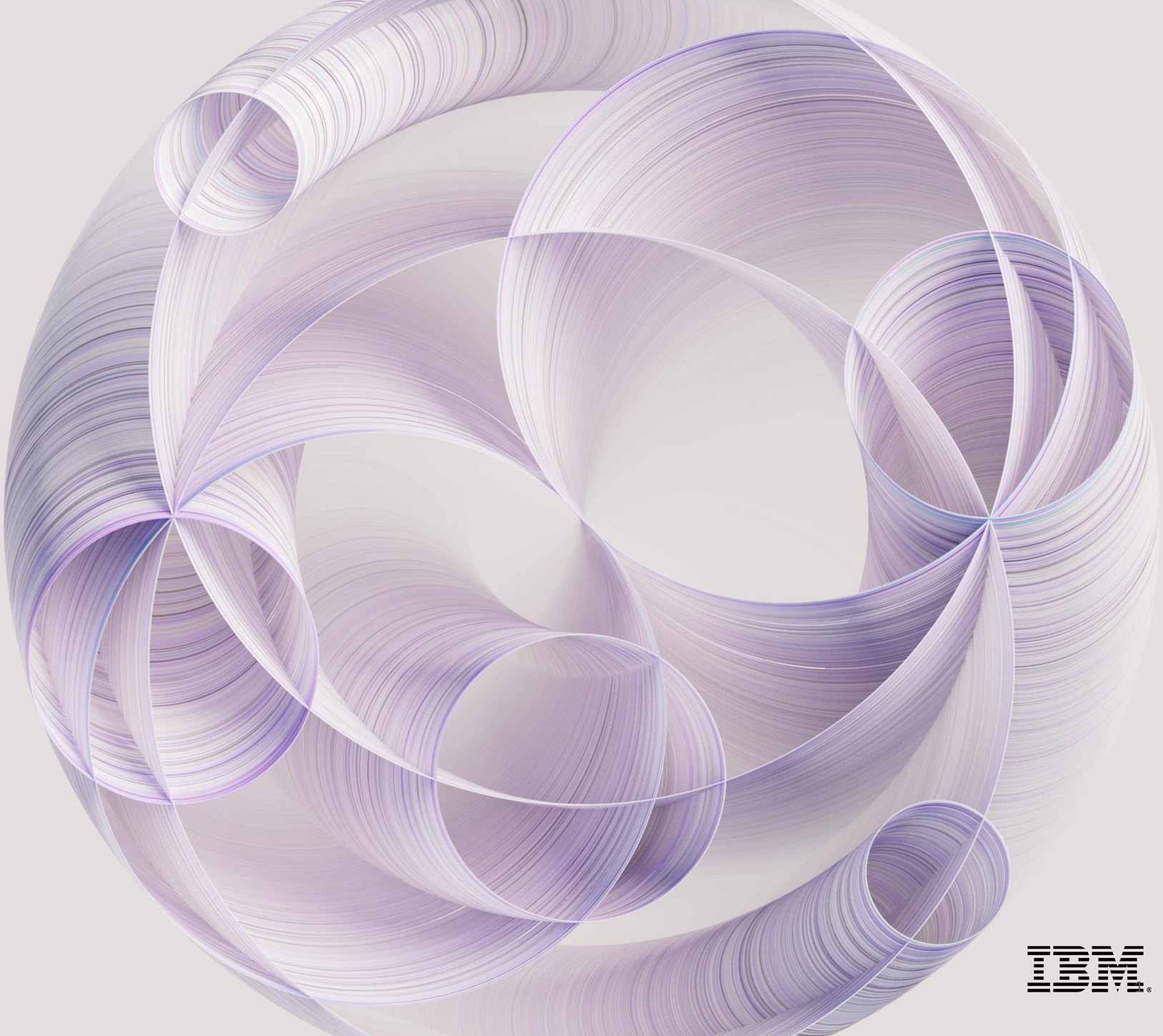
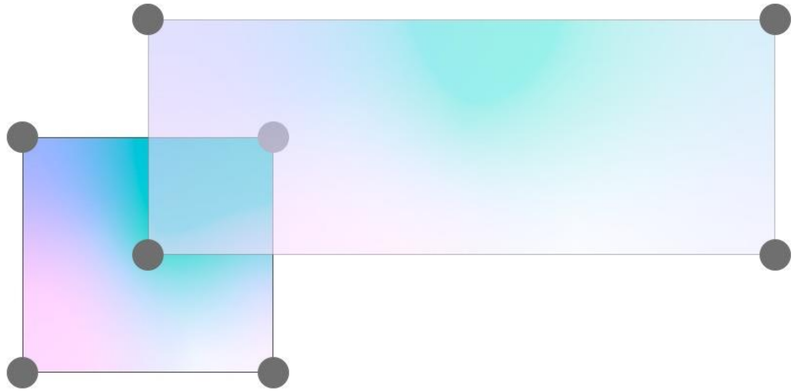
AI development & AI governance

An end-to-end perspective

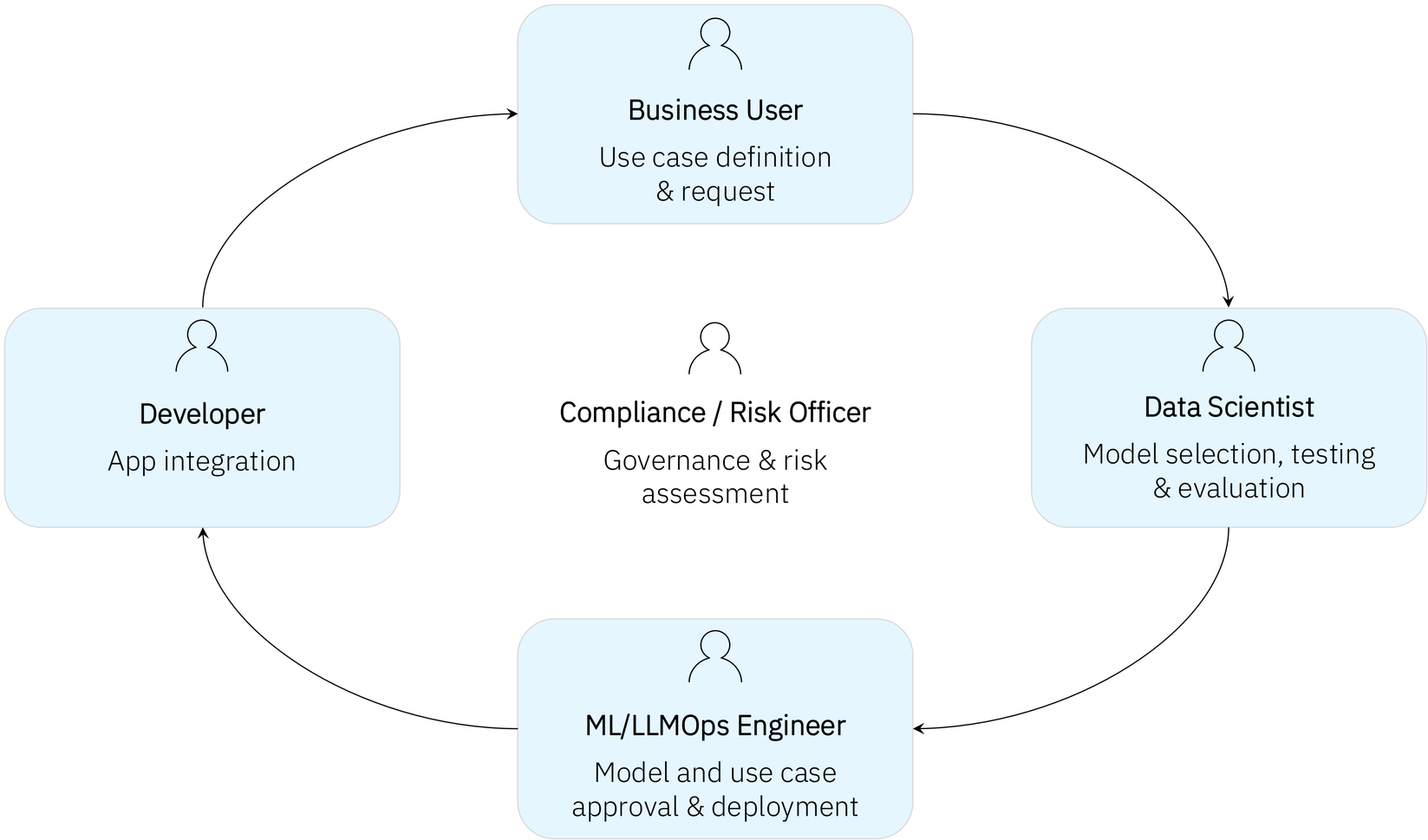


IBM Technology

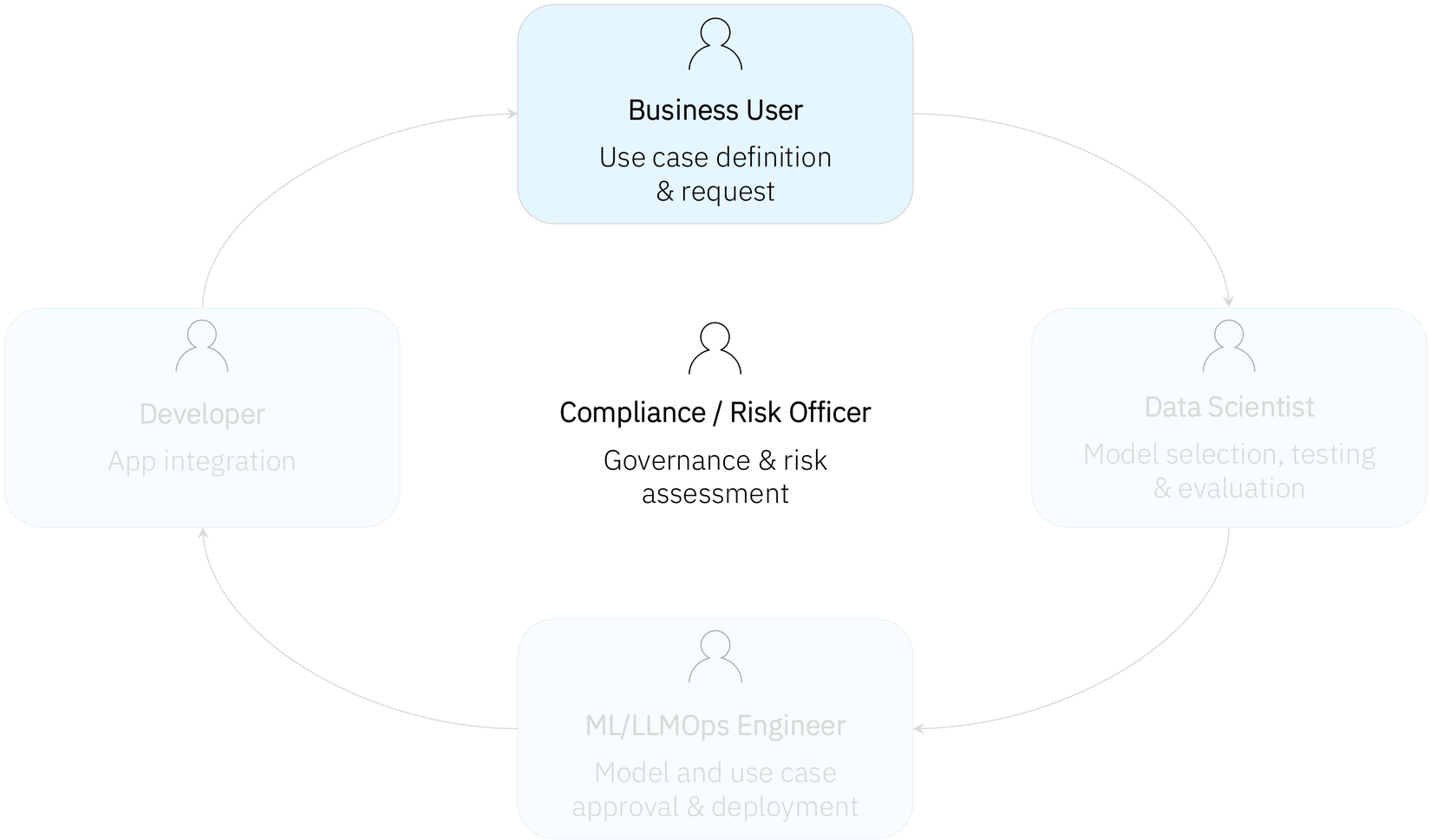
AI development & AI governance
An end-to-end perspective



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

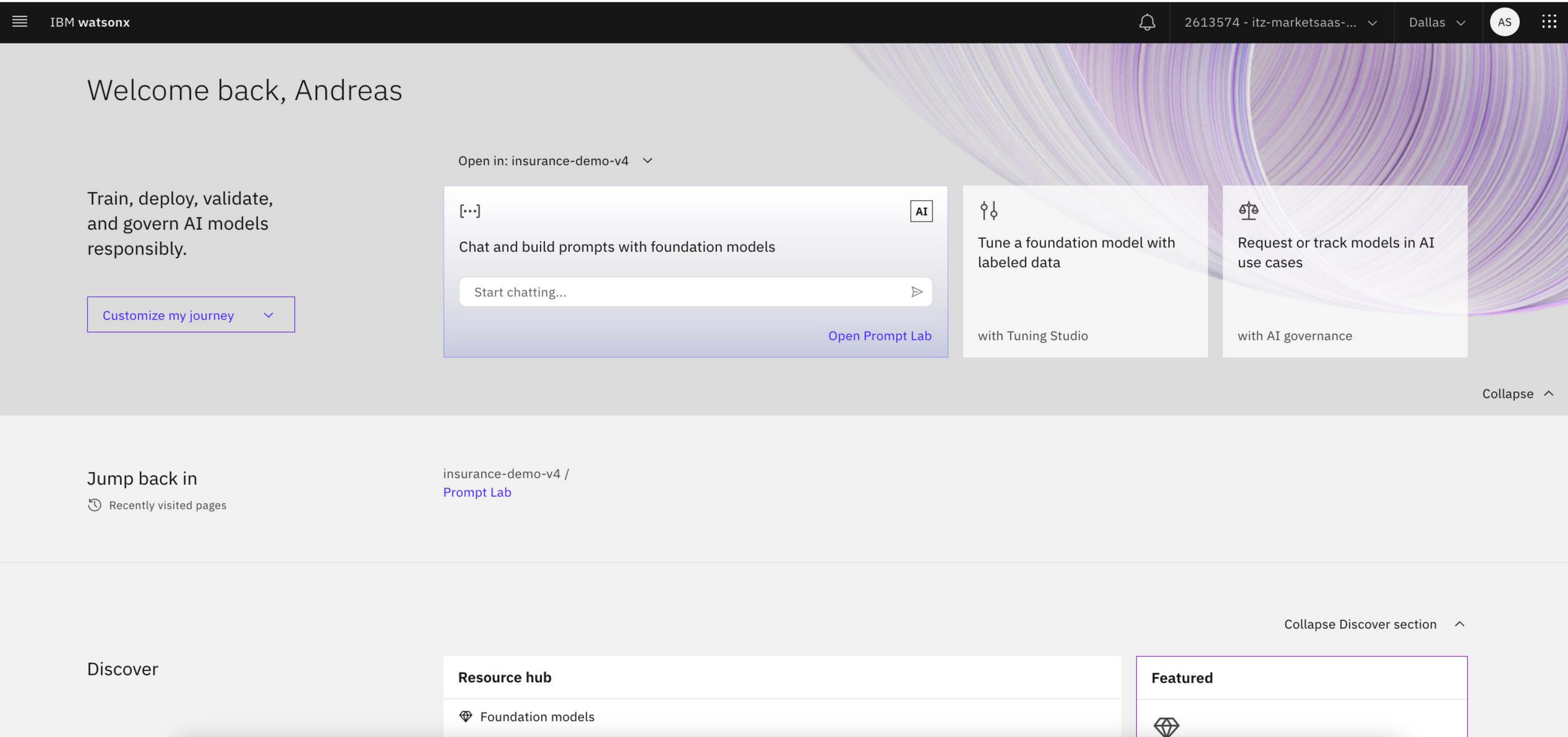
Overview	A fire broke out in the insured's garage, resulting in significant damage to the insured vehicle, a Honda Civic

Next steps

Steps to Remediation	☐ Verify the insured's policy coverage to ensure that it includes coverage for fire damage and the necessary repairs. ☐ 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. ☐ 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. ☐ 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. ☐ 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. ☐ 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

