(applicable for all Task Types)

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

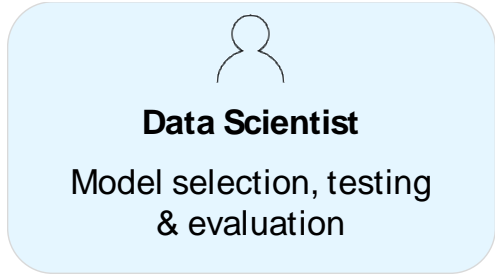
### Model Health Metrics

(applicable for all Task Types)

- Total Scoring Requests
- Number of scoring records (total / min/ max/ median/avg)
- Input token count (total / min/ max/ median/avg)
- Output token count (total / min/ max/ median/avg)
- API Latency (median /avg/min /max)
- API Throughput (median /avg/min /max)
- Record Latency (median /avg/min /max)
- Record Throughput (median /avg/min /max)
- User count

# AI development & AI governance

## Data Scientist evaluates prompt template and creates an AI Factsheet



Projects / Getting started with watsonx.gov... / Insurance claim suggested next ... Open in Prompt Lab

### Configure evaluations

**Evaluations**

- Generative AI Quality
- Model health

#### Settings

Generative AI Quality

Description

Ensure that your minimum sample size is large enough to accurately represent the variety of requests the deployment receives.

Task type

Content generation

Maximum sample size (optional) ⓘ

Minimum sample size (optional) ⓘ

Select metrics to evaluate


- ROUGE
- METEOR
- Text quality
- BLEU
- Output data PII
- Input data PII
- Output data HAP
- Input data HAP
- Readability

Cancel Save

- Depending on the use case, watsonx.governance provides a subset of meaningful metrics that the data scientist can select from to evaluate the model against.

# AI development & AI governance

## Data Scientist evaluates prompt template and creates an AI Factsheet

  
**Data Scientist**  
Model selection, testing  
& evaluation

Projects / Getting started with watsonx.gov... / Insurance claim suggested next ... Open in Prompt Lab

### Configure evaluations

**Evaluations**

- Generative AI Quality
- Model health

#### Generative AI Quality

**Description**

The Generative AI Quality monitor calculates a variety of metrics based on prompt template task type. Some metrics compare model output to the reference output you provide. Other metrics analyze model input and output and do not require reference output.

#### Settings

Task type  
Content generation

Selected metrics  
ROUGE, METEOR, Text quality, BLEU, Output data PII, Input data PII, Readability

#### ROUGE

**Lower thresholds**

ROUGE-1	0.8
ROUGE-2	0.8
ROUGE-L	0.8
ROUGE-Lsum	0.8

Cancel Save

- Thresholds for the evaluated metrics can be defined as needed or approved by the data scientist.

# AI development & AI governance


## Data Scientist evaluates prompt template and creates an AI Factsheet

Projects / Getting started with watsonx.gov... / Insurance claim key information ... Open in Prompt Lab

### Evaluate prompt template

Choose the evaluation dimensions and select the test data. [Learn more](#)

- Select dimensions
- Select test data
- Map variables
- Review and evaluate



**Drop a file here or browse for a file to upload**

Add a CSV file that includes input and expected output (ground-truth). Test data for this deployment can optionally include model output to remove the need for additional model transactions. Maximum size is 8 MB. Maximum number of records is 1000. Minimum number of records is 10.

[Browse](#) [Select from project](#)

[Cancel](#) [Back](#) [Next](#)

- Many of the provided evaluation metrics require a golden source training set (csv). This set needs to include test data of input and output. Watsonx.governance will calculate the metrics based on this input.

# AI development & AI governance

## Data Scientist evaluates prompt template and creates an AI Factsheet

Projects / Getting started with watsonx.gov... / Insurance claim suggested next ... Open in Prompt Lab

### Evaluate prompt template

Choose the evaluation dimensions and select the test data. [Learn more](#)

- ✔ Select dimensions
- ✔ Select test data
- **Map variables**
- Review and evaluate

#### Map prompt variables to columns

For each prompt variable, select the associated column. [Learn more](#)

**Field separation** ⓘ

Select delimiter

Comma (,)

**Input**

input

Insurance\_Claim

**Reference output**

Reference output

Summary

Cancel Back Next

# AI development & AI governance

## Data Scientist evaluates prompt template and creates an AI Factsheet

Projects / Getting started with watsonx.gov... / Insurance claim summarization [...]

Autosave on New prompt + AI guardrails on

### Evaluate prompt template

Choose the evaluation dimensions and select the test data. [Learn more](#)

- ✓ Select dimensions
- ✓ Select test data
- ✓ Map variables
- ⌚ Review and evaluate

#### Review

Task: Text summarization

Test data: Insurance claim summarization test data.csv


Evaluations: Generative AI Quality  
Model health

**Note:** Evaluation can take a several minutes to complete. You can continue to work on other things while your evaluation is in progress.

Cancel Back Evaluate

# AI development & AI governance

## Data Scientist evaluates prompt template and creates an AI Factsheet

  
**Data Scientist**  
Model selection, testing & evaluation

  
**Compliance / Risk Officer**  
Governance & risk assessment

Projects / Validation project / Insurance claim summarization Open in Prompt Lab

AI Factsheet **Evaluate**

Last evaluation: Fri, Aug 2, 2024, 2:16 PM EDT Actions

### Deployment details

**Test data set**  
Insurance claim summarization test data.csv

### Test details

**1** Tests run

Tests passed **0** Tests failed **1**

### Model health

**Records** ⓘ  
**10** Records

**Latency (record)** ⓘ  
**2,315.5** ms Median record latency

**Token count** ⓘ  
**1,938** Total input token count **336** Total output token count

### Generative AI Quality - Text summarization

Alerts triggered

Alerts **1**

Metric	Score	Violation
▼ Rouge		
SARI	86.45	none
METEOR	0.87	none

- Results of the evaluation can be assessed after the evaluation has run successfully
- All results of the evaluation will also be documented in the fact sheet of the use case.

# AI development & AI governance

## Data Scientist evaluates prompt template and creates an AI Factsheet

AI Factsheet

Governance

Foundation model

Prompt template

Prompt parameters

Evaluation

Develop

insurance-demo-v4

Test results

Attachments

Other attachments

### insurance-claims-summarization-use-case

Draft | Andreas Schneider | AI Model Inventory DACH | medium | 9180e620-b6e3-451e-9b61-d94ad8d5c85c

Approach  
Default approach | Version 1.0.0

A default approach for tracking your AI assets.

Lifecycle

01 Develop | 02 Validate | 03 Operate

Untrack

granite-13b-chat-v2

Publisher  
IBM

Decoding

Decoding method  
Greedy

Temperature 0.70

Top P (nucleus sampling) 1

Top K 50

Random seed  
None

Repetition penalty 1

Stopping criteria

Stop sequences  
None

Min tokens 0

Max tokens 200

### insurance-claim-summarization-prompt-template

363a4f27-9376-493c-9dcd-7f356f08ed1b

Task  
Summarization

Created by  
Andreas Schneider

Date  
April 02, 2024 at 05:02:45 PM

Prompt

```
<[system]> You are Granite Chat, an AI language model developed by IBM. You are a cautious assistant. You carefully follow instructions. You are helpful and harmless and you follow ethical guidelines and promote positive behavior.
<[/system]> You are an insurance agent tasked to assess insurance claims.
[Document]
[input]
[End]
Summarize the following insurance claim input. Focus on the car and the damage. Make the summary at least 3 sentences long.
<[/assistant]>
```

Variables

[input]

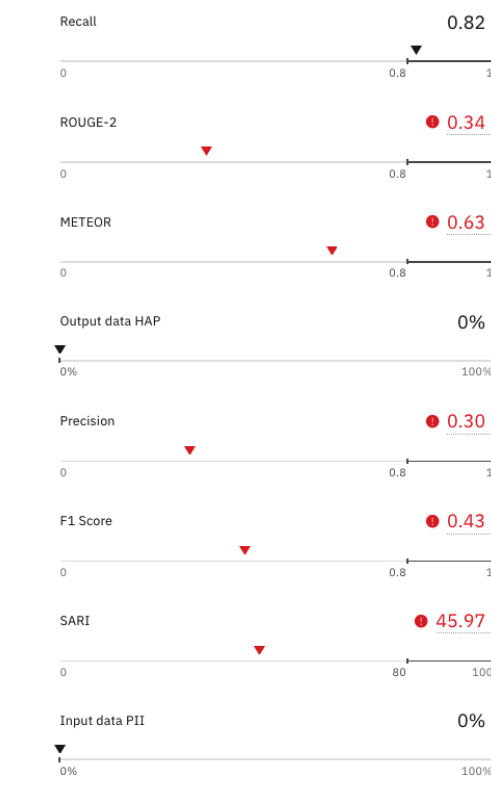
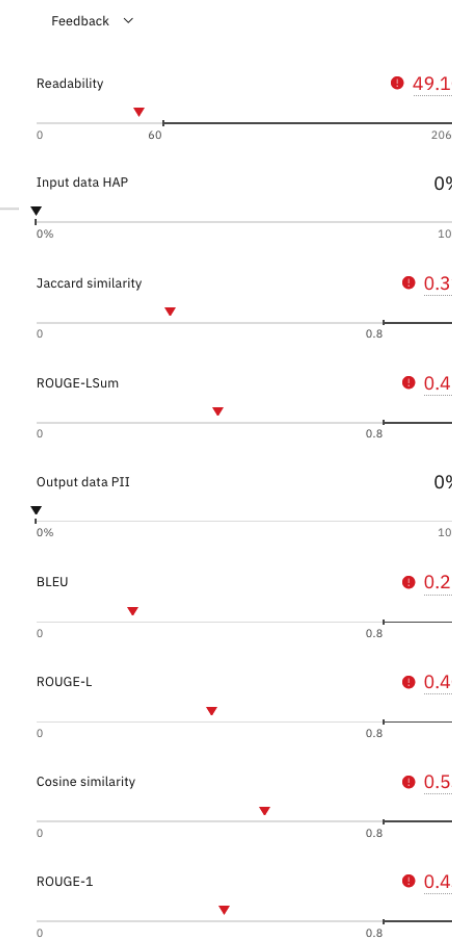
Governance & life cycle information