AI development
Business user can request an AI use case

Business User
Use case definition
& request

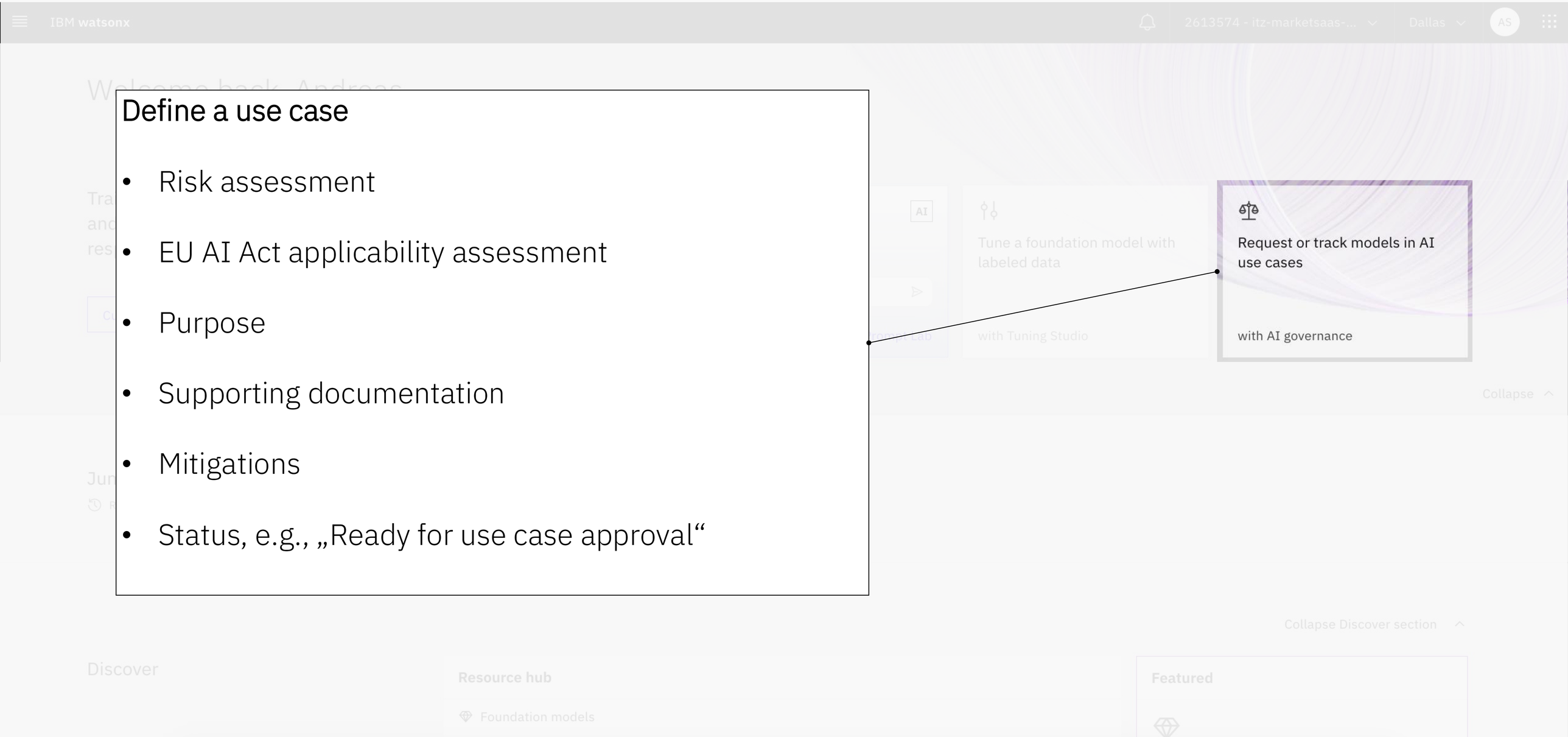


AI development
Business user can request an AI use case

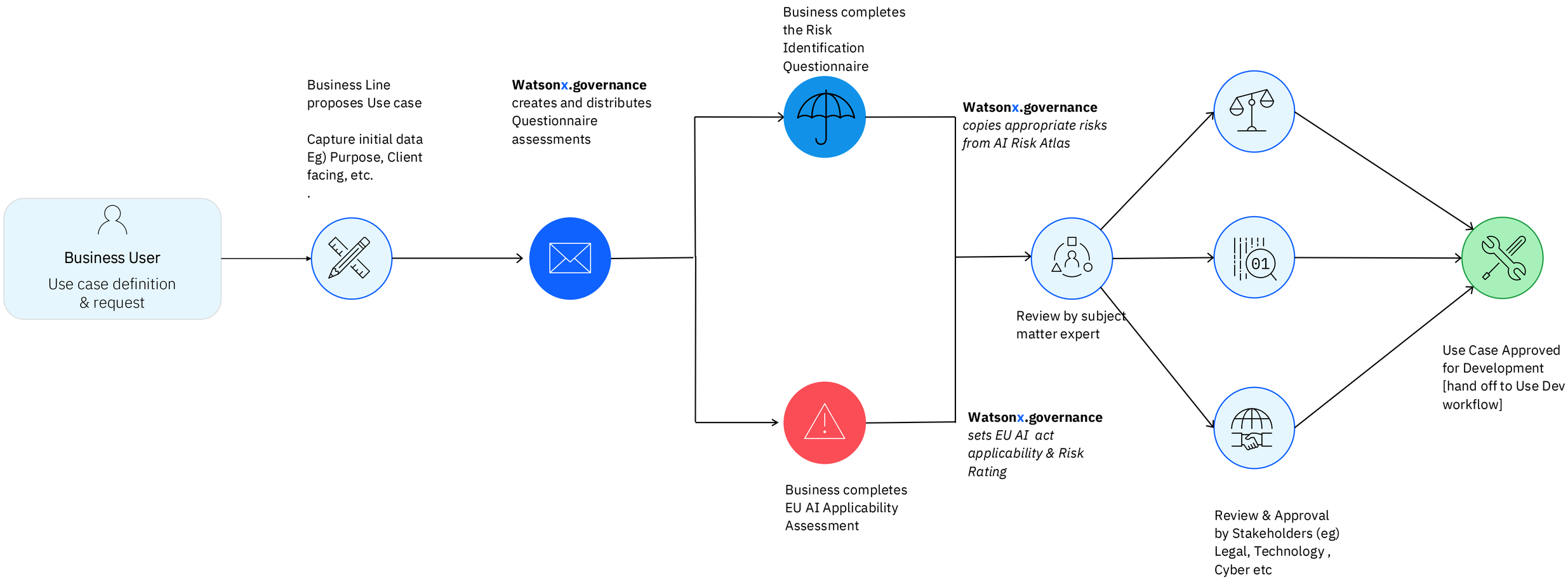
Business User
Use case definition & request

Define a use case

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



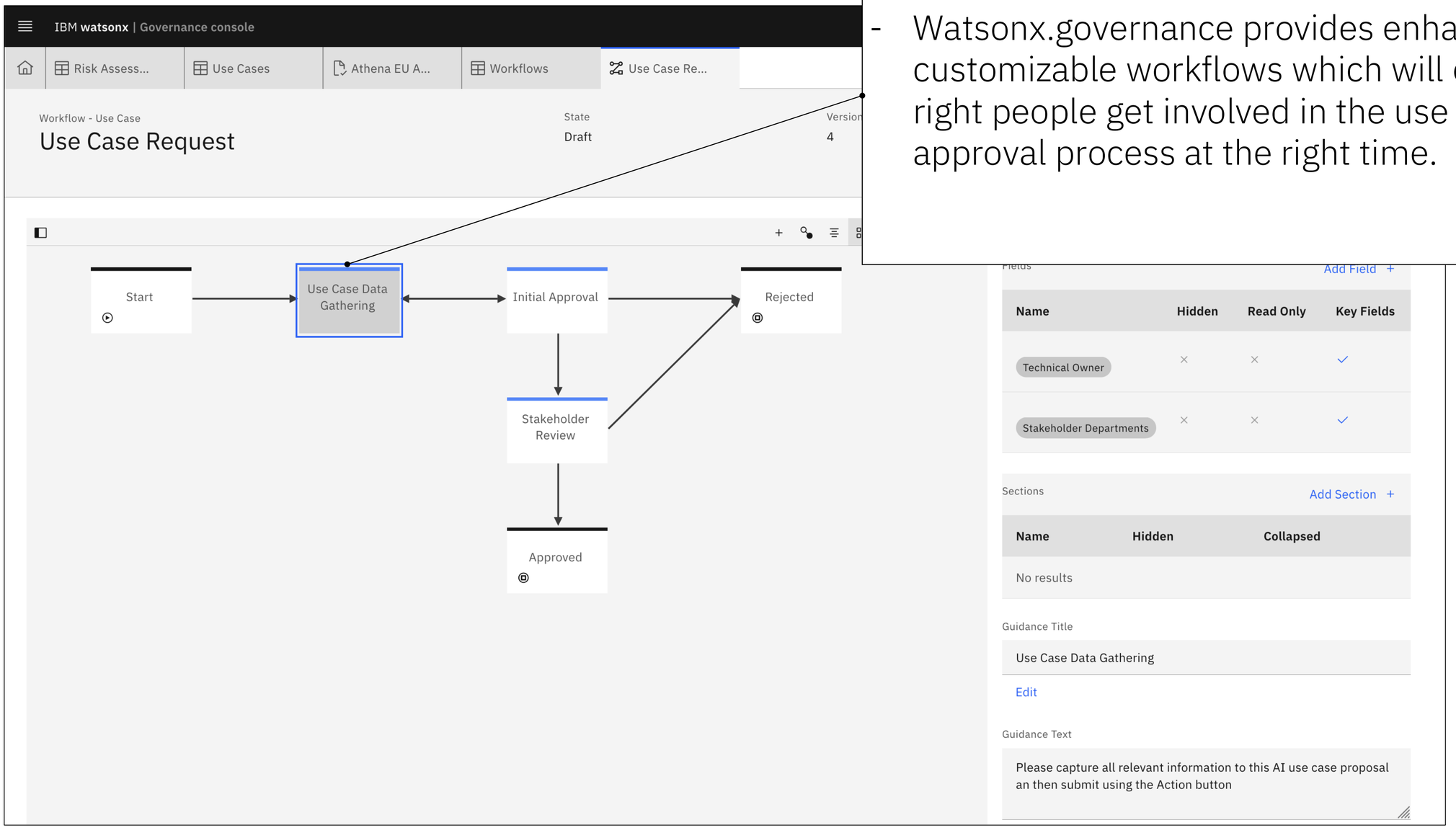
AI governance
Define a use case process



AI governance
Define a use case process

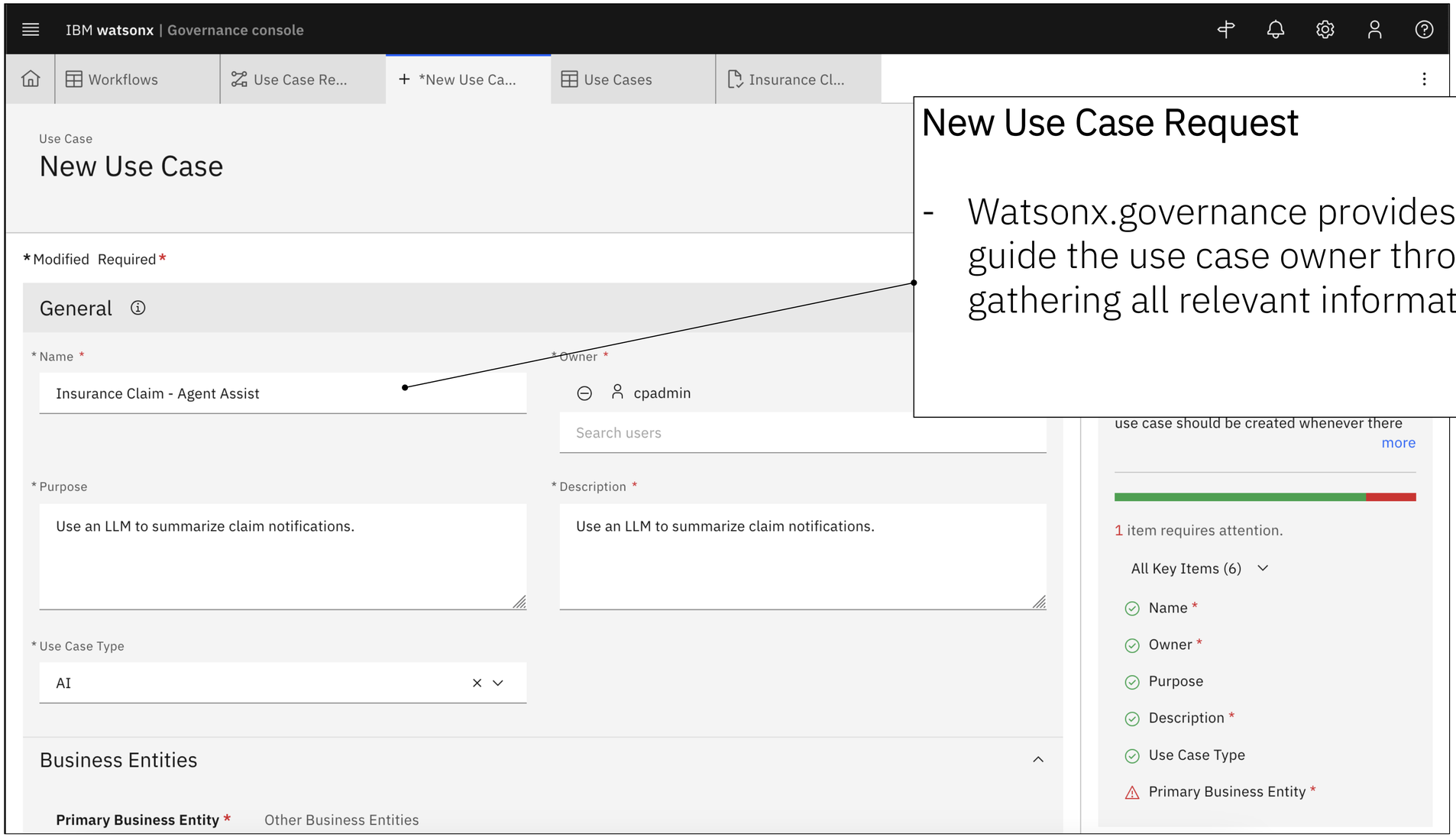
Workflow Management

- 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.



AI governance
New use case request

Business User
Use case definition & request



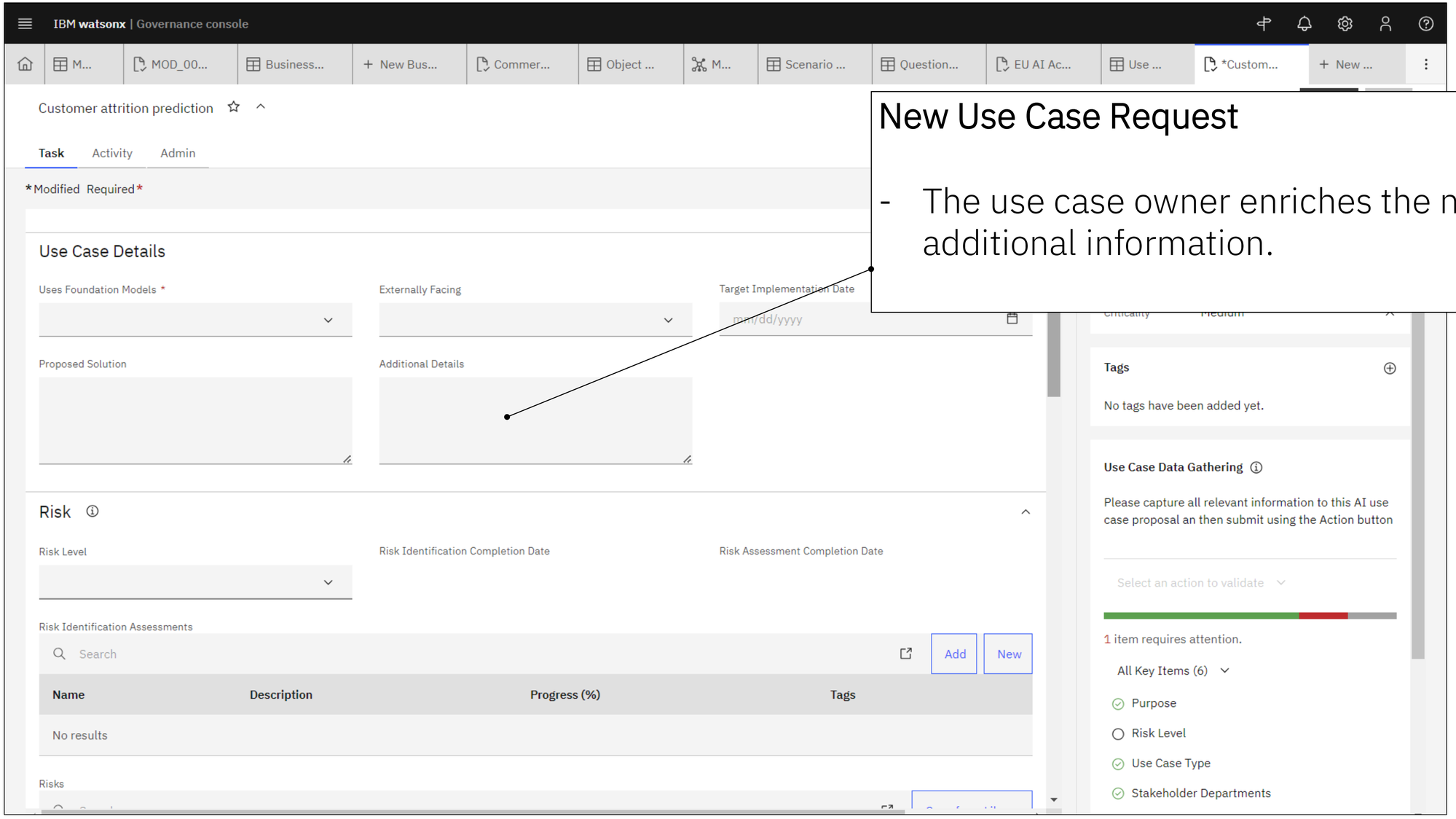
New Use Case Request

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

AI governance
New use case request

Business User

Use case definition & request



New Use Case Request

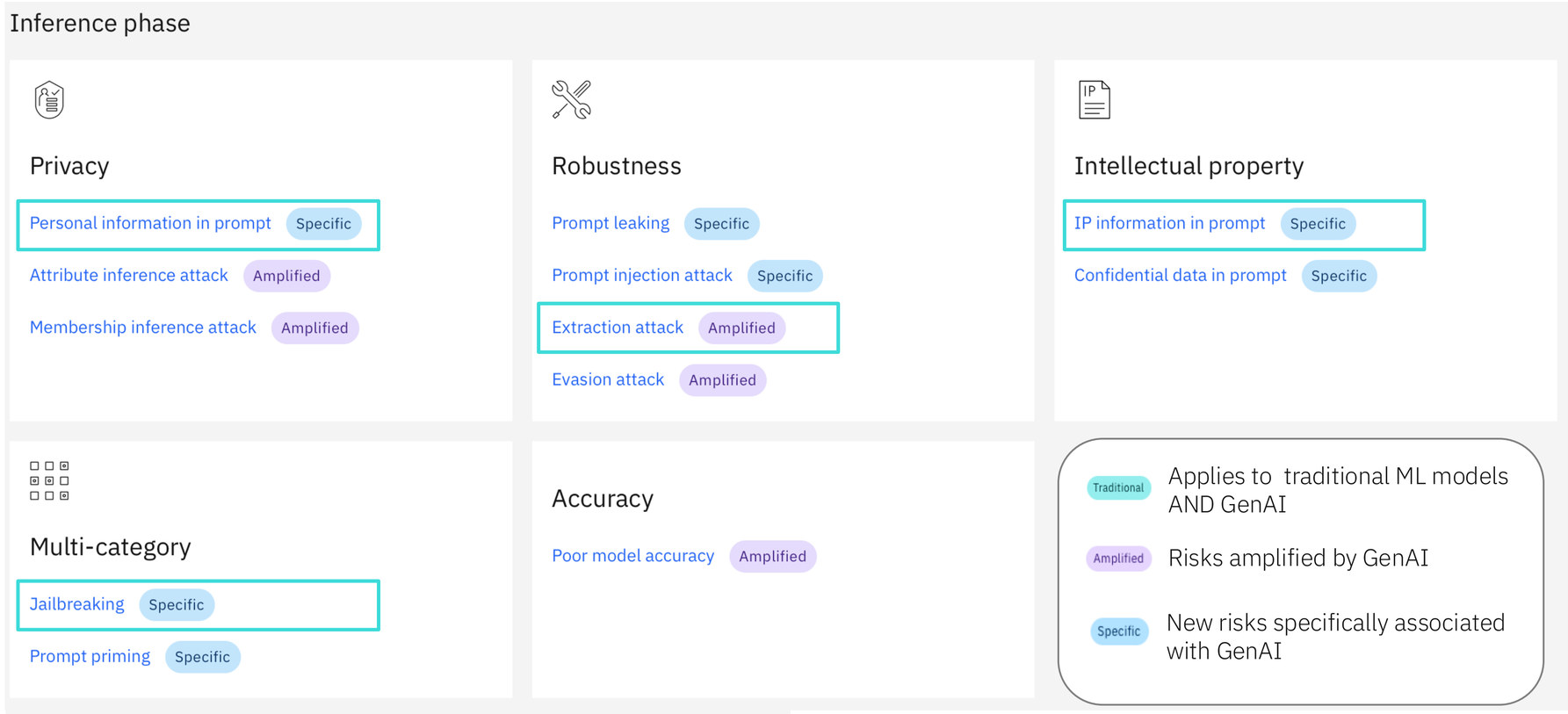
- The use case owner enriches the new use case with additional information.

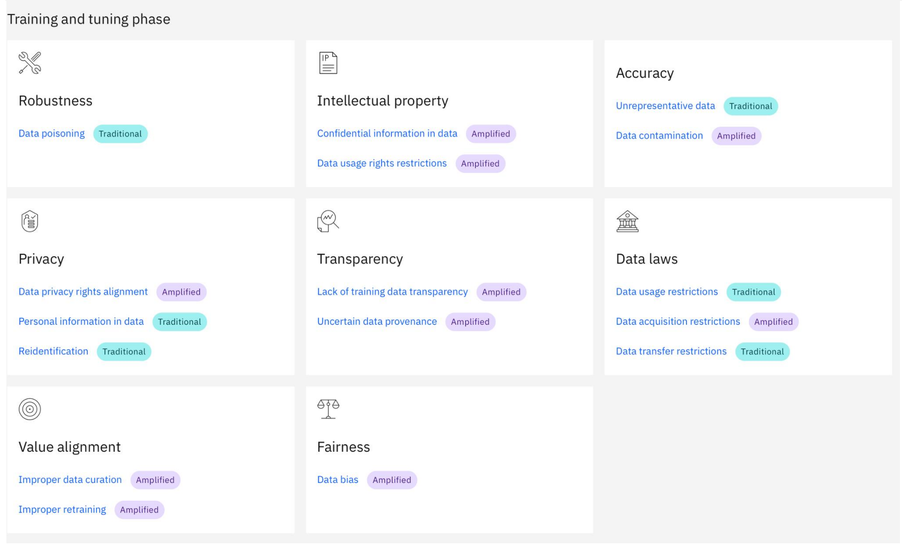
AI Risks

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





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

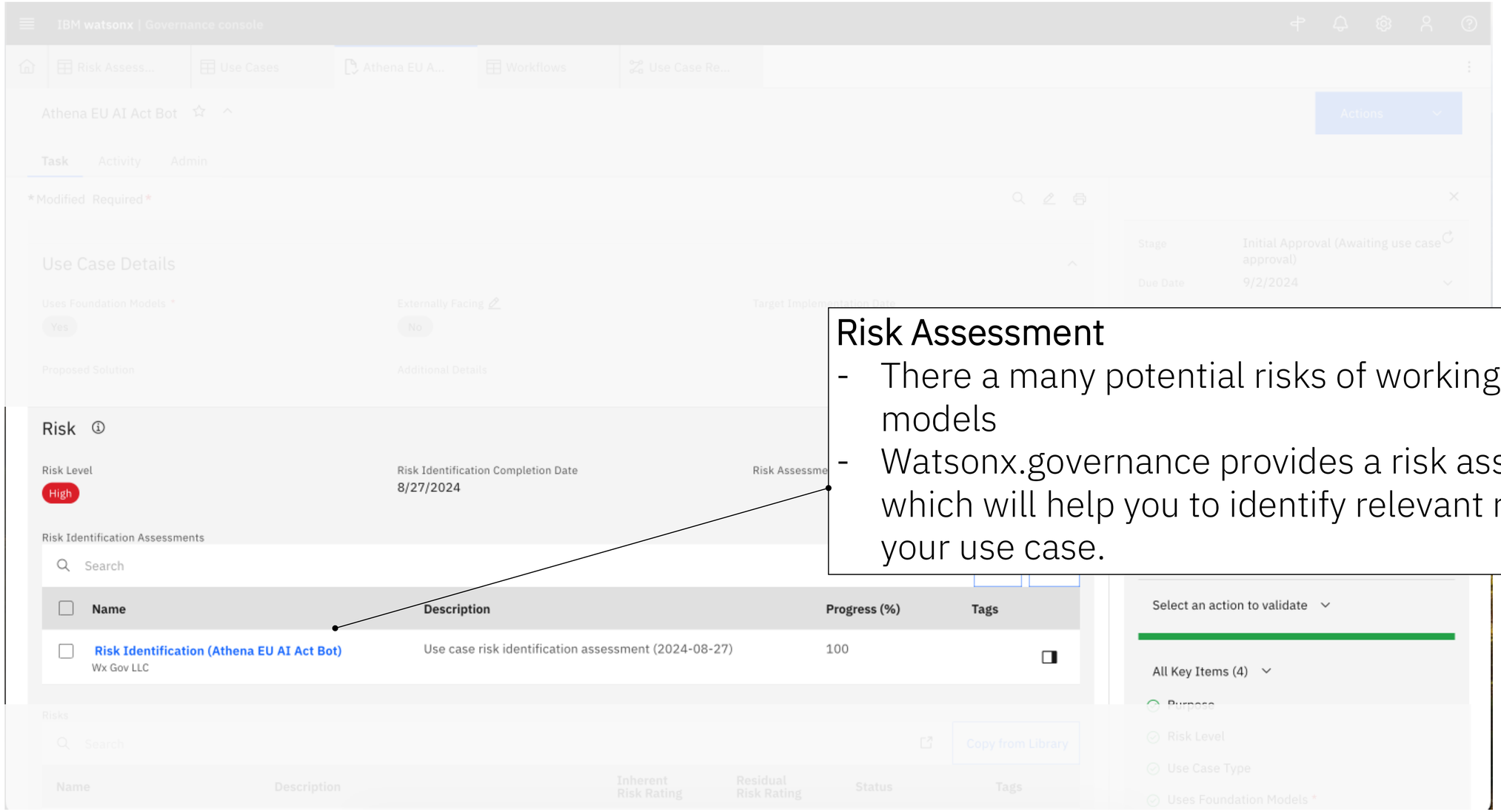
Traditional — Applies to traditional ML models AND GenAI

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

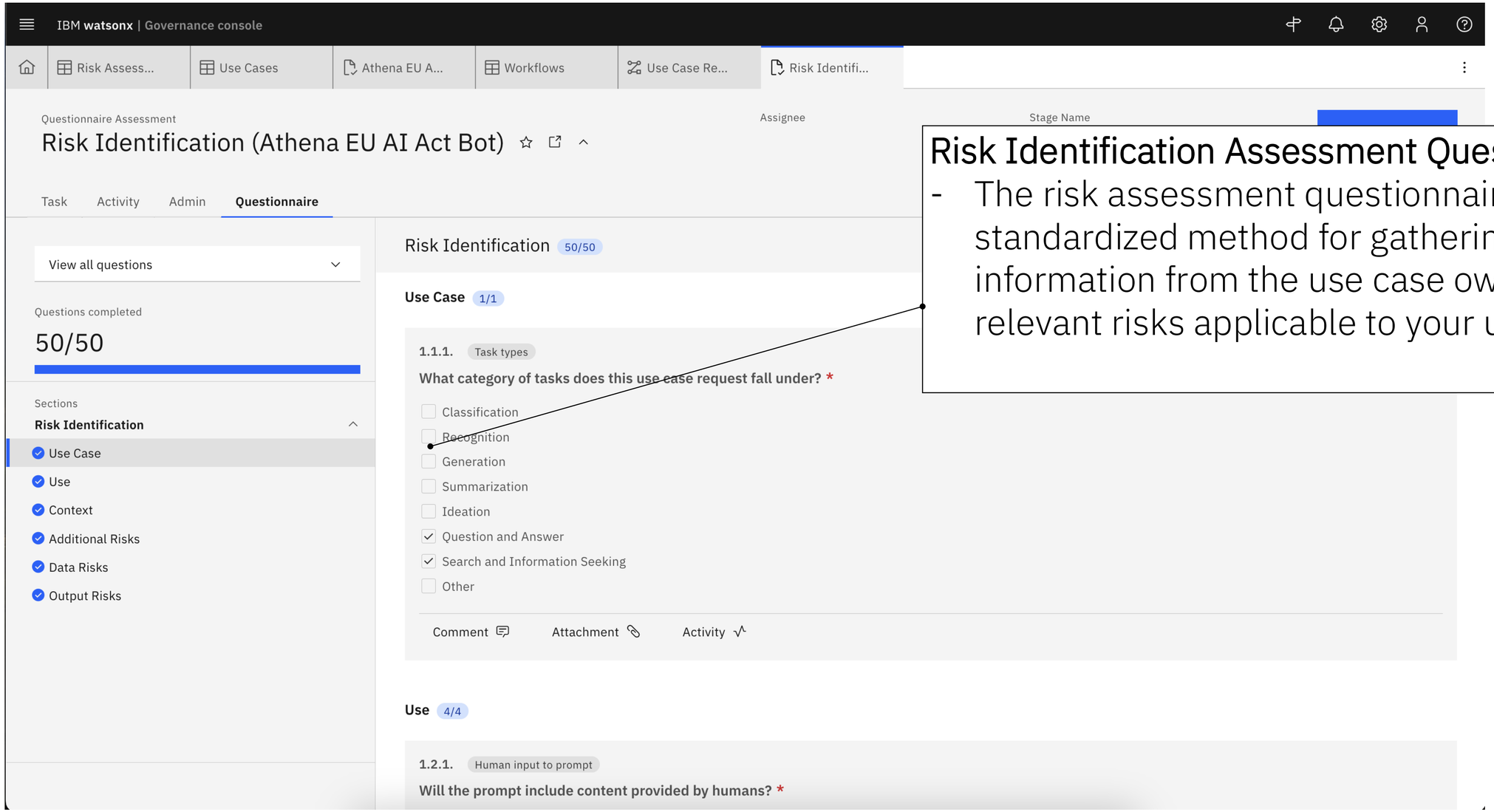


Risk Assessment

- There a 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



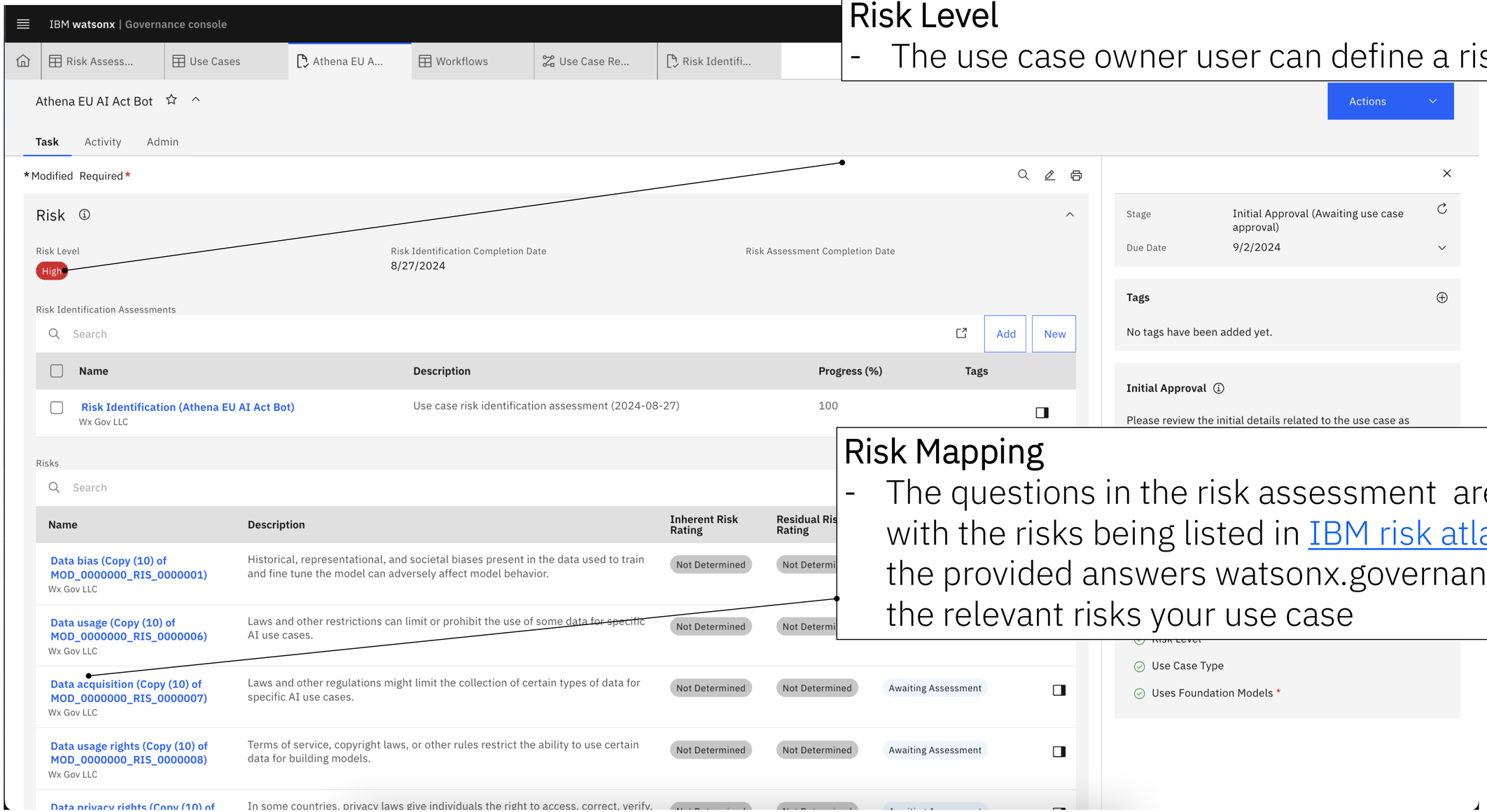
Risk Identification Assessment Questionaire

- 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



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.