Foundation model

Prompt template and parameters

### Test results

#### Generative AI Quality



#### Model Health

Total records	10	Maximum record throughput	0.35	Median API throughput	0.27
Median input token count	246	Average output token count	97.60	Minimum API throughput	0.20
Median record latency	3,795	Minimum output token count	60	Users	1
Maximum API latency	5,130	Maximum API throughput	0.35	Total scoring requests	10
Median records	1	Median record throughput	0.26	Minimum API latency	2,836
Average input token count	240.60	Average record throughput	0.25	Maximum input token count	265
Median output token count	101	Total output token count	976	Maximum records	1
Minimum record latency	2,836	Average record latency	3,996.60	Median API latency	3,669
Minimum records	1	Minimum input token count	205	Average API throughput	0.25
Maximum record latency	5,130	Average API latency	3,996.60	Total input token count	2,406
Minimum record throughput	0.20	Maximum output token count	131	Average records	1

Evaluation results



**Data Scientist**  
Model selection, testing & evaluation



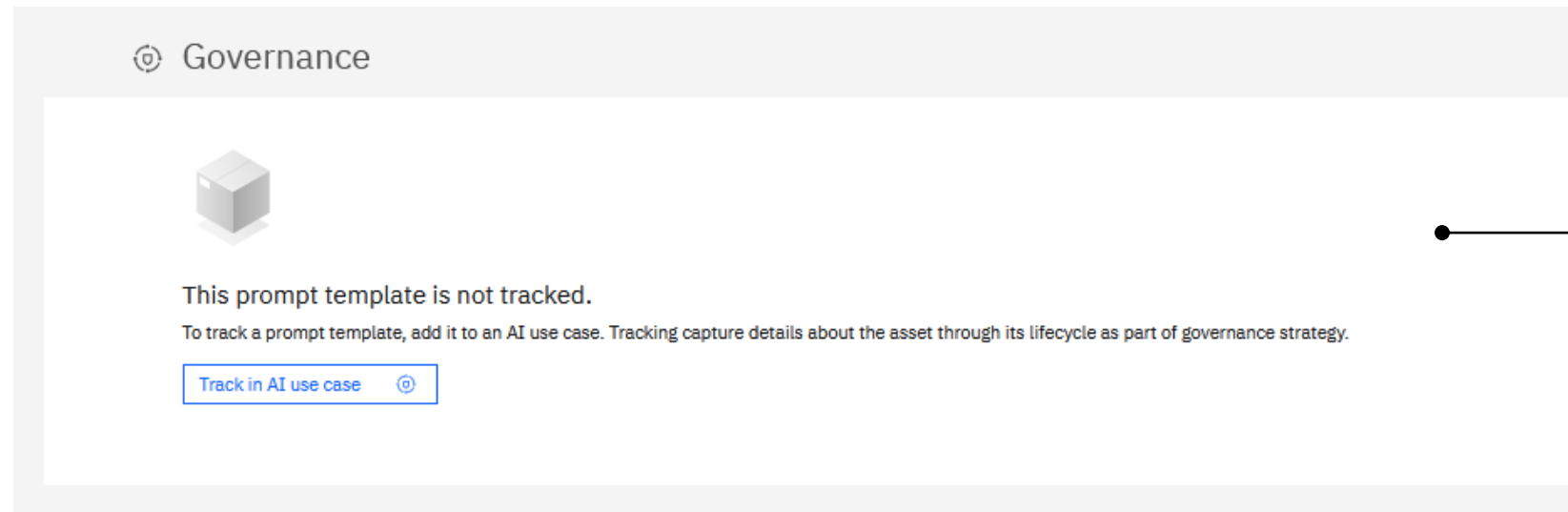
**Compliance / Risk Officer**  
Governance & risk assessment




**Business User**  
Use case definition & request

# AI governance

## Data Scientist evaluates prompt template and creates an AI Factsheet



- Results of the evaluation can be tracked within an AI use case. By tracking the prompt template in the AI use case model information and assessment results will be associated to the use case in the governance console.

  
**Data Scientist**  
Model selection, testing & evaluation

  
**Compliance / Risk Officer**  
Governance & risk assessment

  
**Business User**  
Use case definition & request

IBM watsonx | Governance console

Use Cases: Athena EU A..., Insurance CL...

Insurance Claim - Agent Assist

Task Activity Admin

Read-only

Associations

All Models Prompts and Tunes Associated Foundation Models Deployments Model Links Processes

Name	Description	Model Owner	Model Class	Model Status	Tags
Auto policy risk - P8 Linear Regression - Model-NR001 insurex.ai				Proposed	
Insurance claim key information extraction insurex.ai			Prompt-based	Proposed	
Insurance claim suggested next steps insurex.ai			Prompt-based	Proposed	
Insurance claim summarization insurex.ai			Prompt-based	Proposed	
granite-13b-chat-v2 Library > MRG > Foundation Models	Granite models are designed to be used for a wide range of generative and non-generative tasks with appropriate prompt engineering. They employ a GPT- <a href="#">more</a>		Foundation Model	Awaiting Approval	
granite-13b-instruct-v2	Granite models are designed to be used for a wide		Foundation Model	Awaiting Approval	

Stage: Initial Approval (Awaiting use case approval)  
Due Date: 7/28/2024

Tags: LLM


Initial Approval ⓘ  
Please review the initial details related to the use case as captured by the Use Case Owner.  
Use the Actions button

All Key Items (4)  
✓ Purpose  
✓ Risk Level  
✓ Use Case Type  
✓ Uses Foundation Models \*


- Associated Insurance claim summarization AI model and prompt with the use case

# AI governance

## Data Scientist evaluates prompt template and creates an AI Factsheet

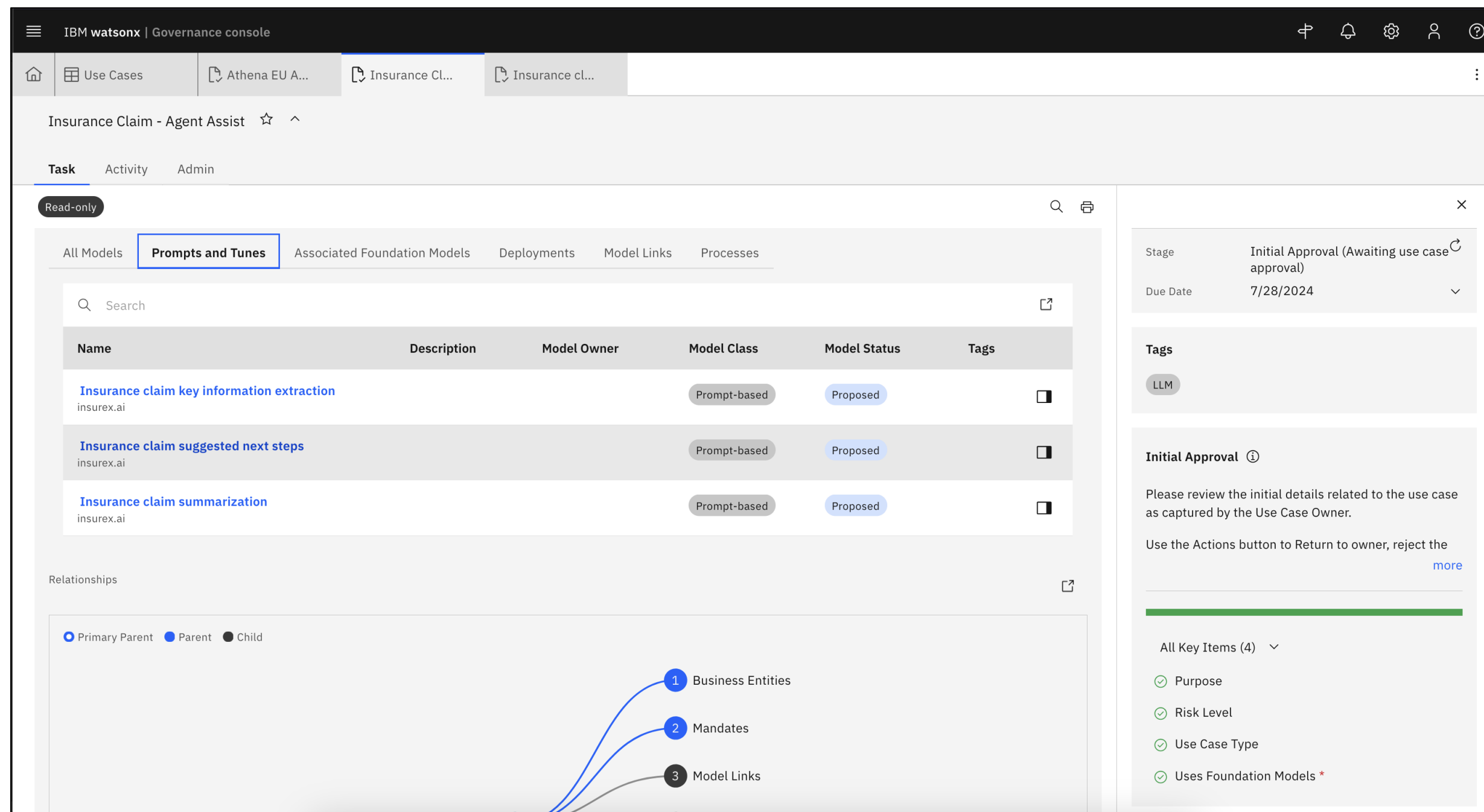
  
**Data Scientist**  
Model selection, testing & evaluation

  
**Compliance / Risk Officer**  
Governance & risk assessment

  
**Business User**  
Use case definition & request

### Associated summarization Prompt with the use case

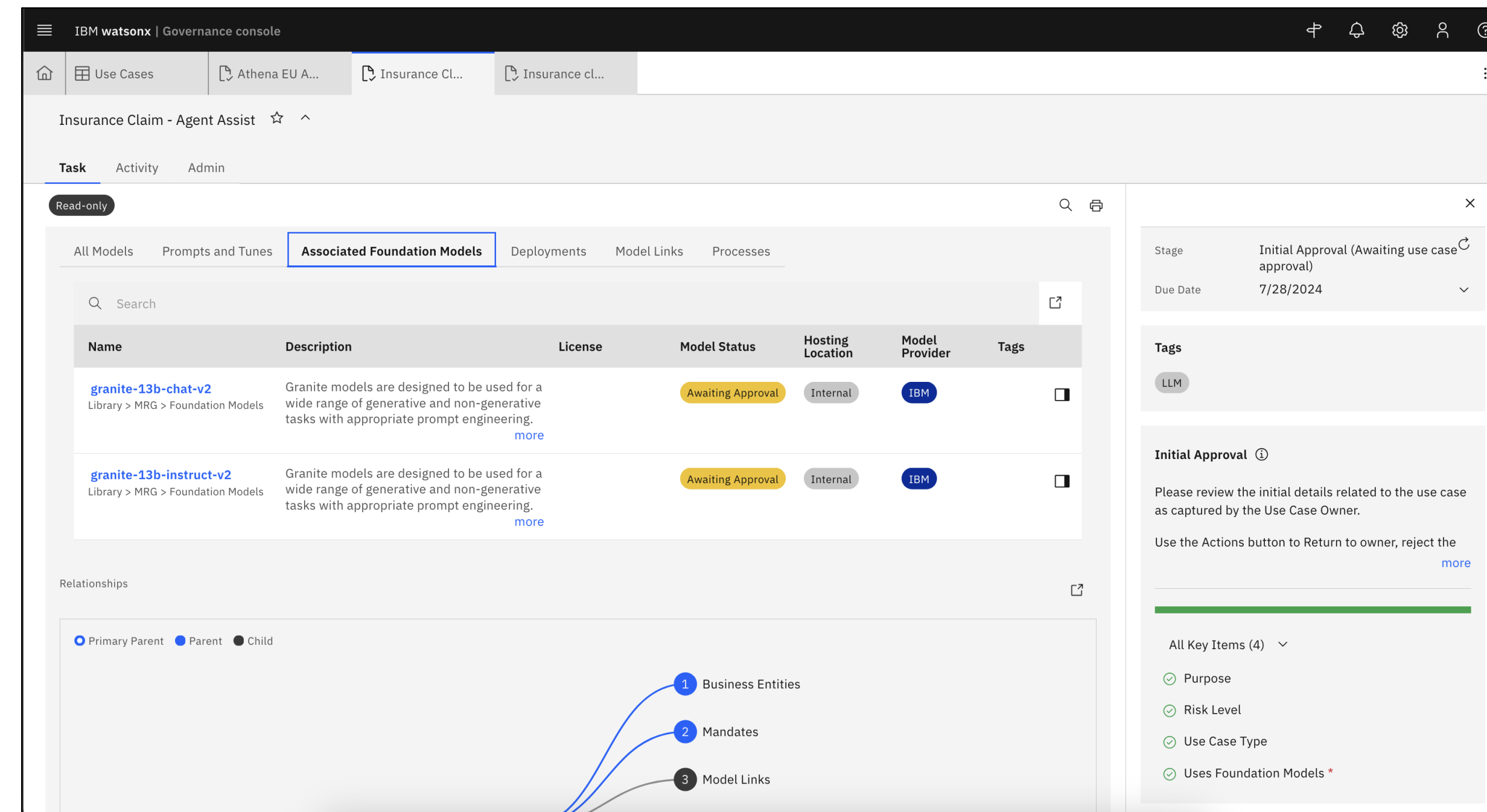
### Associated Foundation Model with the use case



The screenshot shows the 'Prompts and Tunes' section of the IBM Watsonx Governance console. The main table lists three prompts:

Name	Description	Model Owner	Model Class	Model Status	Tags
Insurance claim key information extraction insurex.ai			Prompt-based	Proposed	
Insurance claim suggested next steps insurex.ai			Prompt-based	Proposed	
Insurance claim summarization insurex.ai			Prompt-based	Proposed	

The right-hand sidebar shows the 'Initial Approval' status: 'Initial Approval (Awaiting use case approval)' with a due date of 7/28/2024. Below this, it lists 'All Key Items (4)': Purpose, Risk Level, Use Case Type, and Uses Foundation Models.



The screenshot shows the 'Associated Foundation Models' section of the IBM Watsonx Governance console. The main table lists two foundation models:

Name	Description	License	Model Status	Hosting Location	Model Provider	Tags
granite-13b-chat-v2 Library > MRG > Foundation Models	Granite models are designed to be used for a wide range of generative and non-generative tasks with appropriate prompt engineering. <a href="#">more</a>		Awaiting Approval	Internal	IBM	
granite-13b-instruct-v2 Library > MRG > Foundation Models	Granite models are designed to be used for a wide range of generative and non-generative tasks with appropriate prompt engineering. <a href="#">more</a>		Awaiting Approval	Internal	IBM	

The right-hand sidebar shows the 'Initial Approval' status: 'Initial Approval (Awaiting use case approval)' with a due date of 7/28/2024. Below this, it lists 'All Key Items (4)': Purpose, Risk Level, Use Case Type, and Uses Foundation Models.

# AI governance

## Data Scientist evaluates prompt template and creates an AI Factsheet

The dashboard displays performance monitoring for the 'Athena EU AI Act Bot'. It includes a 'Performance Monitoring' section with a grid of breach status counts and a 'Metrics in Breach' table. A callout box highlights that evaluation results are tracked and documented within the associated use case.

**Performance Monitoring**

All Metrics: 121  
Acceptable Metrics: 3  
Metrics in Warning: 0  
Metric Breaches by Category: 24

**Breach Status**

Breach Status	Quality	Fairness	Drift	Performance	Explainability	Model Health	Other
Not Determined	0	0	0	0	0	0	94
Green	0	0	0	0	0	0	3
Yellow	0	0	0	0	0	0	0
Red	0	0	0	0	0	0	24

**Metrics in Breach**


Name	Description	Evaluation Category	Evaluation Sub-Category	Value	Breach Status	Tags
MET_0000633 Wx Gov LLC	ROUGE-2 metric of generative_ai_quality in develop phase	Other	ROUGE-2	0.2919	Red	Open side panel
MET_0000634 Wx Gov LLC	Faithfulness metric of generative_ai_quality in develop phase	Other	Faithfulness	0.3717	Red	
MET_0000635 Wx Gov LLC	ROUGE-LSum metric of generative_ai_quality in develop phase	Other	ROUGE-LSum	0.42	Red	
MET_0000636 Wx Gov LLC	Answer Relevance metric of generative_ai_quality in develop phase	Other	Answer Relevance	0.5625	Red	
MET_0000639 Wx Gov LLC	BLEU metric of generative_ai_quality in develop phase	Other	BLEU	0.2008	Red	
MET_0000640 Wx Gov LLC	ROUGE-L metric of generative_ai_quality in develop phase	Other	ROUGE-L	0.3456	Red	
MET_0000641 Wx Gov LLC	Exact match metric of generative_ai_quality in develop phase	Other	Exact match	0.00	Red	
MET_0000644 Wx Gov LLC	ROUGE-1 metric of generative_ai_quality in develop phase	Other	ROUGE-1	0.4484	Red	
MET_0000679 Wx Gov LLC	ROUGE-2 metric of generative_ai_quality	Other	ROUGE-2	0.22	Red	
MET_0000680 Wx Gov LLC	Faithfulness metric of generative_ai_quality	Other	Faithfulness	0.1978	Red	