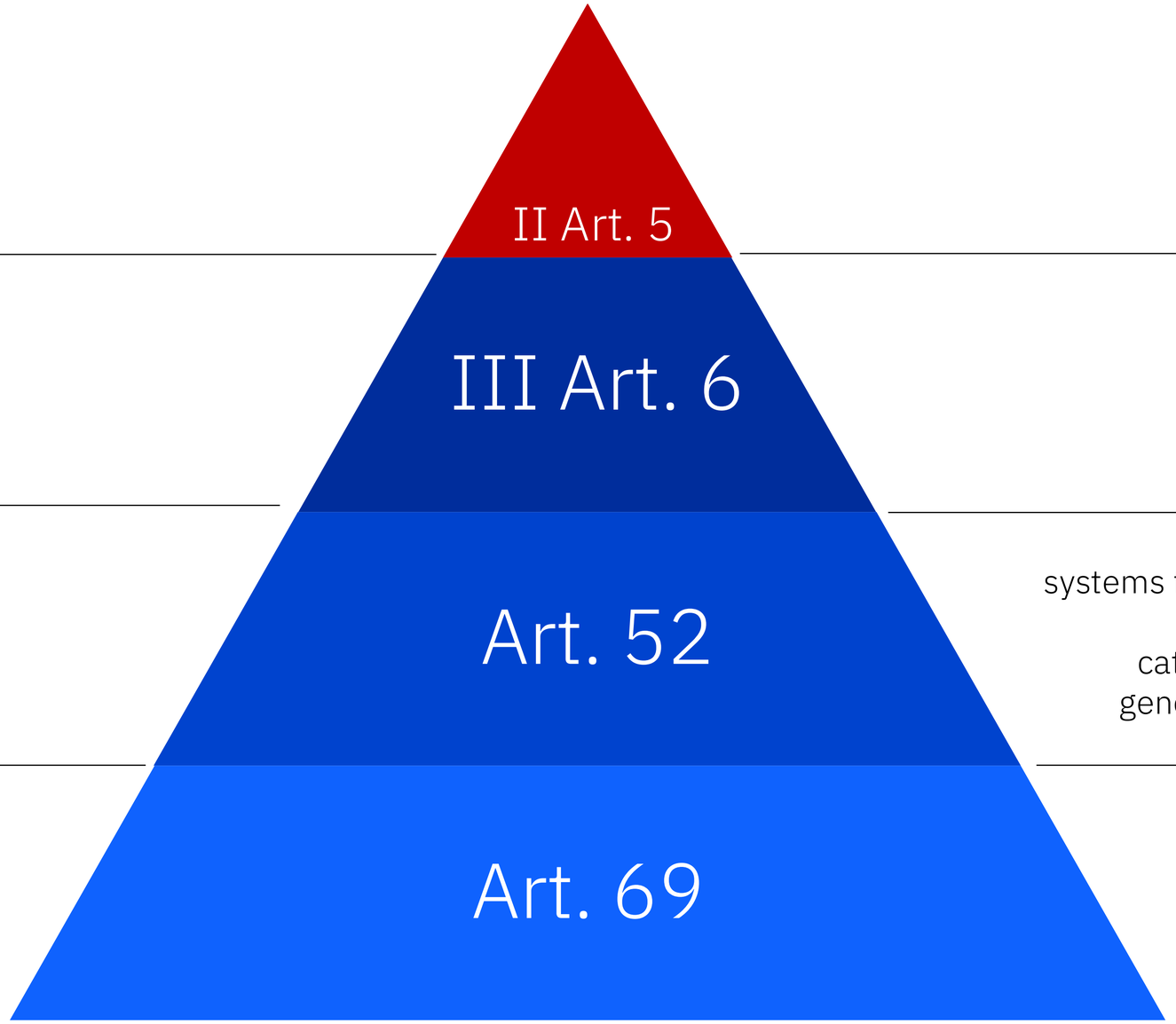
03
Ensure appropriate certain degree of transparency.

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



Prohibited

Unacceptable risk
harmful AI practices considered to be a clear threat to people's safety, livelihoods and rights

Regulated

High risk
AI systems that create adverse impact on people's safety or their fundamental rights

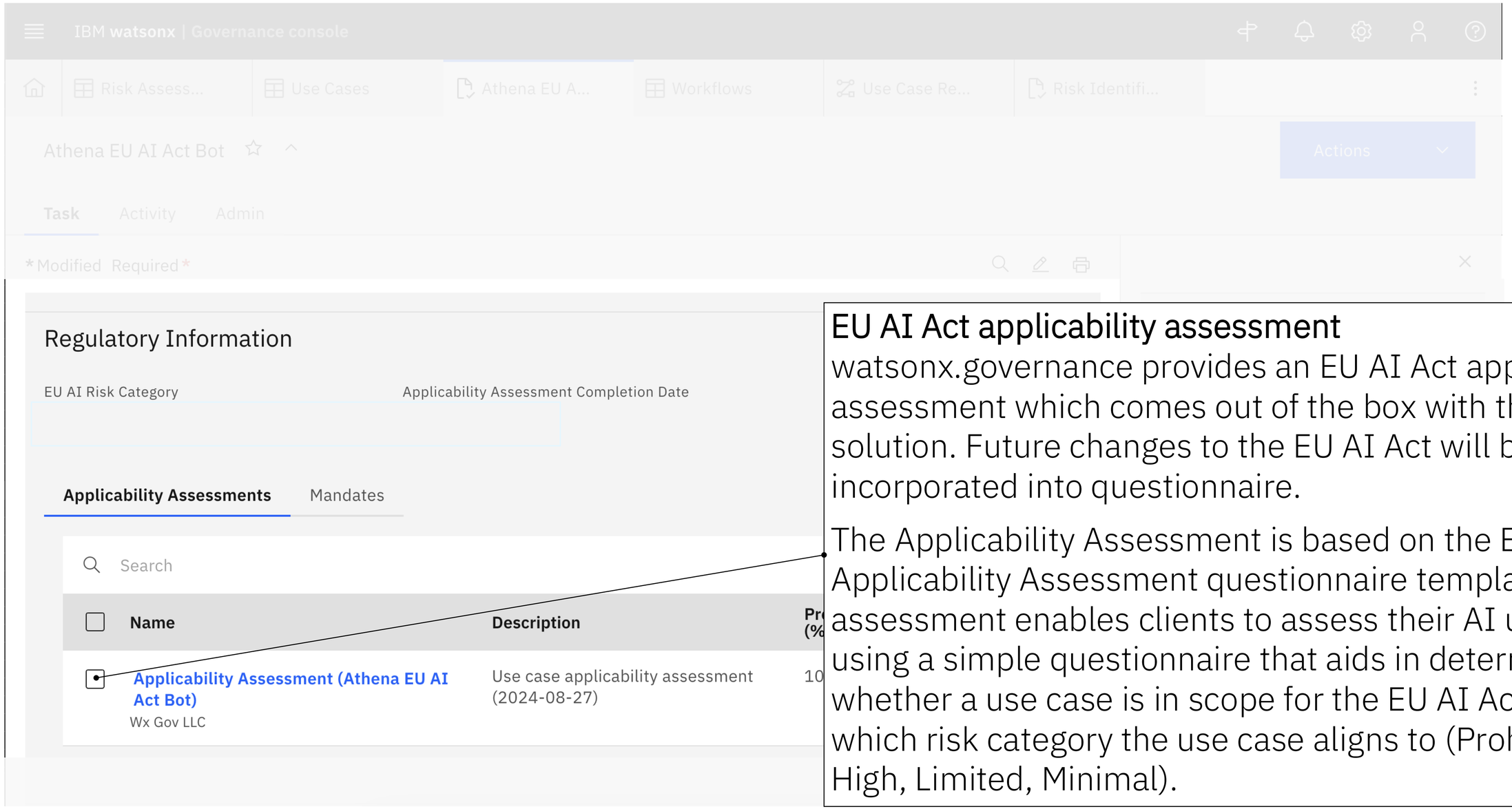
Transparency

Limited risk
systems that interacts with humans (i.e. chatbots), emotion recognition systems, biometric categorization systems, and AI systems that generate or manipulate image, audio or video content (i.e. deepfakes)

Voluntary code of conduct

Low and minimal risk
All other AI systems presenting only low or minimal risk

AI governance
Define a use case – Risk assessment



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.

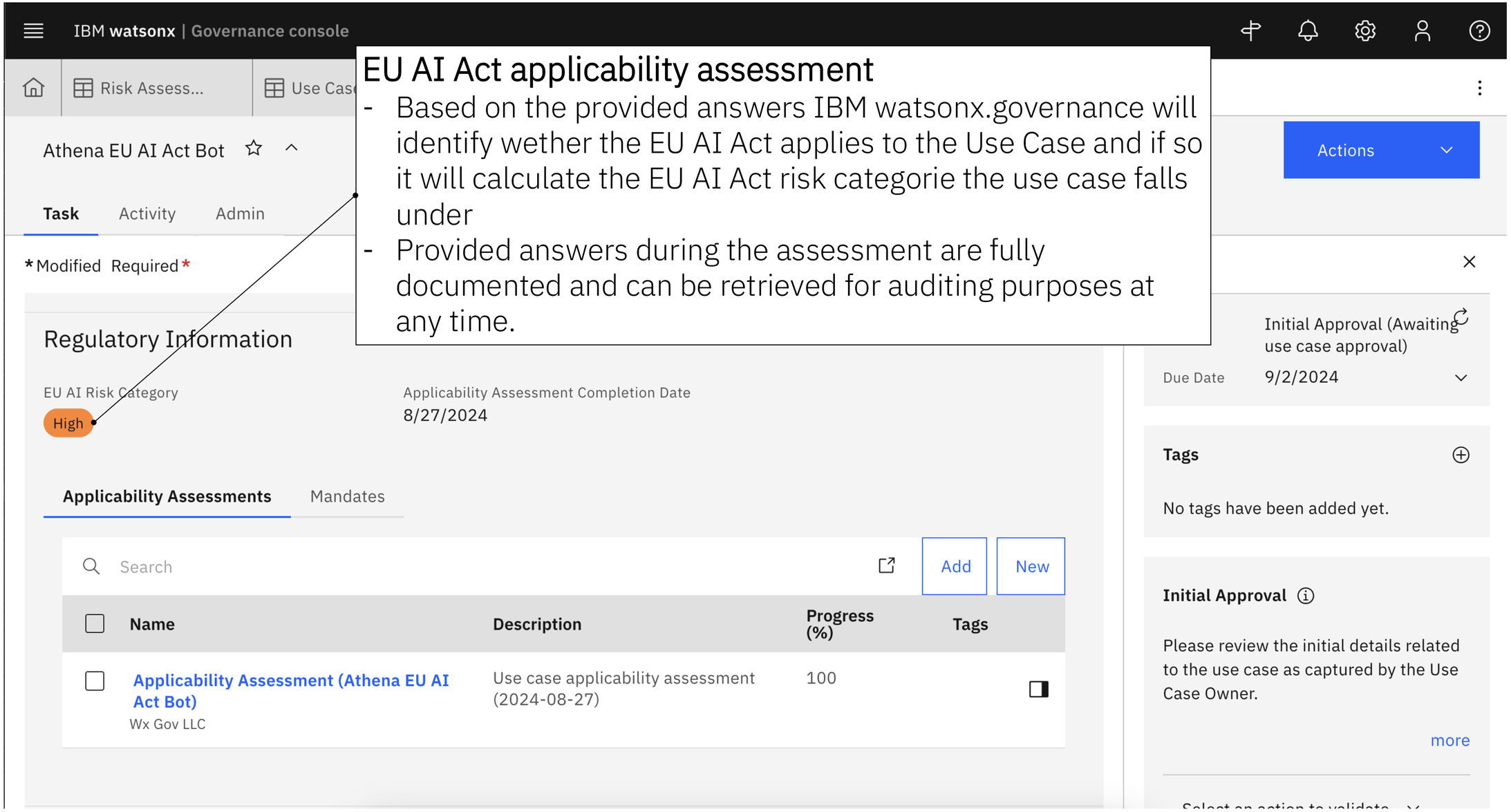
The Applicability Assessment is based on the EU AI Act Applicability Assessment questionnaire template. This assessment enables clients to assess their AI uses cases using a simple questionnaire that aids in determining whether a use case is in scope for the EU AI Act and which risk category the use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

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

Business User
Use case definition & request

EU AI Act applicability assessment

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

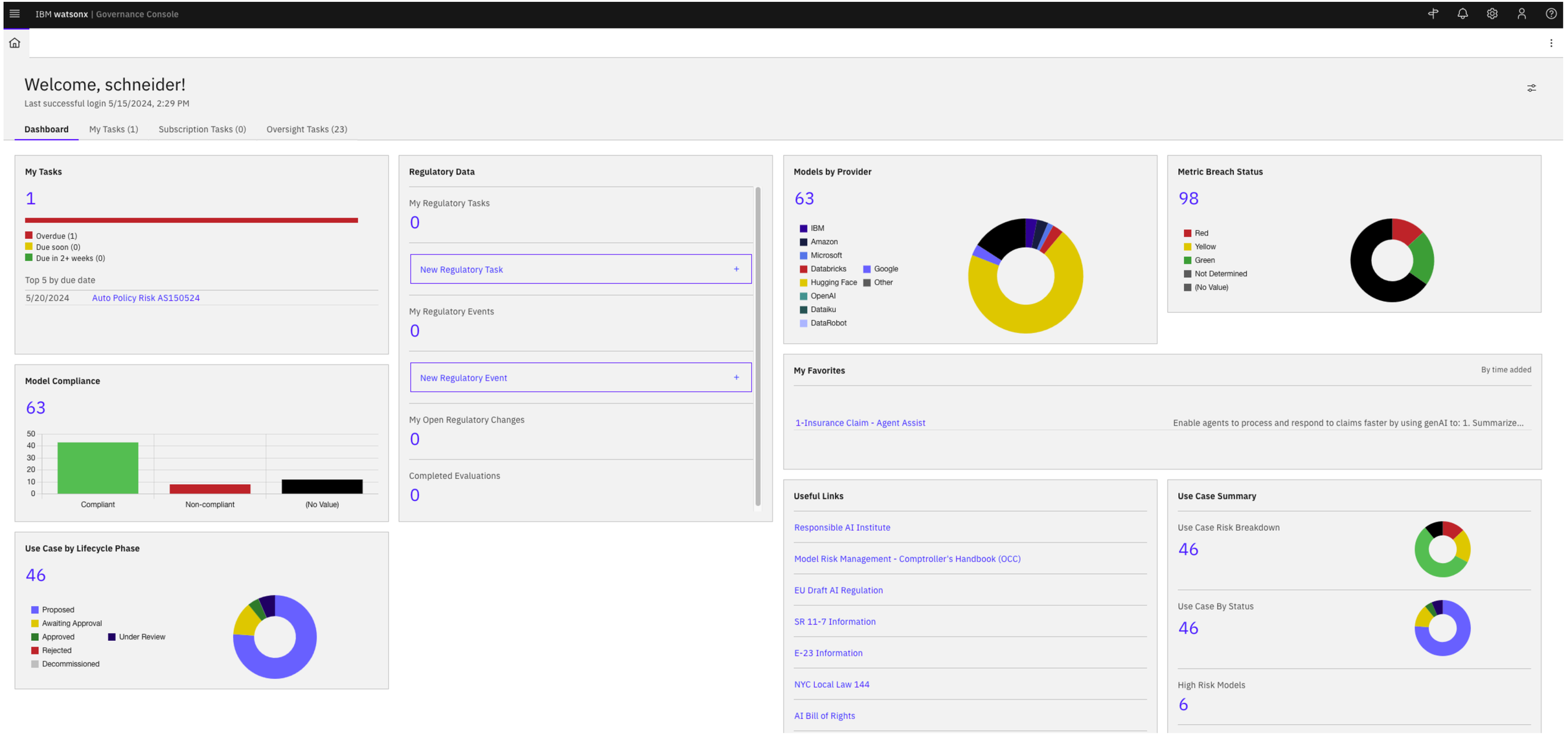


AI governance

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

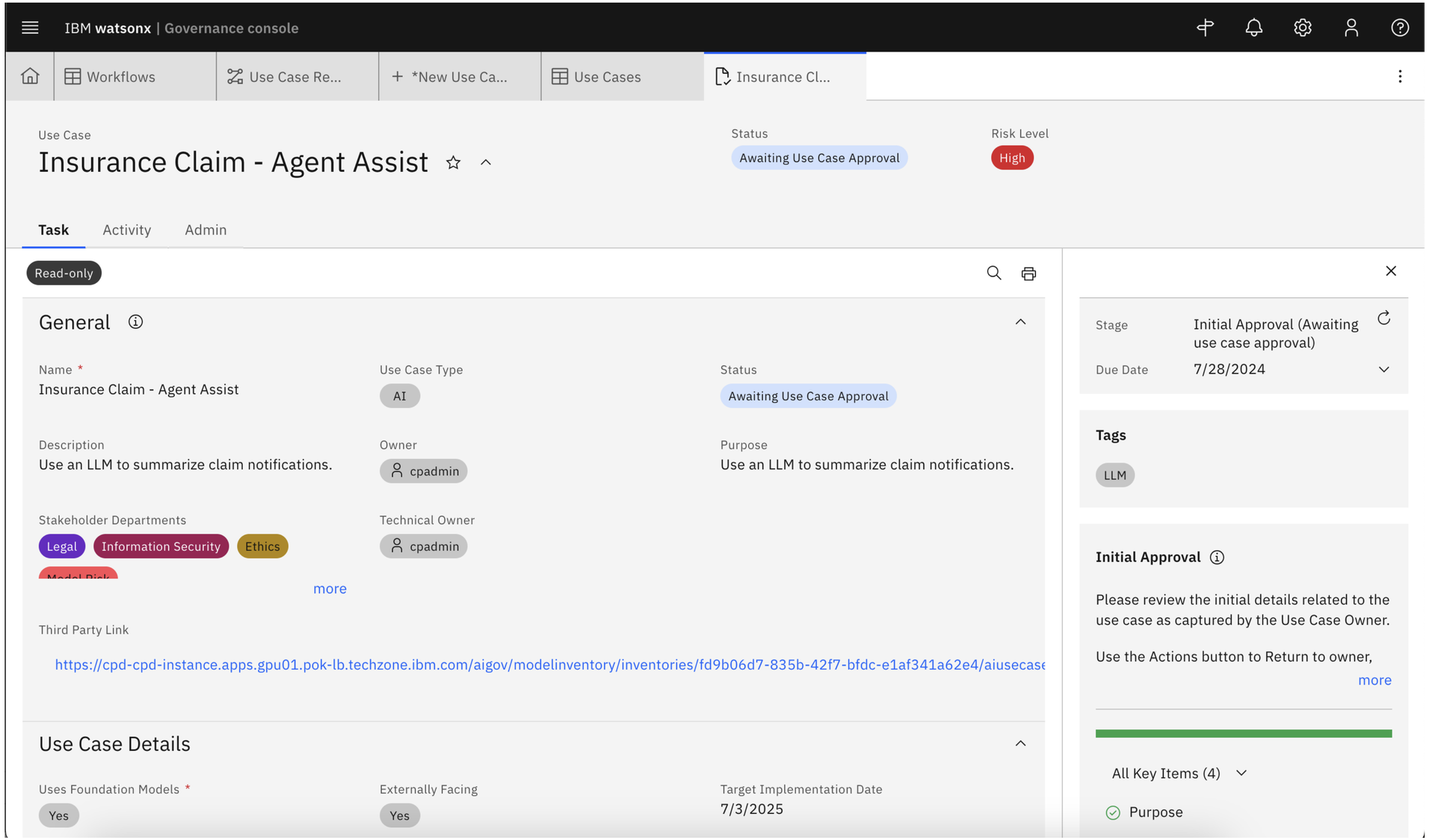
Compliance / Risk Officer

Governance & risk assessment



AI governance
Define a use case – Documented use case

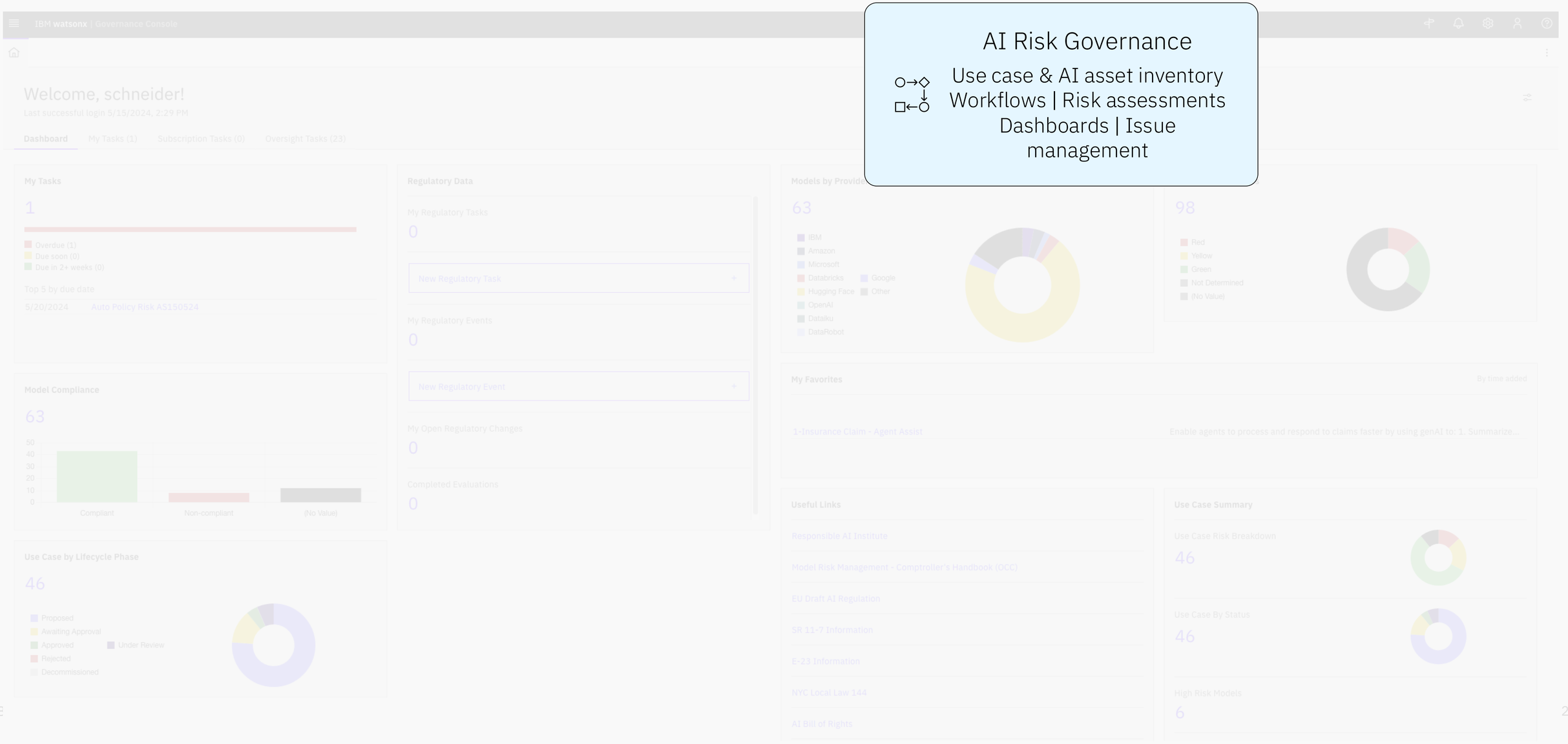
Business User
Use case definition & request



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

Compliance / Risk Officer

Governance & risk assessment



AI Risk Governance

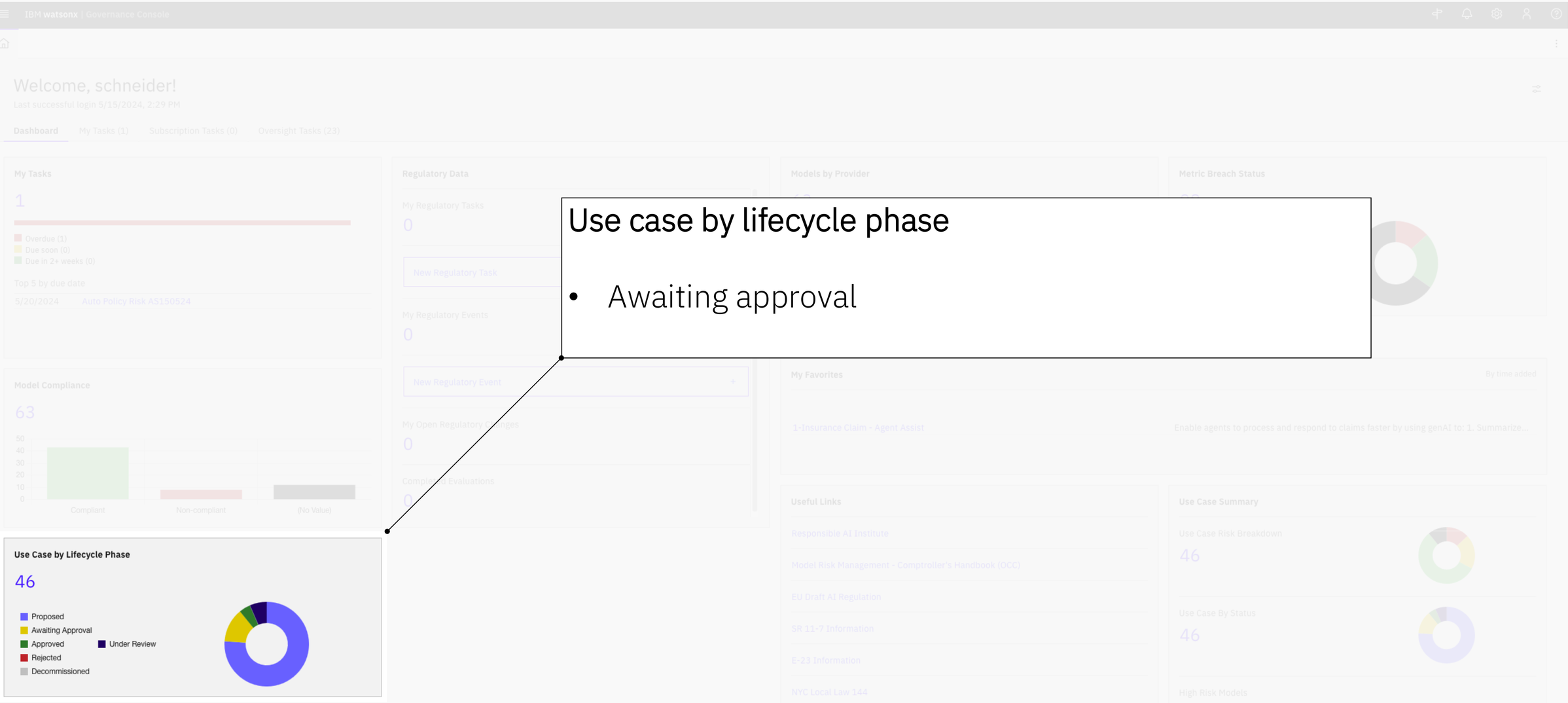
Use case & AI asset inventory
Workflows | Risk assessments
Dashboards | Issue management

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

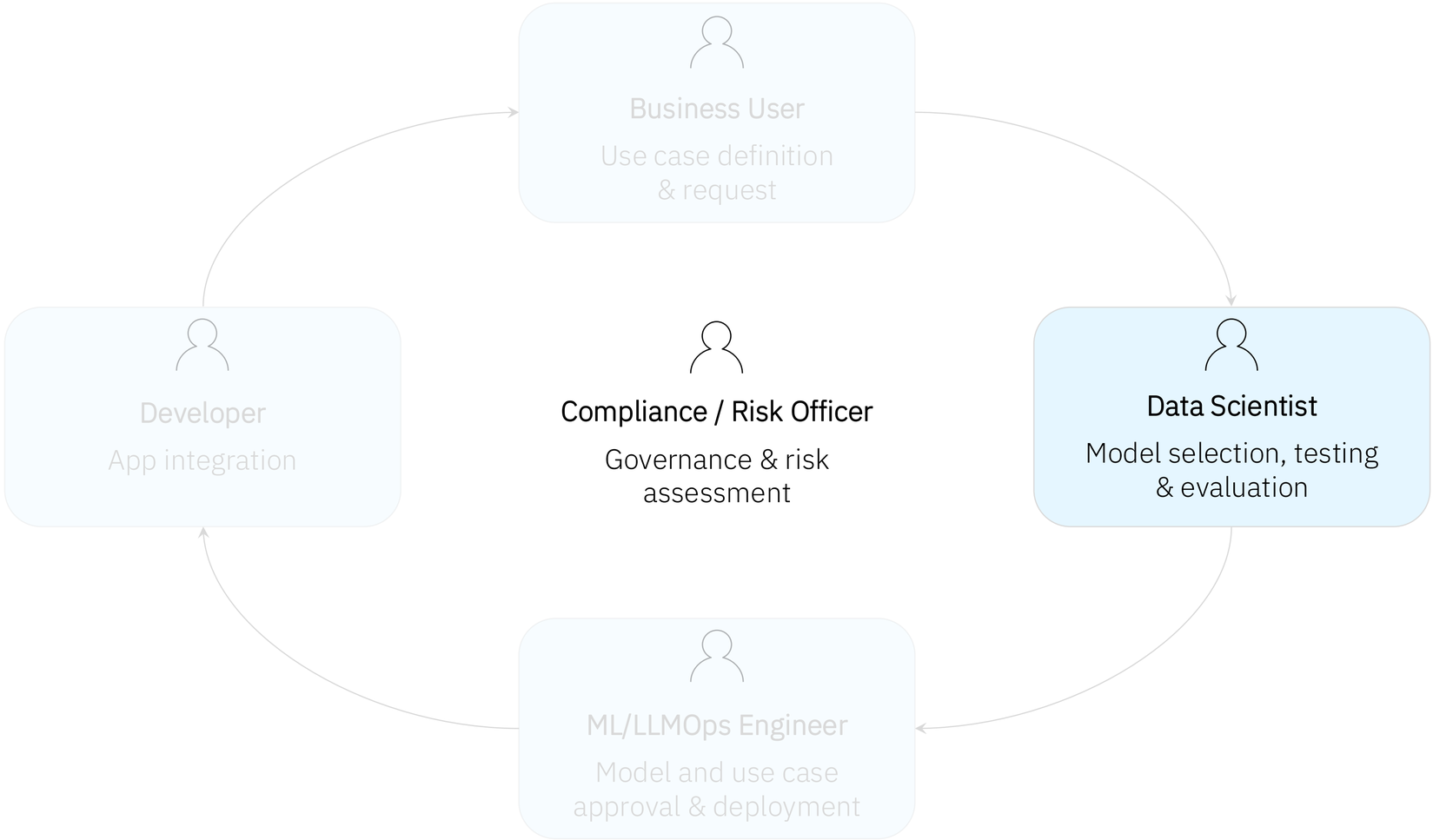
Compliance / Risk Officer
Governance & risk assessment