Items per page: 10 | 1-10 of 24 items | 1 of 3 pages

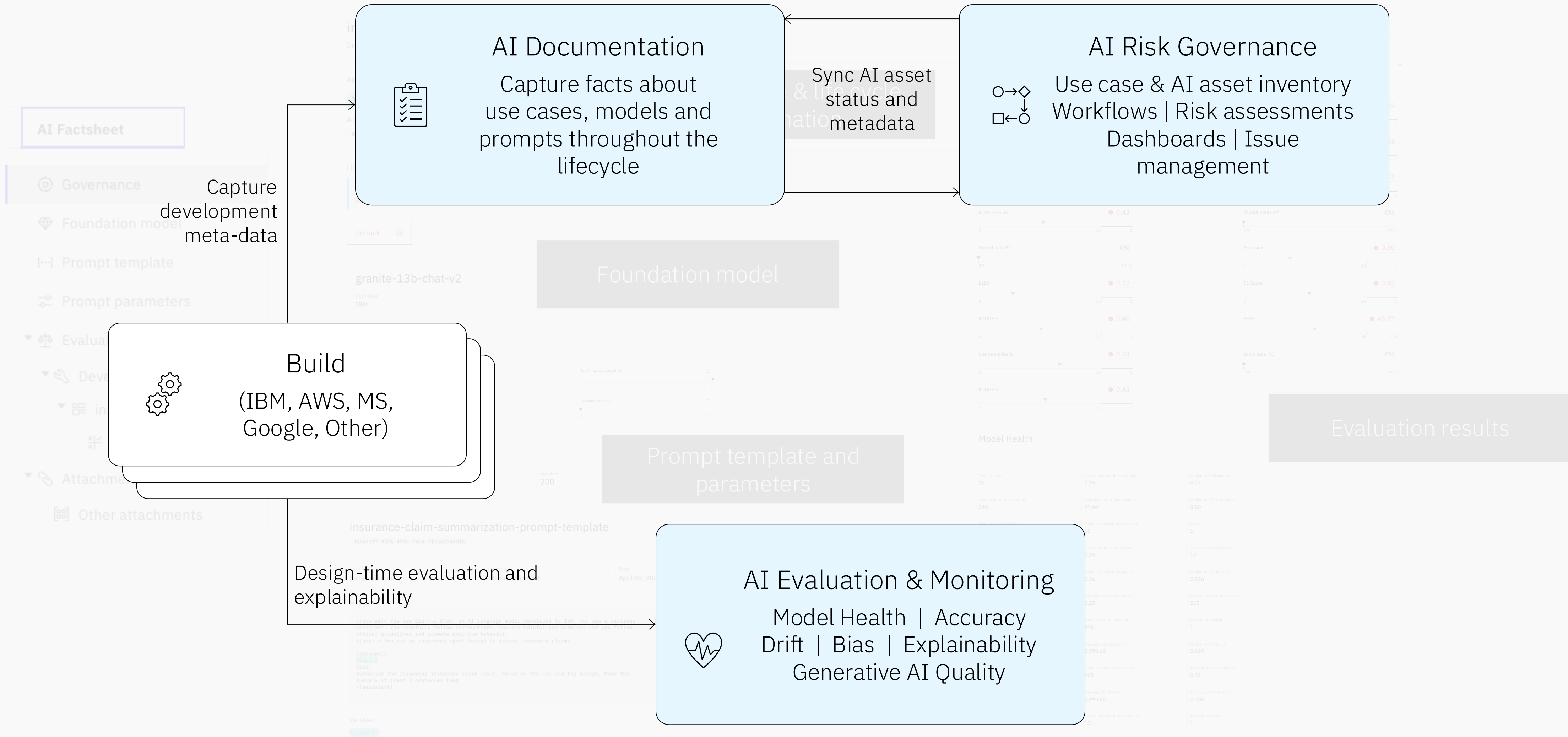
**Callout:** - Evaluation results are also tracked and documented within the associated use case

# AI development & AI governance

## Data Scientist evaluates prompt template and creates an AI Factsheet

  
**Data Scientist**  
Model selection, testing & evaluation

  
**Compliance / Risk Officer**  
Governance & risk assessment



# AI governance

Compliance / Risk Officer checks the status of model compliance, regulatory requirements, metrics and use cases



**Compliance / Risk Officer**  
Governance & risk assessment

IBM watsonx | Governance Console

Welcome, schneider!  
Last successful login 5/15/2024, 2:29 PM

Dashboard My Tasks (1) Subscription Tasks (0) Oversight Tasks (23)

### My Tasks

1

Overdue (1)  
Due soon (0)  
Due in 2+ weeks (0)

Top 5 by due date

5/20/2024	Auto Policy Risk AS150524
-----------	---------------------------

### Regulatory Data

My Regulatory Tasks: 0

My Regulatory Events: 0

My Open Regulatory Changes: 0

Completed Evaluations: 0

### Model Compliance

63

Compliant	40
Non-compliant	10
(No Value)	13

### Use Case by Lifecycle Phase

46

Proposed	30
Awaiting Approval	10
Approved	2
Under Review	2
Rejected	0
Decommissioned	2

### Use Case Summary

Use Case Risk Breakdown: 46

Use Case By Status: 46

High Risk Models: 6

### Use case by lifecycle phase

- Approved

### Useful Links

- Responsible AI Institute
- Model Risk Management - Comptroller's Handbook (OCC)
- EU Draft AI Regulation
- SR 11-7 Information
- E-23 Information
- NYC Local Law 144
- AI Bill of Rights


# AI development & AI governance

## Summarization of car insurance claims



# AI development

## ML/LLMOps Engineer approves & deploys the prompt template as an API endpoint

  
**ML/LLMOps Engineer**  
Model and use case approval & deployment


Deployments /



### insurance-claim-summarization-deployment-v4

- Overview**
- Assets
- Deployments
- Jobs
- Manage

#### Jump back in

 insurance-claim-summarization-prompt-template  
17 hours ago

[View all \(1\)](#)

#### Deployments All

 Deployed	 Failed
<b>1</b>	<b>0</b>


[View deployments](#)

#### Job runs

 Active	 Failed last 24 hours
<b>0</b>	<b>0</b>

[View jobs](#)

#### Space history

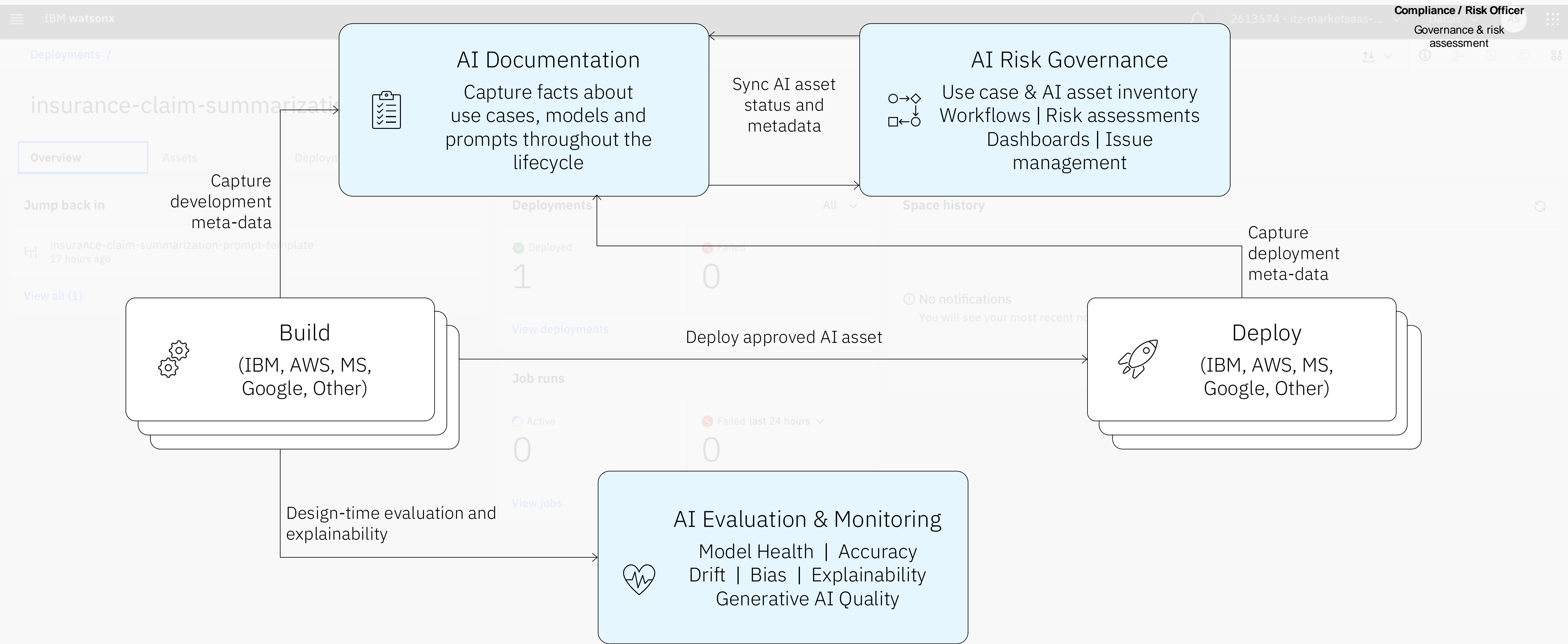
 No notifications  
You will see your most recent notifications here.