Use case by lifecycle phase

- Awaiting approval

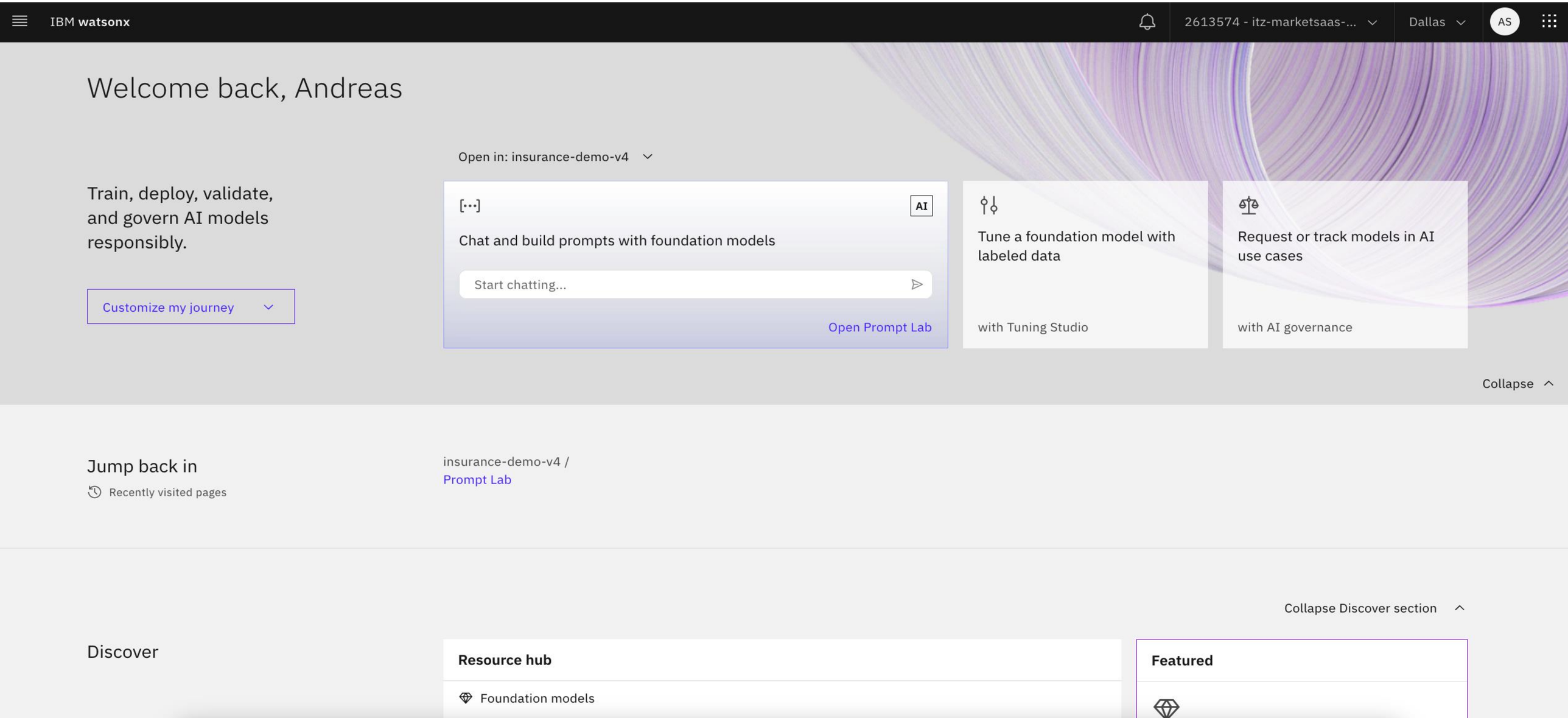


AI development & AI governance
Summarization of car insurance claims



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

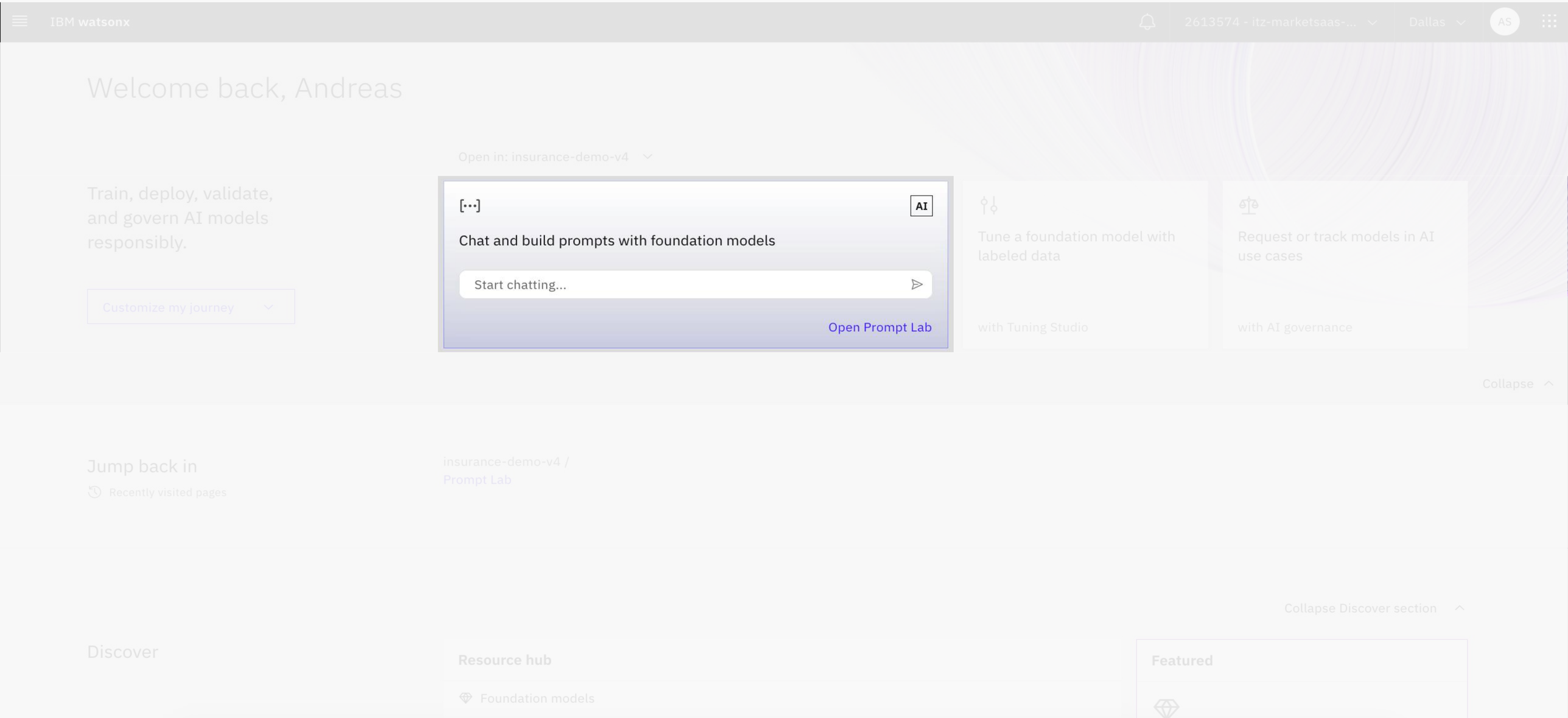
Data Scientist
Model selection, testing & evaluation



AI development

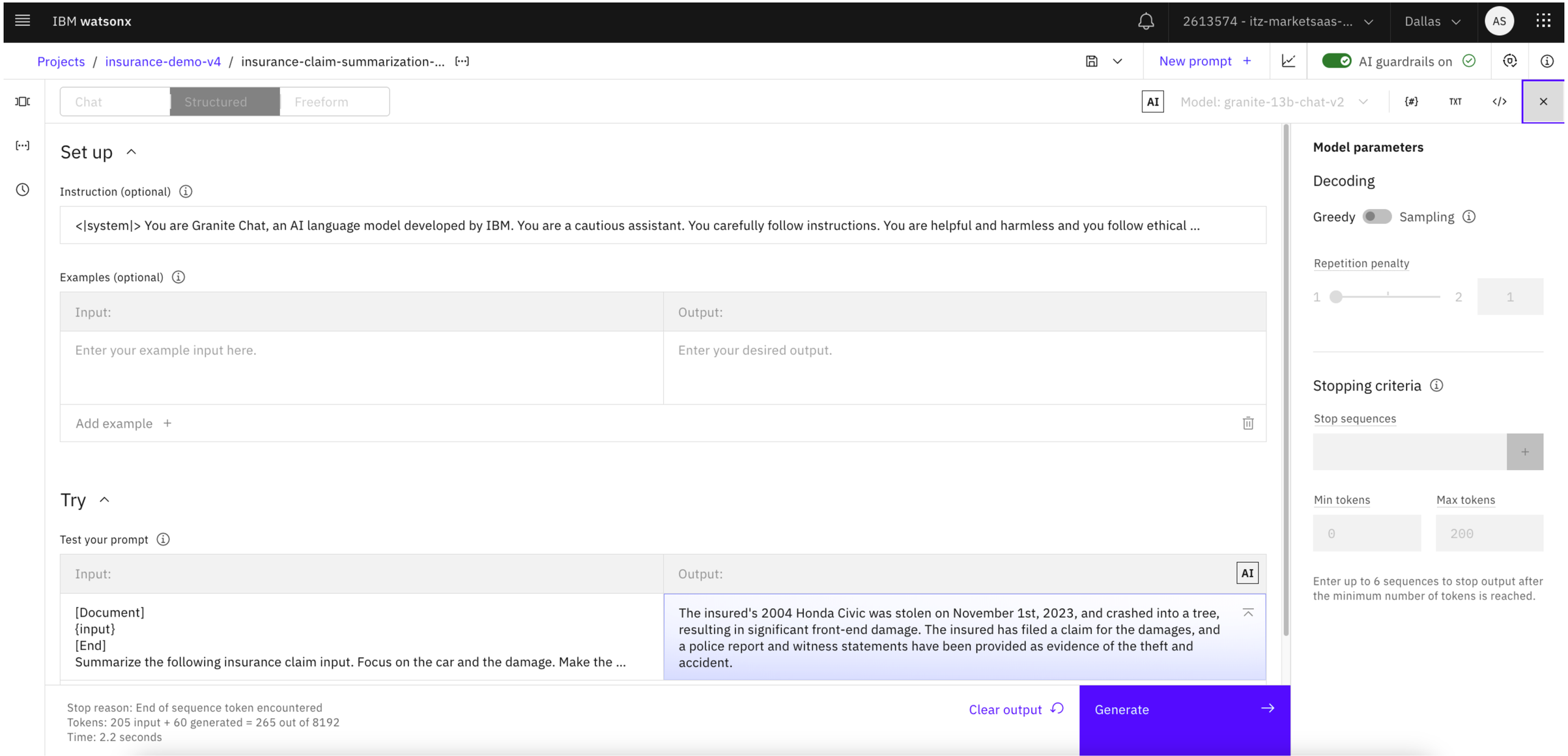
Data Scientist experiments with different prompts and foundation models to evaluate if the desired information can be derived from the claim

Data Scientist
Model selection, testing & evaluation



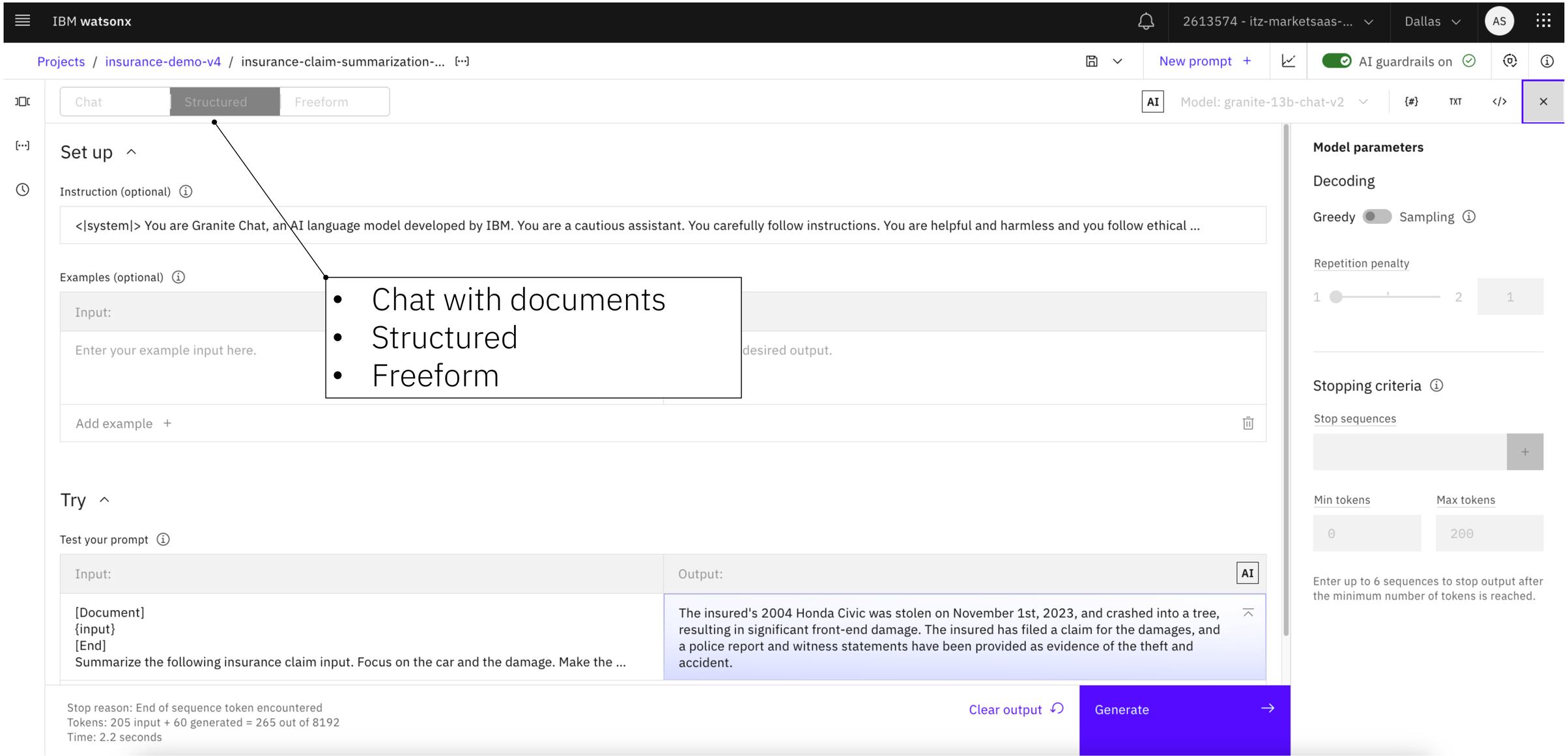
AI development
Data Scientist experiments with different prompts and foundation models to evaluate if the desired information can be derived from the claim

Data Scientist
Model selection, testing & evaluation



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

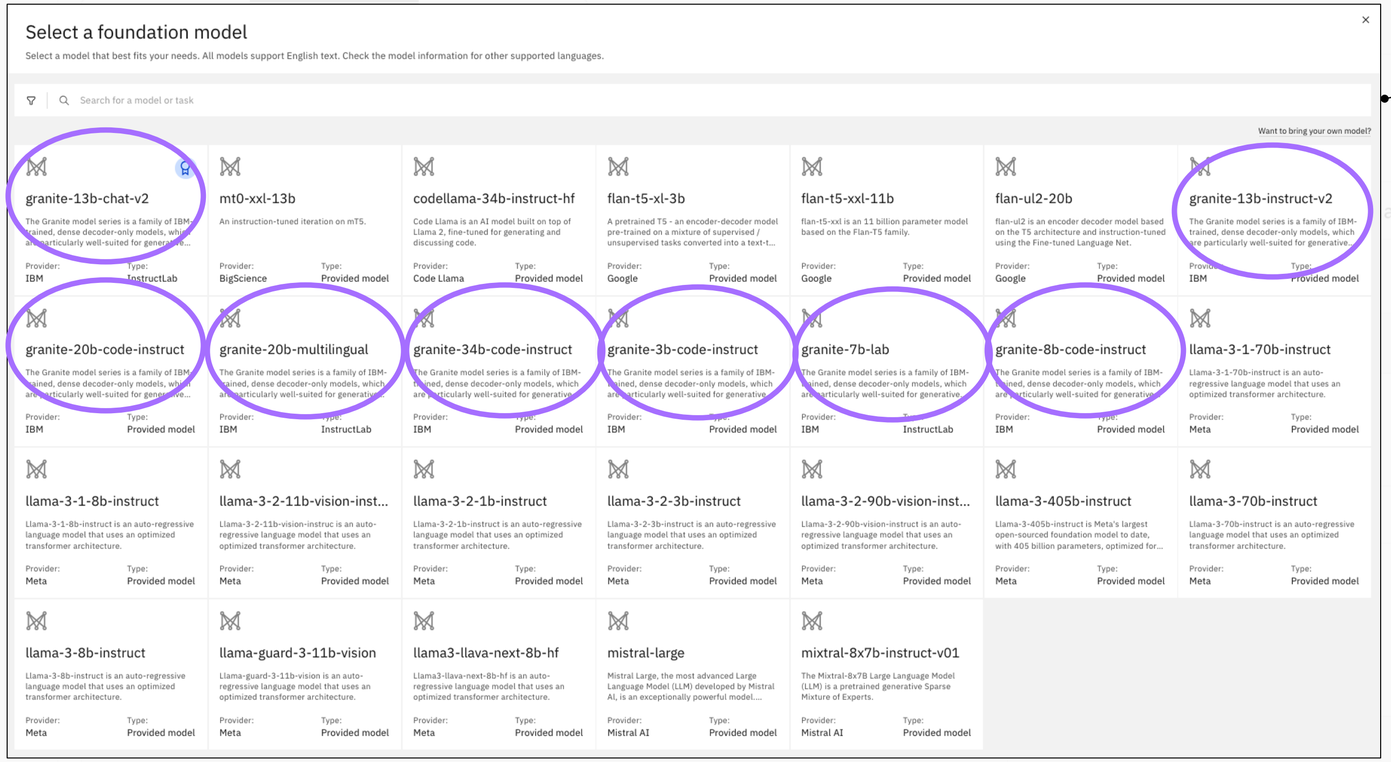
Data Scientist
Model selection, testing & evaluation



- Chat with documents
- Structured
- Freeform

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



Model: granite-13b-chat-v2

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

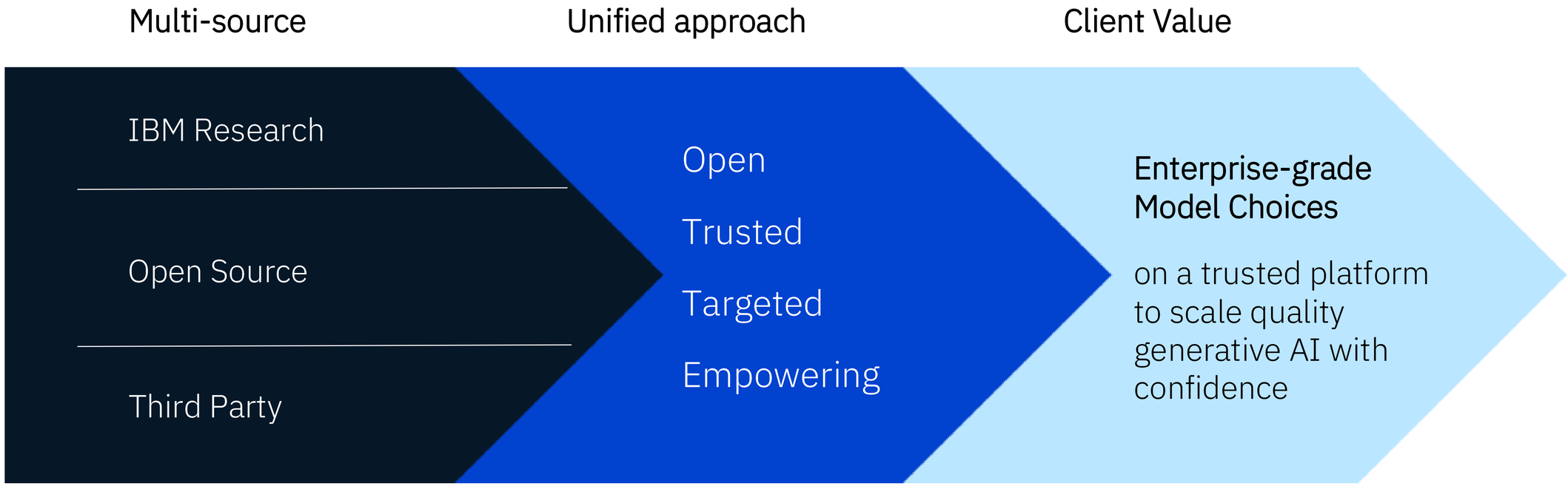
Deploying custom foundation models to watsonx.ai
https://community.ibm.com/community/user/watsonx/blogs/saloni-saluja/2024/03/28/deploying-custom-foundation-models-to-watsonxai