# AI development & AI governance

ML/LLMOps Engineer approves & deploys the prompt template as an API endpoint

  
**ML/LLMOps Engineer**  
Model and use case approval & deployment

  
**Compliance / Risk Officer**  
Governance & risk assessment



# AI development

## ML/LLMOps Engineer monitors deployed prompt template

  
**ML/LLMOps Engineer**  
Model and use case approval & deployment


Deployments /



### insurance-claim-summarization-deployment-v4

- Overview**
- Assets
- Deployments
- Jobs
- Manage

#### Jump back in

 insurance-claim-summarization-prompt-template  
17 hours ago

[View all \(1\)](#)

#### Deployments All

 Deployed <b>1</b>	 Failed <b>0</b>
--	--


[View deployments](#)

#### Job runs

 Active <b>0</b>	 Failed last 24 hours <b>0</b>
--	--

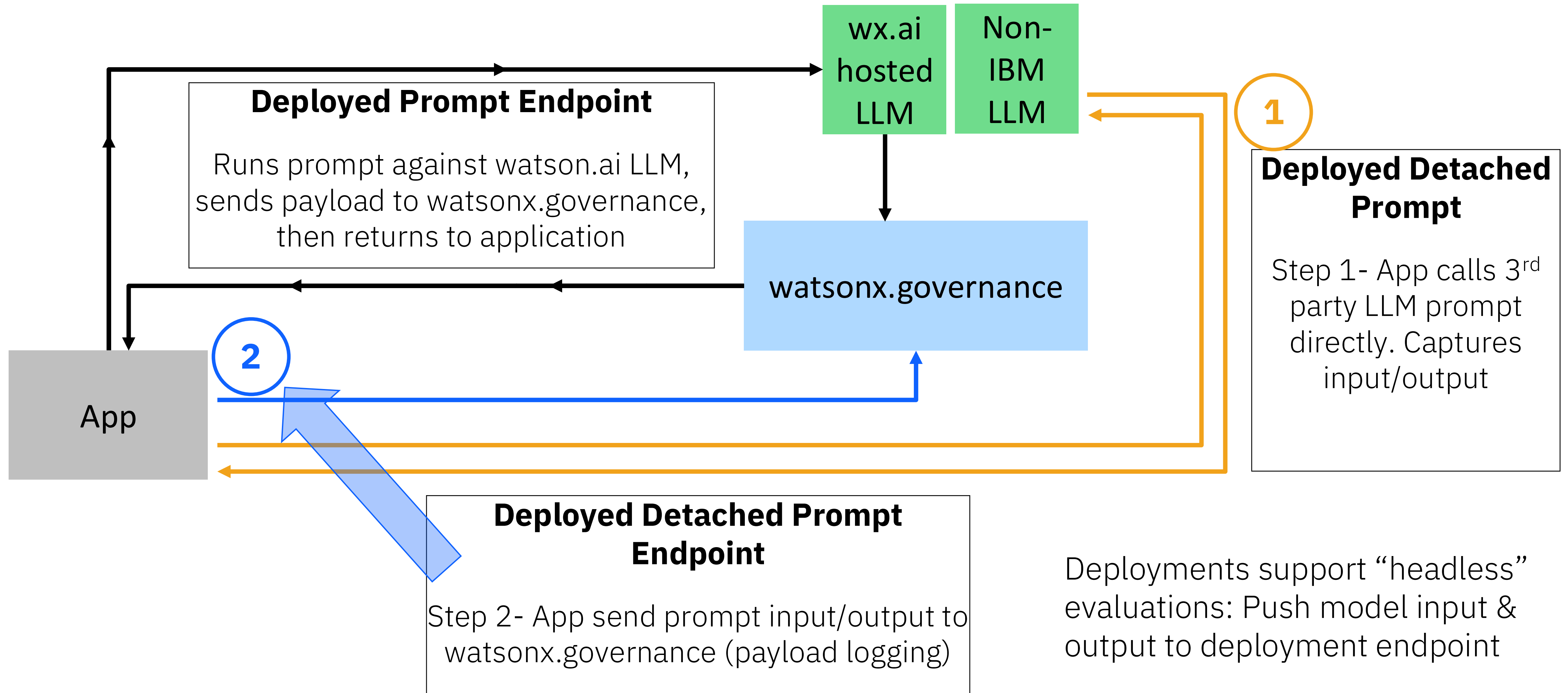
[View jobs](#)

#### Space history

 No notifications  
You will see your most recent notifications here.

# AI development & AI governance

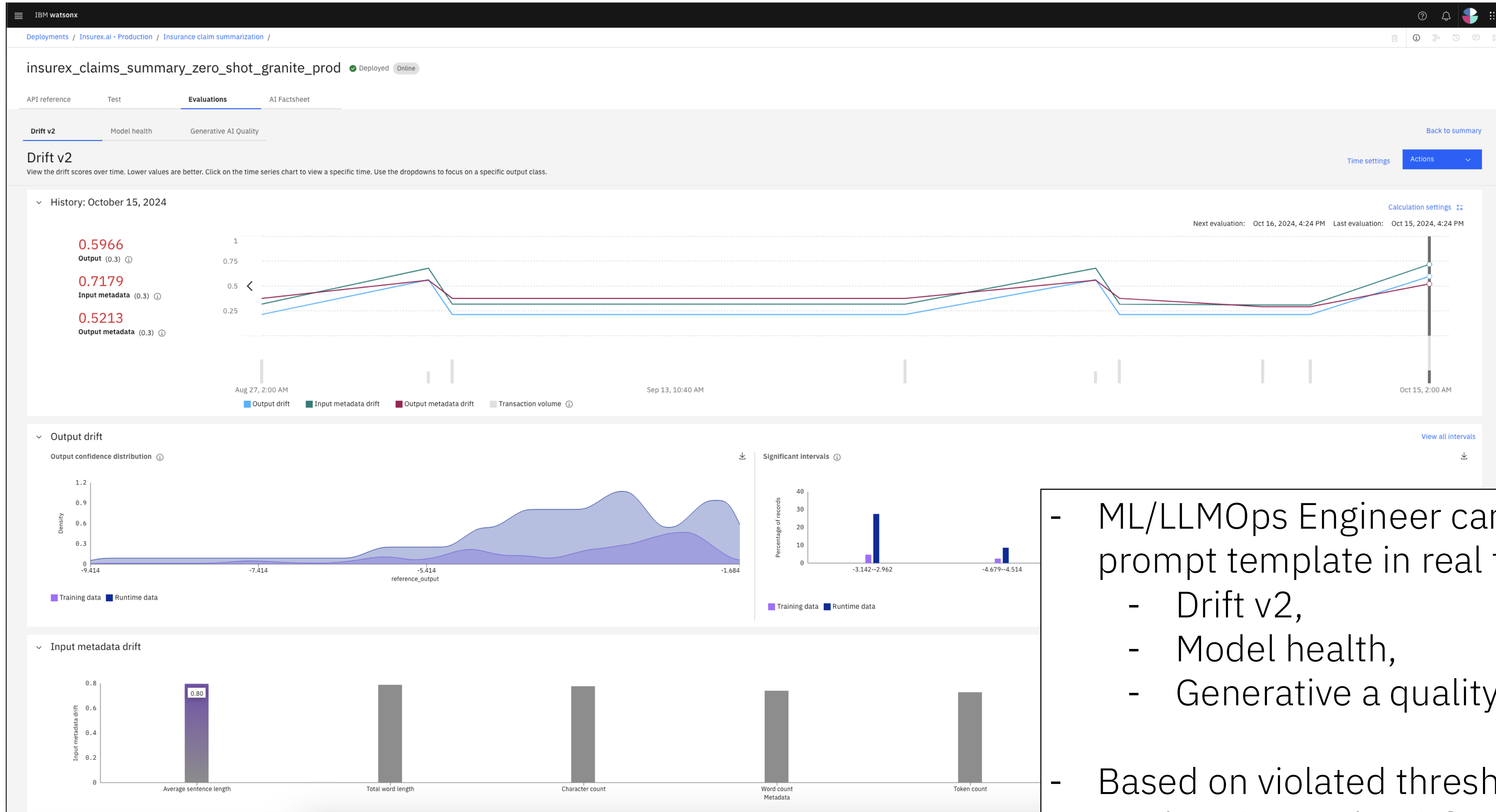
3rd party / external Foundation Model support: watsonx.governance 2.0



# AI governance

## ML/LLMOps Engineer monitors deployed prompt template

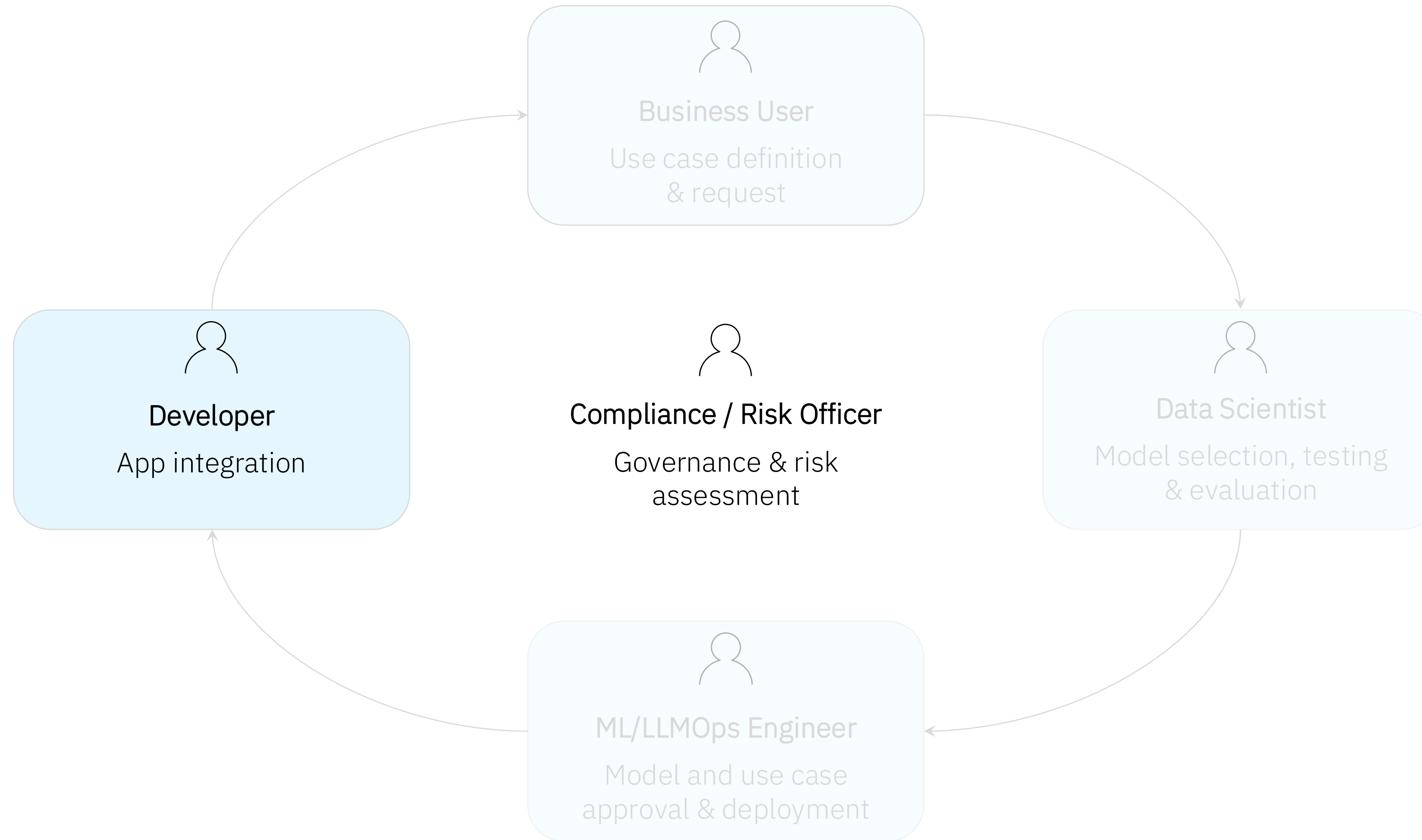
  
**ML/LLMOps Engineer**  
Model and use case  
approval & deployment



- ML/LLMOps Engineer can monitor the deployed prompt template in real time:
  - Drift v2,
  - Model health,
  - Generative a quality metrics
- Based on violated thresholds automatic notifications can be triggered to inform relevant stake holders about a needed change request etc.


# AI development & AI governance

## Summarization of car insurance claims



# AI development

## Developer makes use of the API endpoint to build an app

  
**Developer**  
App integration

Deployments / insurance-claim-summarization-... / insurance-claim-summarization-... /

### insurance-claim-summarization-endpoint Deployed Online

API reference Test Evaluations AI Factsheet

#### Direct link

Private endpoint

<https://private.us-south.ml.cloud.ibm.com/ml/v1/deployments/a1667036-1899-400c-8e1d-f2b0dafa3dc2/text/generation?version=2021-05-01>

[https://private.us-south.ml.cloud.ibm.com/ml/v1/deployments/a1667036-1899-400c-8e1d-f2b0dafa3dc2/text/generation\\_stream?version=2021-05-01](https://private.us-south.ml.cloud.ibm.com/ml/v1/deployments/a1667036-1899-400c-8e1d-f2b0dafa3dc2/text/generation_stream?version=2021-05-01)

Bearer <token> ⓘ

IAM

Public endpoint

<https://us-south.ml.cloud.ibm.com/ml/v1/deployments/a1667036-1899-400c-8e1d-f2b0dafa3dc2/text/generation?version=2021-05-01>

[https://us-south.ml.cloud.ibm.com/ml/v1/deployments/a1667036-1899-400c-8e1d-f2b0dafa3dc2/text/generation\\_stream?version=2021-05-01](https://us-south.ml.cloud.ibm.com/ml/v1/deployments/a1667036-1899-400c-8e1d-f2b0dafa3dc2/text/generation_stream?version=2021-05-01)

[Learn more](#) about the 2021-05-01 version query parameter

#### Code snippets

##### cURL

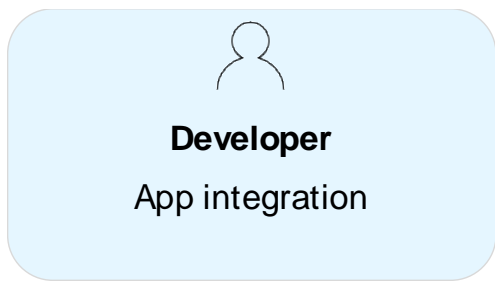
```
# NOTE: you must set $API_KEY below using information retrieved from your IBM Cloud account (https://dataplatfom.cloud.ibm.com/docs/content/wsj/analyze-data/ml-authentication.html)

curl --insecure -X POST --header "Content-Type: application/x-www-form-urlencoded" --header "Accept: \
application/json" --data-urlencode "grant_type=urn:ibm:params:oauth:grant-type:apikey" \
--data-urlencode "apikey=$API_KEY" "https://iam.cloud.ibm.com/identity/token"

# the above CURL request will return an auth token that you will use as $IAM_TOKEN in the scoring request below
# TODO: manually define and pass values to be scored below
curl -X POST --header "Content-Type: application/json" --header "Accept: application/json" --header "Authorization: \
Bearer $IAM_TOKEN" -d '{ "parameters": { "prompt_variables": { "input": "On November 1st, 2023, at 11:00 am, my vehicle, a 2004 Honda Civic, was stolen from its parking spot in Lagos. I immediately reported the incident to the police and o
```

# AI development: Programmatic access

Use the development environment of your choice to interact with the watsonx platform (see some samples below)



## Model inference

```
1 from ibm_watsonx_ai.foundation_models import ModelInference
2 from ibm_watsonx_ai.metanames import GenTextParamsMetaNames as GenParams
3 from ibm_watsonx_ai.foundation_models.utils.enums import ModelTypes, DecodingMethods
4
5 # To display example params enter
6 GenParams().get_example_values()
7
8 generate_params = {
9     GenParams.MAX_NEW_TOKENS: 200
10 }
11
12 model_inference = ModelInference(
13     model_id=ModelTypes.GRANITE_13B_CHAT_V2,
14     params=generate_params,
15     credentials={
16         "apikey": "DHL4To90TGR_ICEWybY0g9mEBfzSxzqJ8Iu-nYwREPI0",
17         "url": "https://us-south.ml.cloud.ibm.com"
18     },
19     project_id="cb1425f8-e2e3-414a-b469-0182932e7473"
20 )
21
22 q = "<|system|> You are Granite Chat, an AI language model developed by IBM. You are
23 <|user|> You are an insurance agent tasked to assess insurance claims. \
24 [Document] \
25 On November 1st, 2023, at 11:00 am, my vehicle, a 2004 Honda Civic, was stolen fr
26 [End] \
27 Summarize the following insurance claim input. Focus on the car and the damage. M
28 <|assistant|>"
29
30 generated_response = model_inference.generate(prompt=q)
31 print(generated_response)
```

### Product documentation

<https://dataplatfom.cloud.ibm.com/docs/content/wsj/analyze-data/fm-python-lib.html?context=wx>

### ibm-watsonx-ai Python library

[https://ibm.github.io/watsonx-ai-python-sdk/foundation\\_models.html](https://ibm.github.io/watsonx-ai-python-sdk/foundation_models.html)

## LangChain integration

```
33 # LangChain integration
34 from langchain_ibm import WatsonxLLM
35
36 flan_ul2_llm = WatsonxLLM(
37     model_id=model_id_1,
38     url=credentials["url"],
39     apikey=credentials["apikey"],
40     project_id=credentials["project_id"],
41     params=parameters
42 )
43
44 flan_t5_llm = WatsonxLLM(
45     model_id=model_id_2,
46     url=credentials["url"],
47     apikey=credentials["apikey"],
48     project_id=credentials["project_id"],
49     params=parameters
50 )
51
52 #print(flan_ul2_llm.dict())
53 #print(flan_t5_llm.dict())
54
55 # Sequential Chain experiment
56
57 from langchain_core.prompts import PromptTemplate
58 prompt_1 = PromptTemplate(
59     input_variables=["topic"],
60     template="Generate a random question about {topic}: Question: "
61 )
62
63 prompt_2 = PromptTemplate(
64     input_variables=["question"],
65     template="Answer the following question: {question}",
66 )
67
68 from langchain.chains import LLMChain
69 prompt_to_flan_ul2 = LLMChain(llm=flan_ul2_llm, prompt=prompt_1, output_key='ar
70
71 from langchain.chains import SequentialChain
72 qa = SequentialChain(chains=[prompt_to_flan_ul2, flan_to_t5], input_va
73
74 print(qa.invoke({"topic": "life"}))
```