AI development: Foundation models
Open, multi-model approach to delivering enterprise-grade foundation models

Data Scientist
Model selection, testing & evaluation

Compliance / Risk Officer
Governance & risk assessment

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



Hugging Face

Meta AI

MISTRAL AI_

InstructLab

Red Hat

watsonx

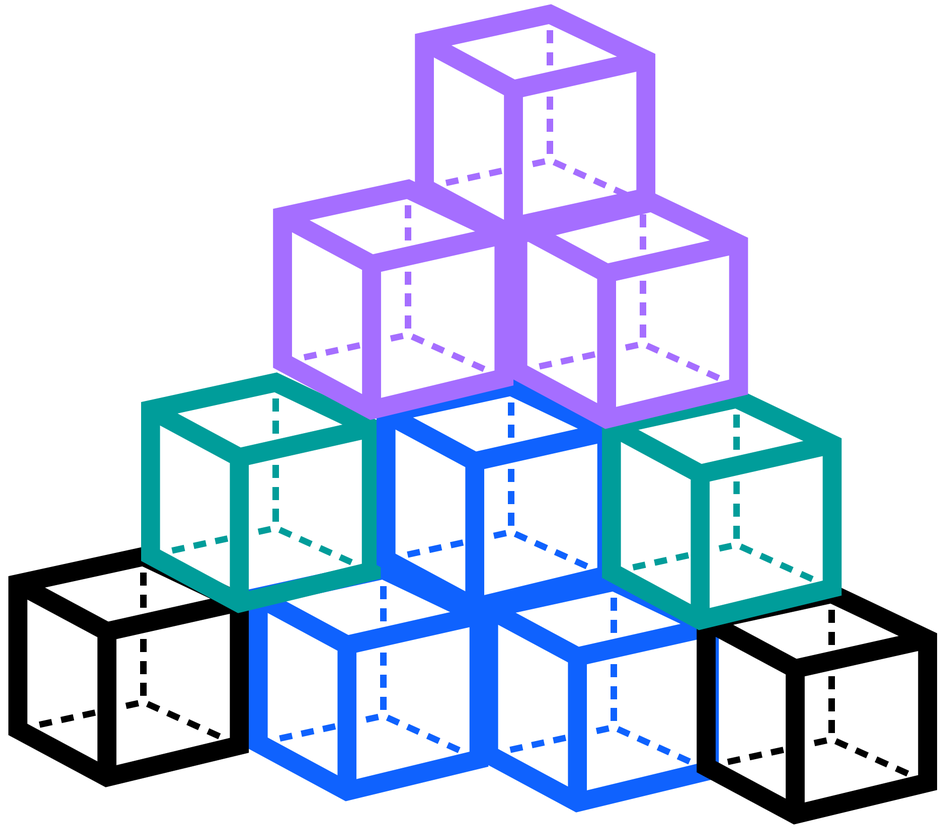
AI development: Foundation models
Approach of selecting third-party / open-source models

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 [1, 2] through comprehensive data management across the training phase to build trusted models

Data Scientist
Model selection, testing & evaluation

Compliance / Risk Officer
Governance & risk assessment

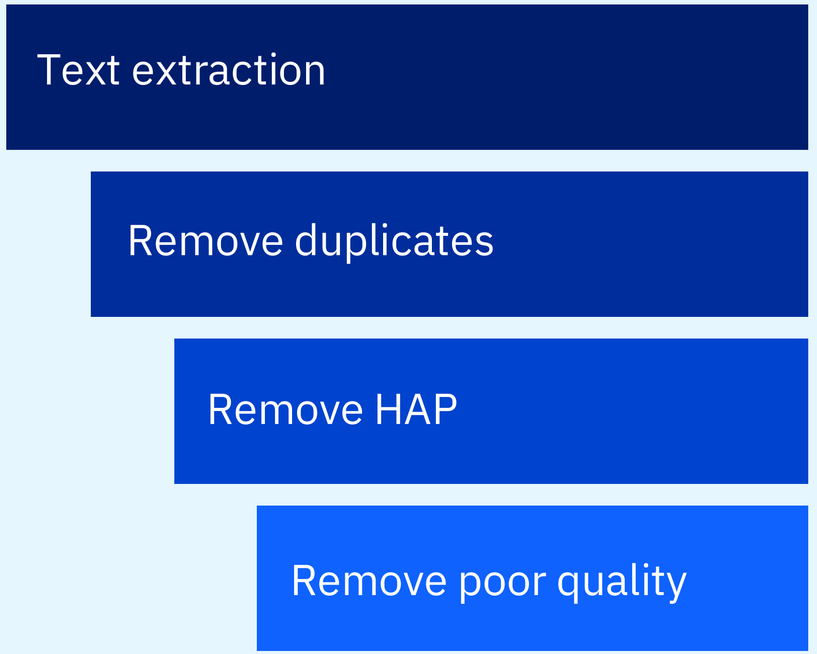
IBM's approach to AI model development is grounded in core principles of trust and transparency.

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



27.19TB of extracted data

7.36TB of deduplicated data

7.35TB of usable data

7.48TB of data for tokenization

2.5 trillion tokens used for training IBM Granite base models

TB = Terabytes

[1] Granite Foundation Models (Research paper)
[2] 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

External datasets

IBM internal datasets

Data collection and extraction

IBM Cloud Object Storage

Data deduplication

- Document ID generation
- Exact dedup
- Fuzzy dedup (threshold)
- Sentence splitting

Data annotation

- Language detection
- HAP filter
- Document quality
- URL blocking
- License detection (for code)

Data filtering

Tokenization

IBM Cloud Object Storage

FM training

Tuning and customization

Customer datasets

Dataset Acquisition

Dataset Preprocessing (model neutral)

Data Preprocessing (model specific)

AI development: Foundation models
Granite models are targeted at specific industry domains like finance and legal

Data Scientist
Model selection, testing & evaluation

Compliance / Risk Officer
Governance & risk assessment

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

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 [4]	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

Source: Granite Foundation Models (Research paper)
https://www.ibm.com/downloads/cas/X9W4O6BM

AI development: Foundation models
IBM stands behind its Granite models with industry-leading IP indemnification

Compliance / Risk Officer

Governance & risk assessment

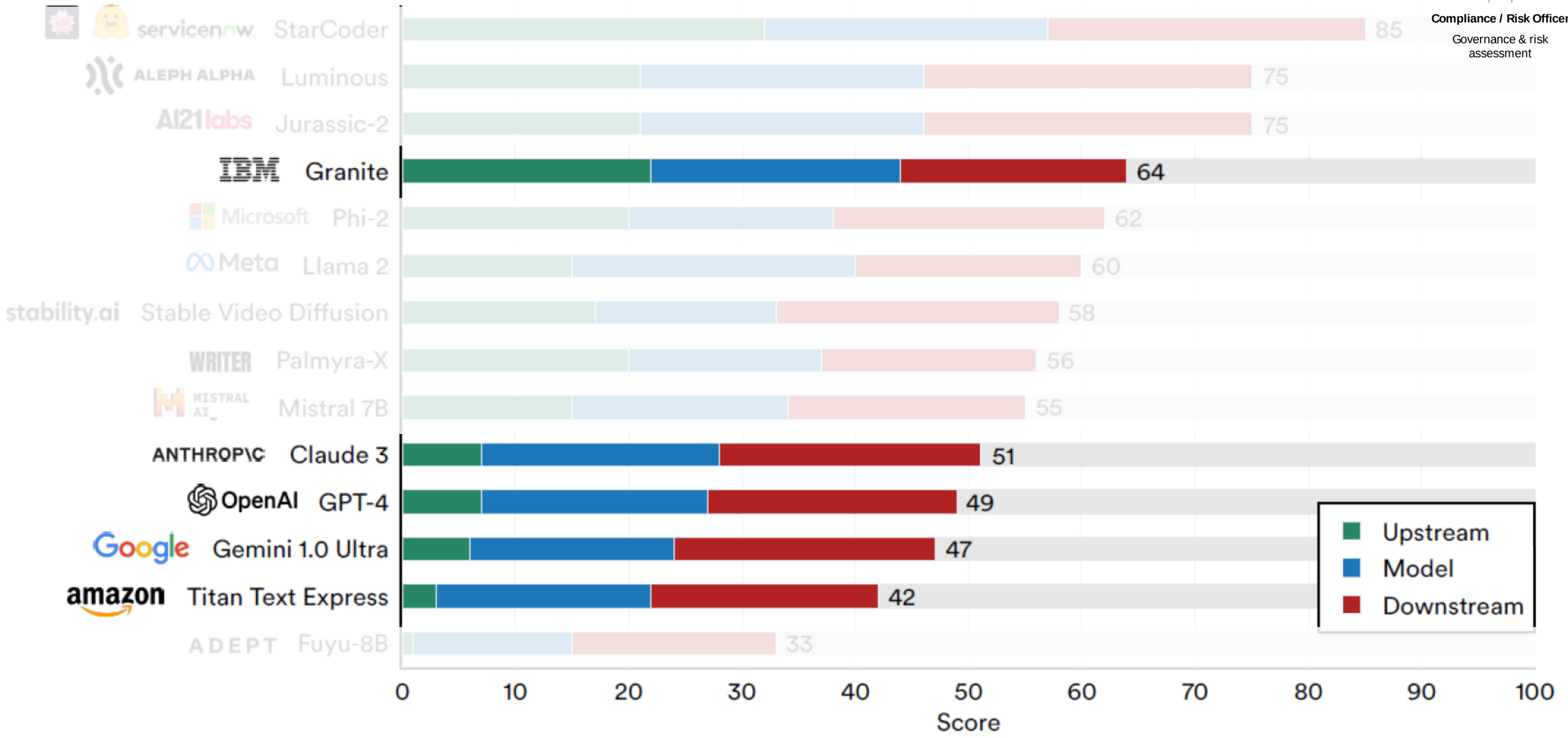
- IBM clients **benefit from** industry-leading **legal protection for Granite models**, backed by decades of experience navigating legal frameworks in tech

- 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



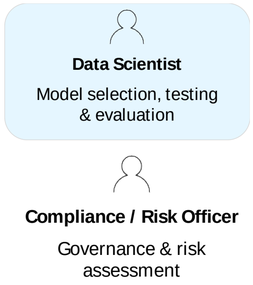
Data Scientist

Model selection, testing & evaluation

Compliance / Risk Officer

Governance & risk assessment

AI development: Foundation models
Granite model benefits



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[1]

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

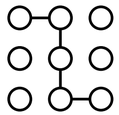
Wide applicability

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

AI development: Foundation models
Granite everywhere

Data Scientist

Model selection, testing & evaluation



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

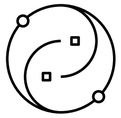
Open-source, model-agnostic toolkit, enables individuals to contribute knowledge and skills to foundation models, facilitating a new “cheaper” approach to alignment tuning that leverages both human and model evaluations for quality assurance

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

Red Hat Enterprise Linux AI (RHEL AI)

Bring togethers 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

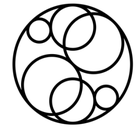
Model IP indemnification for use of Granite open sourced-licensed models provided by Red Hat accessible in RHEL AI



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

Data Scientist
Model selection, testing & evaluation

Compliance / Risk Officer
Governance & risk assessment

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
- 3rd 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[1]

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

Data Scientist

Model selection, testing & evaluation

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:

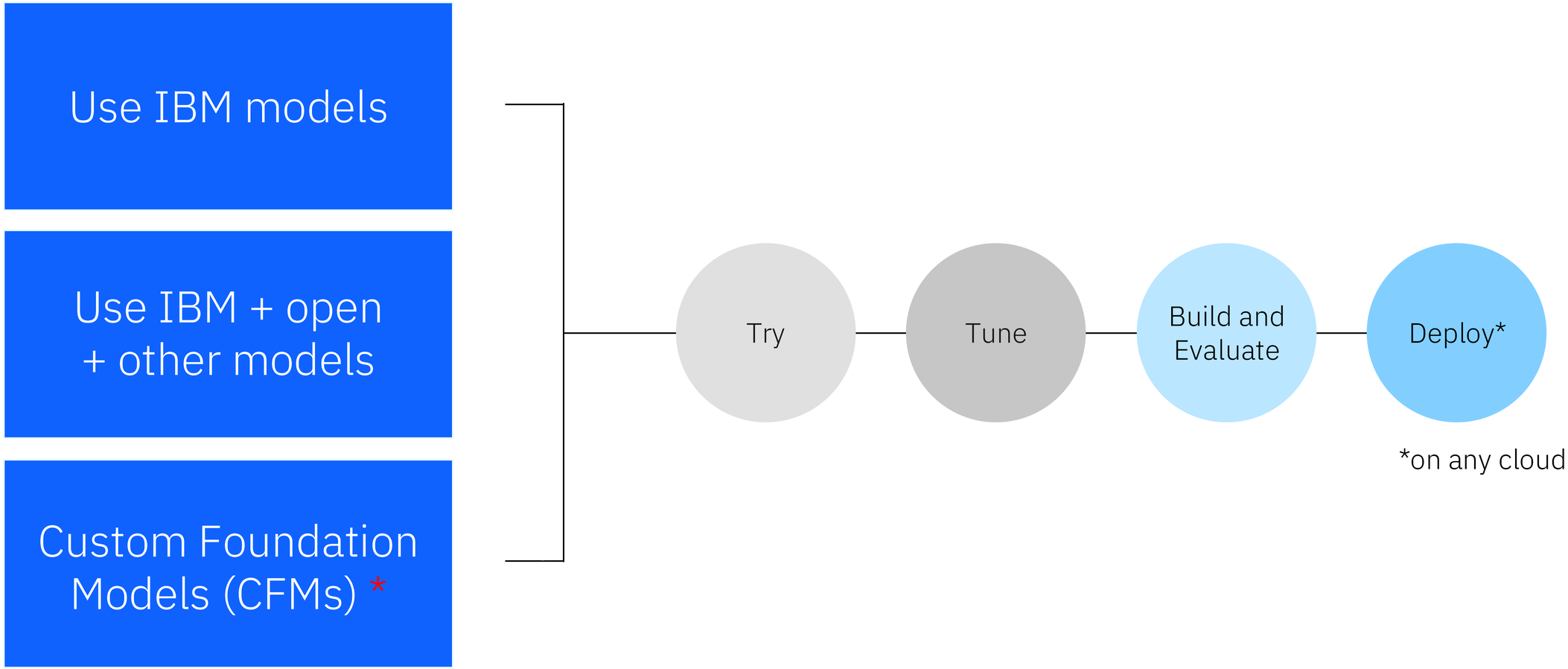
- Works well out-of-the-box on various NLP tasks
- Energy intensive to train and run
- More expensive
- Less efficient

AI development: Foundation models

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

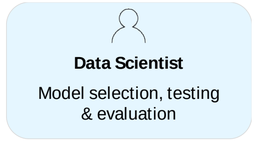
Data Scientist

Model selection, testing & evaluation



* 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

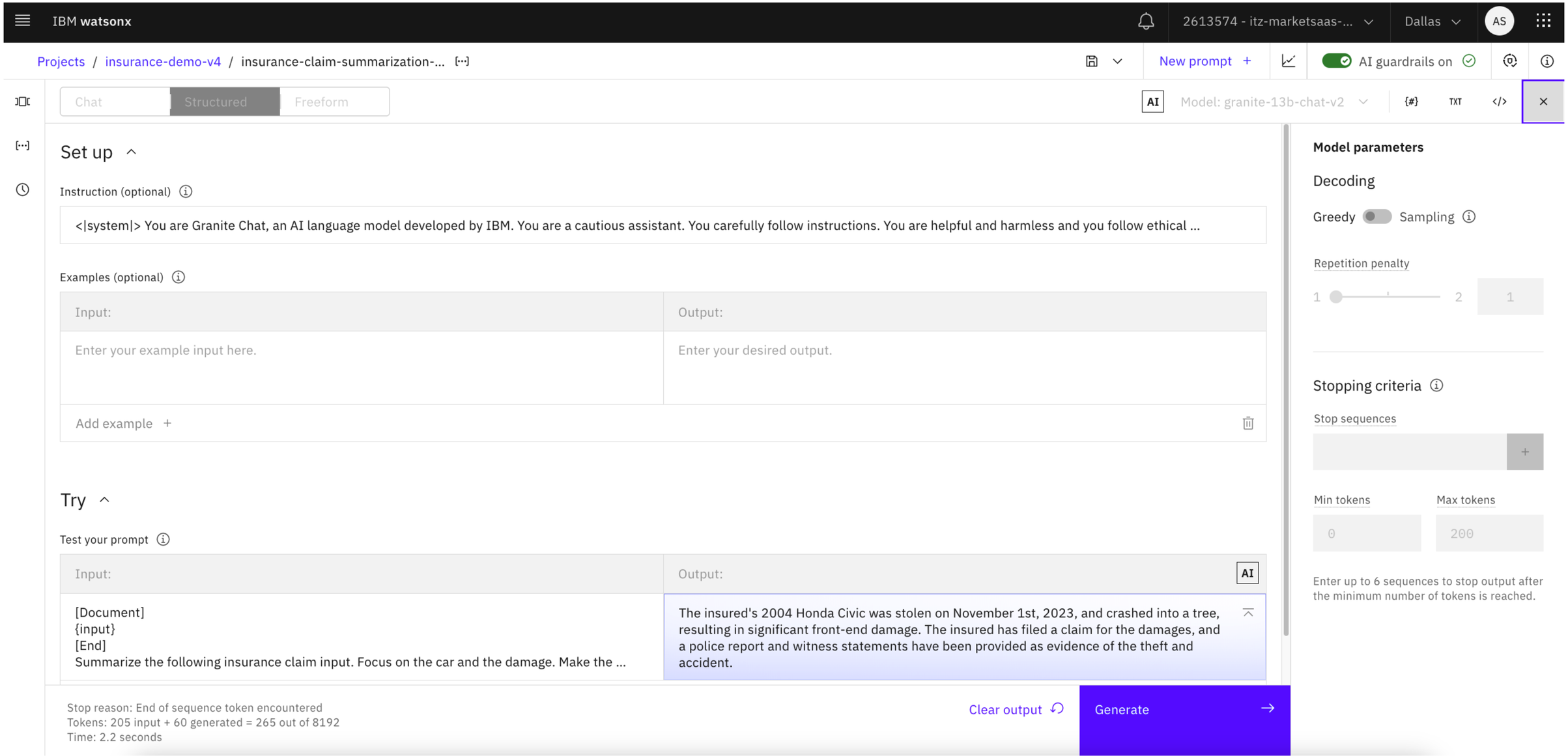
- 1000+ more model choices
- Access to more language support
- Bring your own fine-tuned model catered to an industry or business domain