### Product documentation

<https://dataplatfom.cloud.ibm.com/docs/content/wsj/analyze-data/fm-python-lib.html?context=wx>

### LangChain

[https://python.langchain.com/docs/integrations/llm/s/ibm\\_watsonx/](https://python.langchain.com/docs/integrations/llm/s/ibm_watsonx/)

## Governance metrics toolkit

### Azure OpenAI related

```
import os
import openai

# Get the Azure OpenAI deployment details from Azure OpenAI Studio
openai.api_type = "azure"
openai.api_base = "<Replace with Azure OpenAI API Base URL>"
openai.api_version = "<Replace with Azure OpenAI API Version>"
openai.api_key = "<Replace with Azure OpenAI API Key>"

def get_prompt(text):
    prompt = f"<<<Please provide a summary of the following text with maximum of 20 words.
    {text}
    Summary:>>>"
    return prompt
```

### Run the prompt evaluation

```
# Replace with the prompt config values, as needed
def get_completion(prompt_text):
    response = openai.Completion.create(
        engine = "azure-openai-deployment-001",
        prompt=get_prompt(prompt_text),
        temperature=0.1,
        max_tokens=30,
        top_p=0.5,
        frequency_penalty=0,
        presence_penalty=0,
        stop='\n'
    )
    return response.choices[0].text
    #return response
```

### Evaluate Metrics

#### IBM watsonx.governance authentication

```
from ibm_cloud_sdk_core.authenticators import IAMAuthenticator, BearerTokenAuthenticator, CloudPakForDataAuthentic
from ibm_watsonx.governance import *
from ibm_watsonx.governance.supporting_classes.enums import *
from ibm_watsonx.governance.supporting_classes import *

if use_cpd:
    authenticator = CloudPakForDataAuthenticator(
        url=WOS_CREDENTIALS['url'],
        username=WOS_CREDENTIALS['username'],
        apikey=WOS_CREDENTIALS['api_key'],
        disable_ssl_verification=True
    )
else:
    authenticator = IAMAuthenticator(apikey=CLOUD_API_KEY)
client = APIClient(authenticator=authenticator)
print(client.version)
```

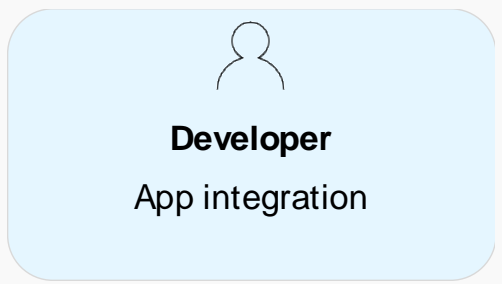
### Metrics configuration for evaluation

```
metric_config = {
    "configuration": {
        LLMTxtMetricGroup.SUMMARIZATION.value: {
            LLMSummarizationMetrics.ROUGE_SCORE.value: {},
            LLMSummarizationMetrics.SARI.value: {},
            LLMSummarizationMetrics.METEOR.value: {},
            LLMSummarizationMetrics.NORMALIZED_RECALL.value: {},
            LLMSummarizationMetrics.NORMALIZED_PRECISION.value: {},
            LLMSummarizationMetrics.NORMALIZED_F1_SCORE.value: {},
            LLMSummarizationMetrics.COSINE_SIMILARITY.value: {},
            LLMSummarizationMetrics.JACCARD_SIMILARITY.value: {},
            LLMSummarizationMetrics.BLEU.value: {},
            LLMSummarizationMetrics.FLESH.value: {}
        }
    }
}
```

Name	Description	Owner	Status	Branch Status	Indication Trend	Tags
cosine_similarity	watsonx.governance.metrics for 'cosine_similarity'		Active		Not Determined	
jaccard_similarity	watsonx.governance.metrics for 'jaccard_similarity'		Active		Not Determined	
meteor	watsonx.governance.metrics for 'meteor'		Active		Not Determined	
rouge1	watsonx.governance.metrics for 'rouge1'		Active		Not Determined	
rouge2	watsonx.governance.metrics for 'rouge2'		Active		Not Determined	
rougeL	watsonx.governance.metrics for 'rougeL'		Active		Not Determined	
rougeLsum	watsonx.governance.metrics for 'rougeLsum'		Active		Not Determined	
sari	watsonx.governance.metrics for 'sari'		Active		Not Determined	

# AI development: Programmatic access

Use the development environment of your choice to interact with the watsonx platform (see some samples below)



Model inference

```

1 from ibm_watsonx_ai.foundation_models import ModelInference
2 from ibm_watsonx_ai.metanames import GenTextParamsMetaNames as GenParams
3 from ibm_watsonx_ai.foundation_models import ModelInference
4
5 # To display example params enter:
6 GenParams().get_example_values()
7
8 generate_params = {
9     GenParams.MAX_NEW_TOKENS: 80
10 }
11
12 model_inference = ModelInference(
13     model_id=ModelTypes.GPT4_13B_CHAT_V2,
14     params=generate_params,
15     credentials={
16         "apikey": "DHI4T1J0TGR_ICEWybY0g9mEBfz5xzqJ8Iu-nYwREPI0",
17         "url": "https://us-south.ml.cloud.ibm.com"
18     },
19     project_id="cb1425f9-14a-b469-0182932e7473"
20 )
21
22 q = "<|system|> You are Watsonx Chat, an AI language model developed by IBM. You are
23 <|user|> You are an insurance agent tasked to assess insurance claims. \
24 [Doc] \
25 On Nov 15, 2004, a 2004 Honda Civic, was stolen fr
26 [End] \
27 Summarize the following insurance claim input. Focus on the car and the damage. M
28 <|assistant|> "
29
30 generated_response = model_inference.generate(prompt=q)
31 print (generated_response)
    
```

Monitor model

Deploy model

Test model

MLOps lifecycle

LangChain integration

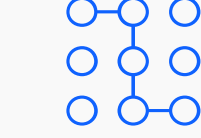
```

33 # LangChain integration
34 from langchain_ibm import WatsonxLLM
35
36 llm = WatsonxLLM(
37     model_id=model_id_1,
38     url=credentials["url"],
39     apikey=credentials["apikey"],
40     project_id=credentials["project_id"],
41     params=parameters
42 )
43
44 flan_t5_llm = WatsonxLLM(
45     model_id=model_id_2,
46     url=credentials["url"],
47     apikey=credentials["apikey"],
48     project_id=credentials["project_id"],
49     params=parameters
50 )
51
52 # Print the output of the llm
53 #print(flan_ul2_llm.invoke(prompt))
54 #print(flan_t5_llm.invoke(prompt))
55
56 # Sequential Chain
57 from langchain.prompts import PromptTemplate
58 prompt_1 = PromptTemplate(
59     input_variables=["topic"],
60     template="Generate a random question about {topic}: Question: "
61 )
62 prompt_2 = PromptTemplate(
63     input_variables=["question"],
64     template="Answer the following question: {question}",
65 )
66
67 from langchain.chains import LLMChain
68 prompt_to_flan_ul2 = LLMChain(llm=flan_ul2_llm, prompt=prompt_1, output_key='answer')
69 flan_to_t5 = LLMChain(llm=flan_t5_llm, prompt=prompt_2, output_key='answer')
70
71 from langchain.chains import SequentialChain
72 qa = SequentialChain(chains=[prompt_to_flan_ul2, flan_to_t5], input_variables=["topic"])
73
74 print(qa.invoke({"topic": "life"}))
    
```

Find data

Prepare data

Train model



Integration

Testing

Deployment

Governance metrics toolkit

Azure OpenAI related

Get the Azure OpenAI deployment details from Azure OpenAI Studio

```

openai.api_type = "azure"
openai.api_base = "Replace with Azure OpenAI API Base URL"
openai.api_version = "Replace with Azure OpenAI API Version"
openai.api_key = "Replace with Azure OpenAI API Key"

def get_prompt(text):
    prompt = f"""Please provide a summary of the following text with maximum of 28 words.
    {text}
    Summary:"""
    return prompt
    
```

Run the prompt evaluation

```

# Replace with the prompt
def get_completion(prompt):
    response = openai.Completion(
        engine="gpt-4o-mini",
        prompt=get_prompt(prompt_text),
        temperature=0.1,
        max_tokens=28,
        top_p=0.5,
        frequency_penalty=0,
        presence_penalty=0,
        stop="!"
    )
    return response.choices[0].text
    
```

Metrics configuration for evaluation

```

metric_config = {
    "configuration": {
        LLMTextMetricGroup.SUMMARIZATION.value: {
            LLMSummarizationMetrics.ROUGE_SCORE.value: {},
            LLMSummarizationMetrics.SARI.value: {},
            LLMSummarizationMetrics.METEOR.value: {},
            LLMSummarizationMetrics.NORMALIZED_RECALL.value: {},
            LLMSummarizationMetrics.NORMALIZED_PRECISION.value: {},
            LLMSummarizationMetrics.NORMALIZED_F1_SCORE.value: {},
            LLMSummarizationMetrics.COSINE_SIMILARITY.value: {},
            LLMSummarizationMetrics.JACCARD_SIMILARITY.value: {},
            LLMSummarizationMetrics.BLEU.value: {},
            LLMSummarizationMetrics.F1ESCH.value: {}
        }
    }
}
    
```

Monitoring\*

Large Language Models

Prompt Engineering

Tuning

Product documentation  
<https://dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/fm-python-lib.html?context=wx>

ibm-watsonx-ai Python library  
[https://ibm.github.io/watsonx-ai-python-sdk/foundation\\_models.html](https://ibm.github.io/watsonx-ai-python-sdk/foundation_models.html)

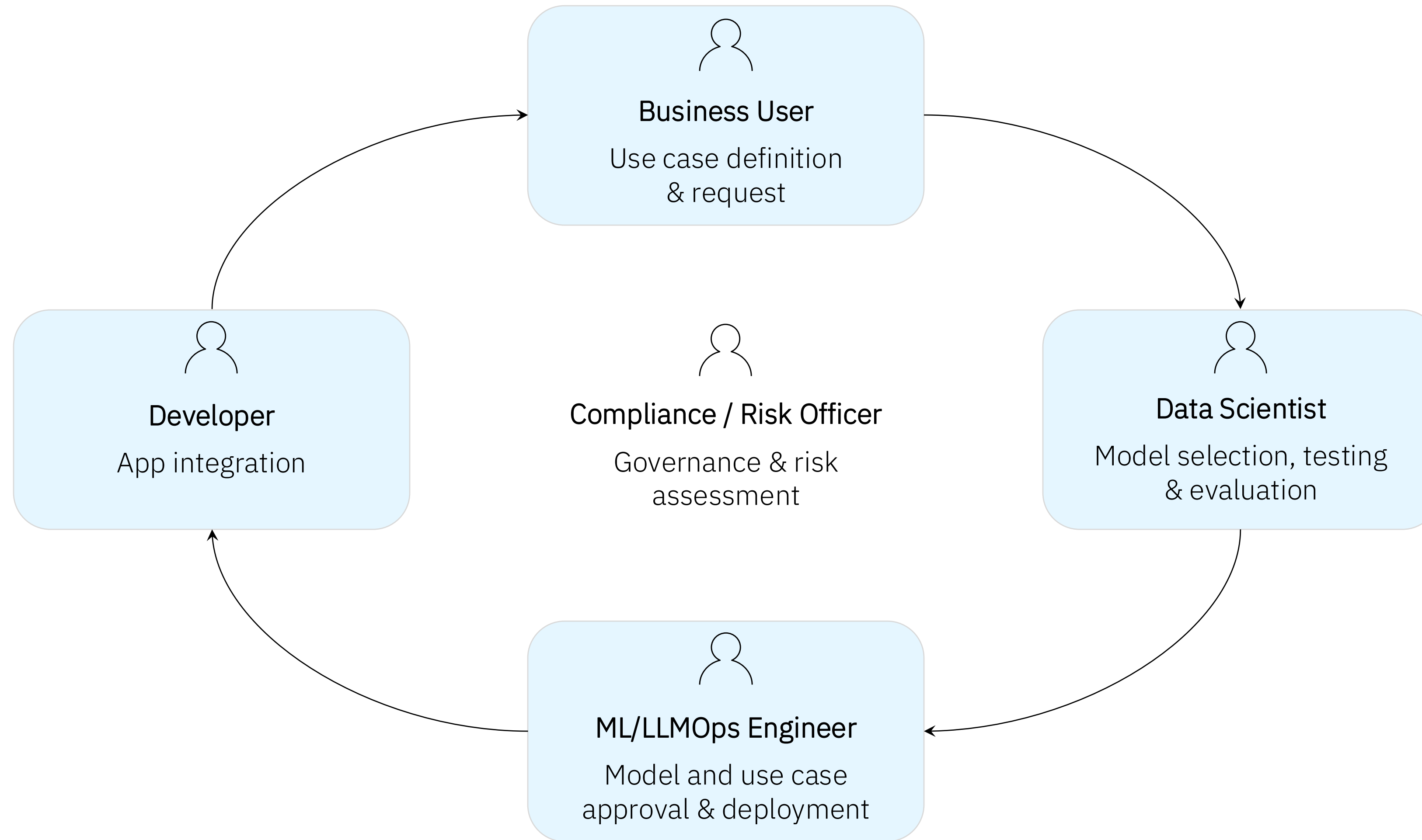
Product documentation  
<https://dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/fm-python-lib.html?context=wx>

LangChain  
[https://python.langchain.com/docs/integrations/llm/s/ibm\\_watsonx/](https://python.langchain.com/docs/integrations/llm/s/ibm_watsonx/)

Model	Deployment	Model	Deployment	Model	Deployment
gpt-4o-mini	Deployed	gpt-4o-mini	Deployed	gpt-4o-mini	Deployed
gpt-4o-mini	Deployed	gpt-4o-mini	Deployed	gpt-4o-mini	Deployed
gpt-4o-mini	Deployed	gpt-4o-mini	Deployed	gpt-4o-mini	Deployed
gpt-4o-mini	Deployed	gpt-4o-mini	Deployed	gpt-4o-mini	Deployed
gpt-4o-mini	Deployed	gpt-4o-mini	Deployed	gpt-4o-mini	Deployed


# AI development & AI governance

## Summarization of car insurance claims



# AI development & AI governance









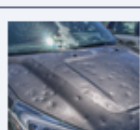
## Example application from the business user perspective

  
**Business User**  
 Use case definition  
 & request

Digital Claims Agent - Powered by watsonx.ai


### Welcome, Alex!

#### Your Inbox

Claim ID	Image	Priority	Description	
CIC202107863		1	On November 5th, 2023, a fire broke out ...	<a href="#">Work on this claim</a>
CIC202105895		1	On August 10th, 2023, the insured vehicl...	<a href="#">Work on this claim</a>
CIC202108708		1	The insured vehicle, a Chevrolet Silvera...	<a href="#">Work on this claim</a>
CIC202109875		2	On May 30th, 2023, heavy rainfall caused...	<a href="#">Work on this claim</a>
CIC202105995		2	While driving on September 8th, 2023, th...	<a href="#">Work on this claim</a>
CIC202109635		3	On January 15th, 2023, the insured vehic...	<a href="#">Work on this claim</a>
CIC202123405		3	The insured vehicle, a Honda Accord, was...	<a href="#">Work on this claim</a>
CIC202103405		3	The insured vehicle, a BMW 3 Series, was...	<a href="#">Work on this claim</a>
CIC202105005		3	The insured vehicle, a Toyota Camry, was...	<a href="#">Work on this claim</a>
CIC202105505		3	The insured vehicle, a Tesla Model S, wa...	<a href="#">Work on this claim</a>

#### Your Overview

**Outstanding Claims:** You have **10 of 10** still to do, of which **3** are high priority.

**Claim Distribution:** 

Digital Claims Agent - Powered by watsonx.ai



Name of Caseworker: **Alex MacLeod**

Caseworker ID: **42425EBC**

Evaluated Priority: **1**

[Generate Evaluation](#)

#### Observations:

On November 5th, 2023, a fire broke out in the insured's garage, resulting in significant damage to the insured vehicle, a Honda Civic. The insured immediately contacted the fire department, and the fire was extinguished, but not before the vehicle sustained extensive fire damage. The fire damaged the exterior paint, melted parts of the body, and caused smoke and soot damage to the interior. The insured promptly reported the incident to their insurance company and is filing a claim for the repairs. The insurance company has arranged for an assessment of the damages by a qualified auto repair specialist. The insured is providing any necessary documentation, including photographs of the damaged vehicle and a description of the fire incident. The insurance company will cover the cost of repairs or replacement, subject to the terms and conditions of the policy.

**Case Number: CIC202107863**

Generated at: 10:50 04/04/2024

Make & Model	Honda Civic
Location	Not Found
Date	November 5th, 2023
Time of incident	Not Found
Overview	A fire broke out in the insured's garage, resulting in significant damage to the insured vehicle, a Honda Civic
Steps to Remediation	<ul style="list-style-type: none"> <li><input type="checkbox"/> Verify the insured's policy coverage to ensure that it includes coverage for fire damage and the necessary repairs.</li> <li><input type="checkbox"/> Review the provided information regarding the fire incident, including the date of occurrence, location (in the insured's garage), and a description of the damages.</li> <li><input type="checkbox"/> Request the insured to provide the police report documenting the fire incident. The police report will serve as crucial evidence and help establish the validity of the claim.</li> <li><input type="checkbox"/> Engage a reputable auto repair shop to assess the damages and provide an estimate for the necessary repairs and replacement parts. Consider obtaining multiple estimates to ensure accuracy and fairness.</li> <li><input type="checkbox"/> Carefully review all supporting documents, including the police report and the repair shop's estimate. Verify the estimated cost of repairs and validate that the damages align with the incident described by the insured.</li> <li><input type="checkbox"/> Maintain regular communication with the insured, providing updates on the claim process and addressing any questions or concerns they may have. Keep them informed about the progress and</li> </ul>

[Finalize Claim](#)

# AI development & AI governance

## Example application from the business user perspective



Compliance / Risk Officer  
Governance & risk assessment