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



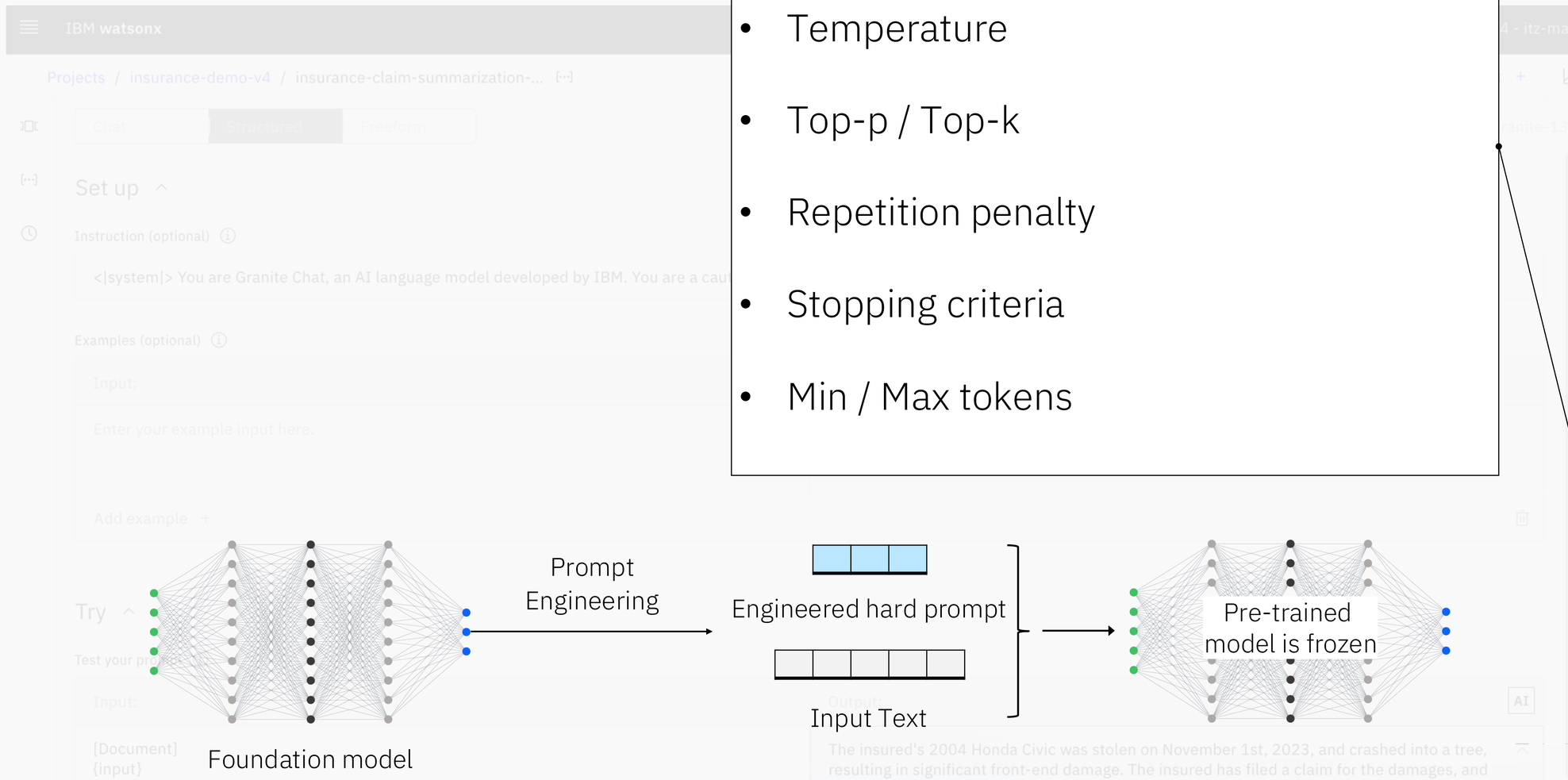
AI development

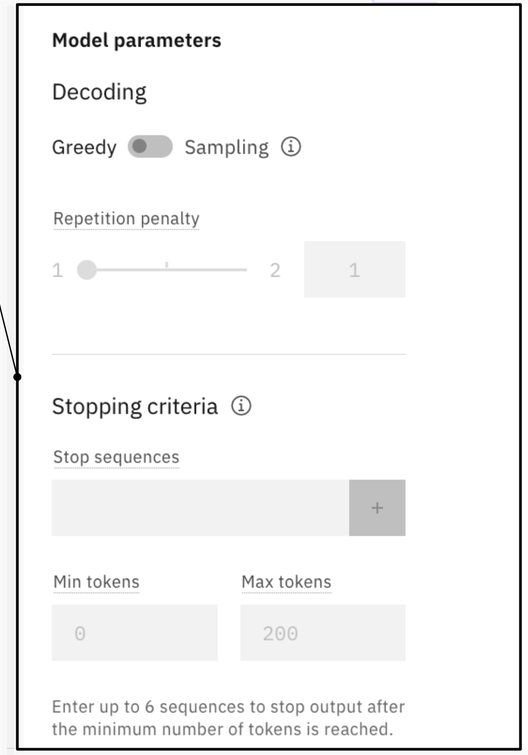
Data Scientist experiments with diff … s to evaluate if the desired information c…

Data Scientist
Model selection, testing & evaluation

Model parameters

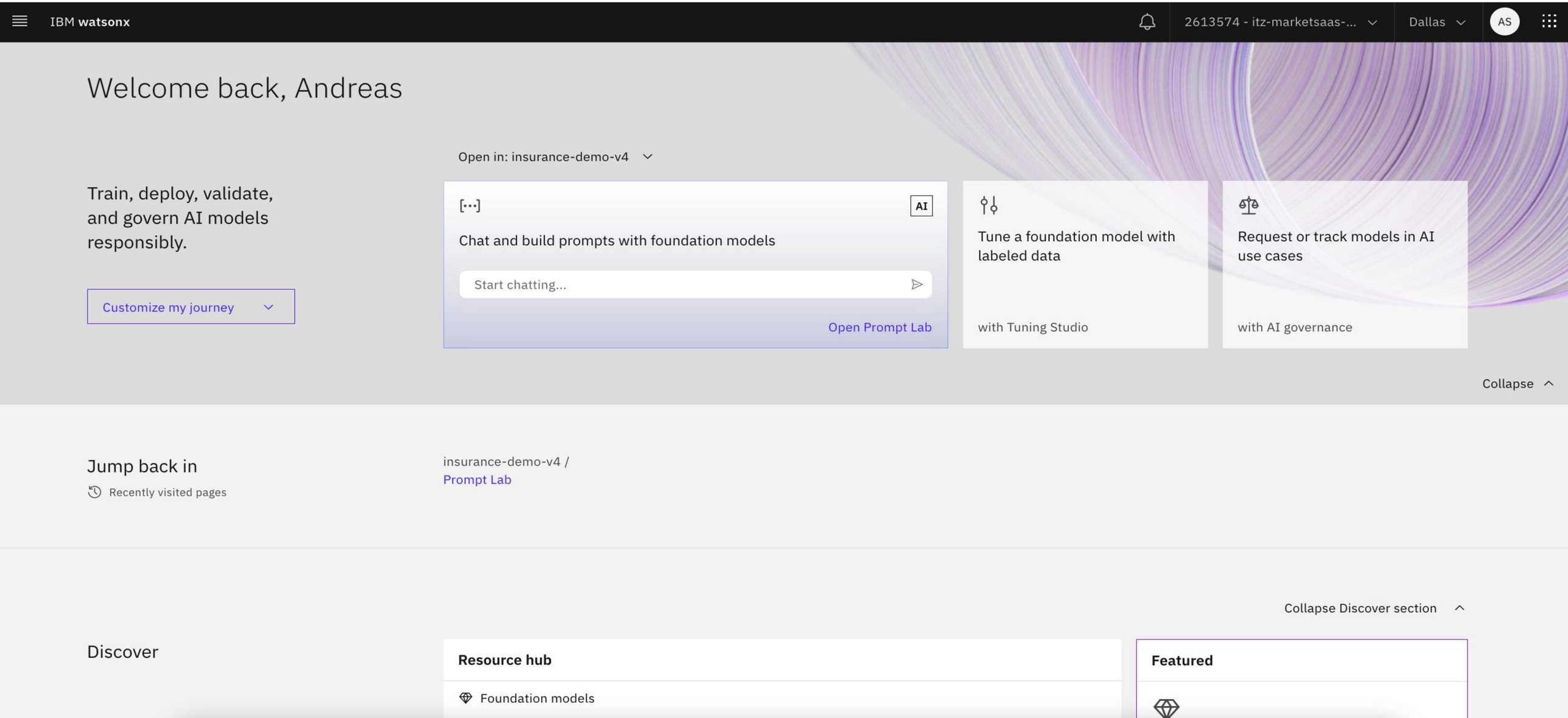
- Greedy vs. Sampling
- Temperature
- Top-p / Top-k
- Repetition penalty
- Stopping criteria
- Min / Max tokens





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



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

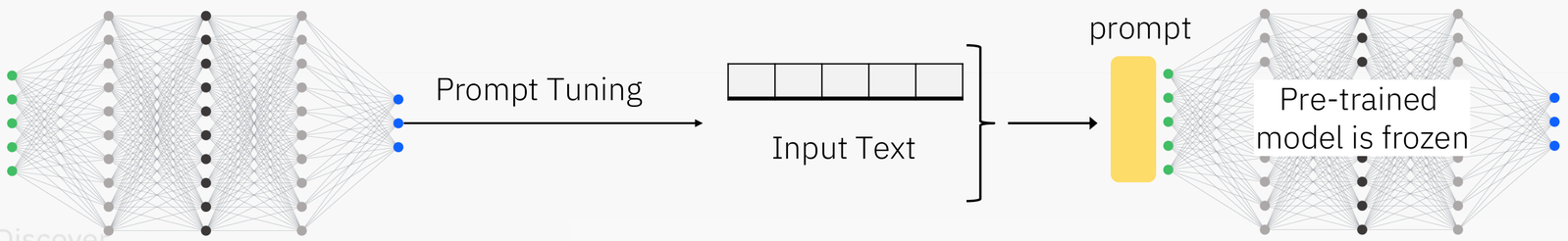
Prompt Tuning parameters

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

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

Taxonomy-driven Small Language Model instruction fine-tuning based on synthetic data generation: Instruct Lab [1]

Data Scientist
Model selection, testing & evaluation

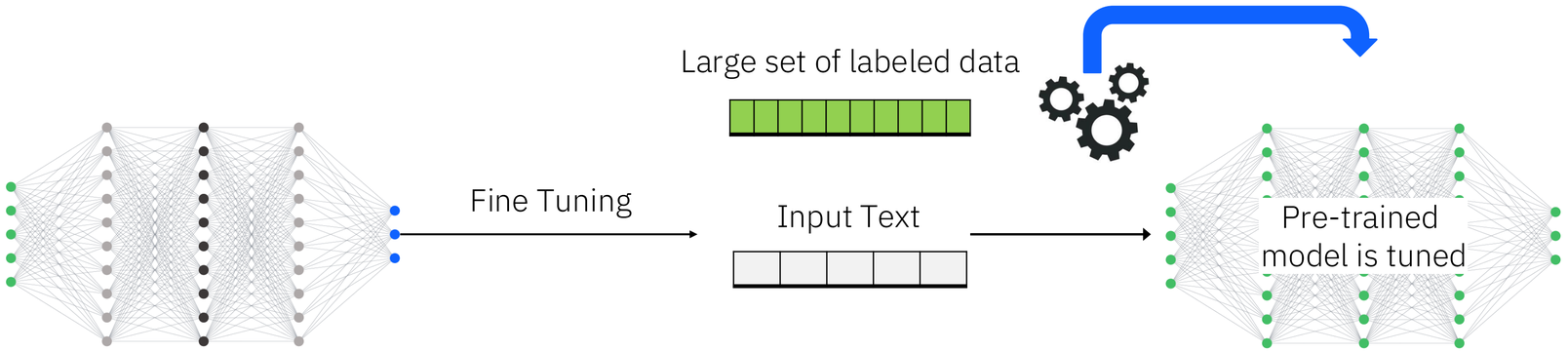
Business User
Use case definition & request

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** [2] 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*

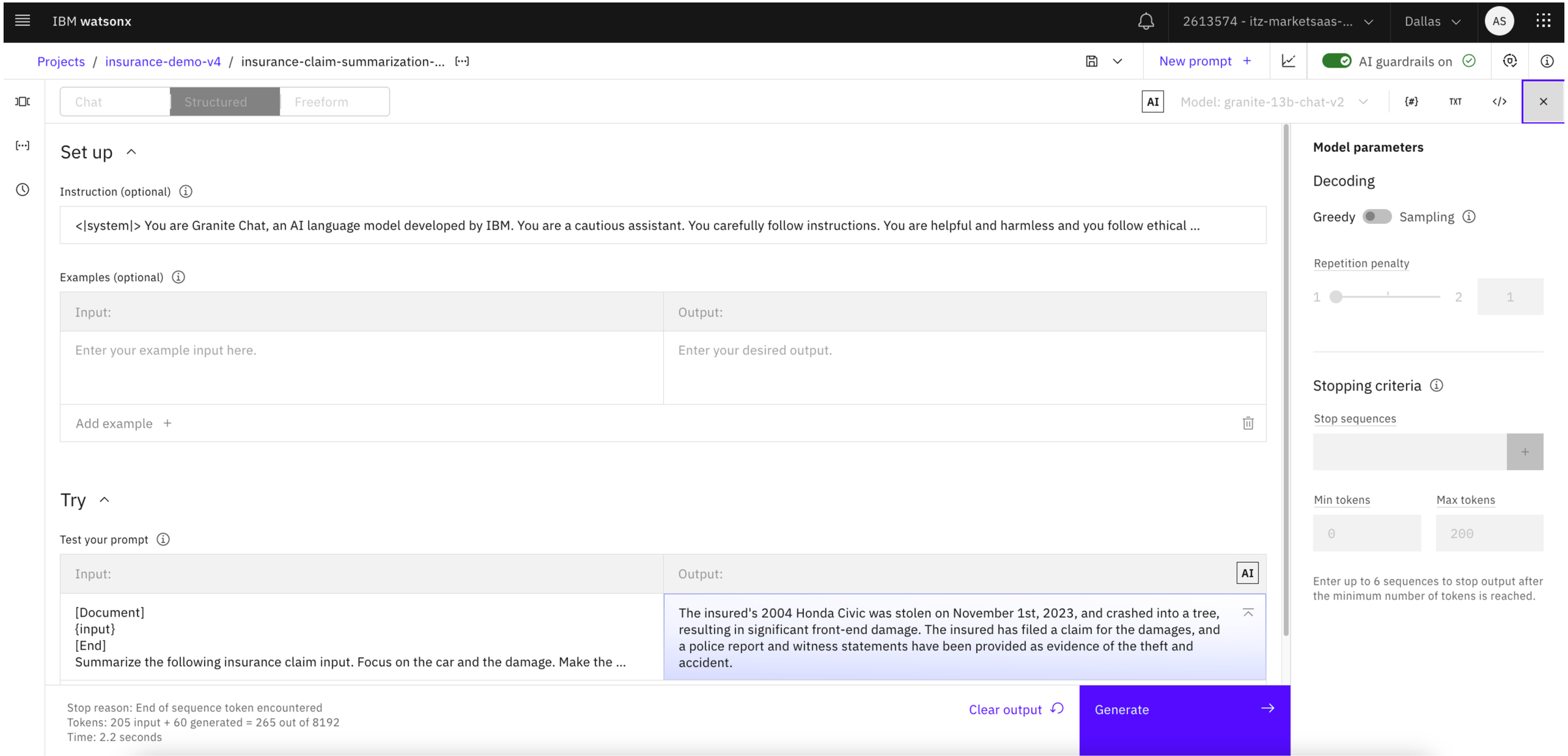
→ *Adding content to the taxonomy is a business user task*



[1] LAB: Large-Scale Alignment for ChatBots
[2] 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

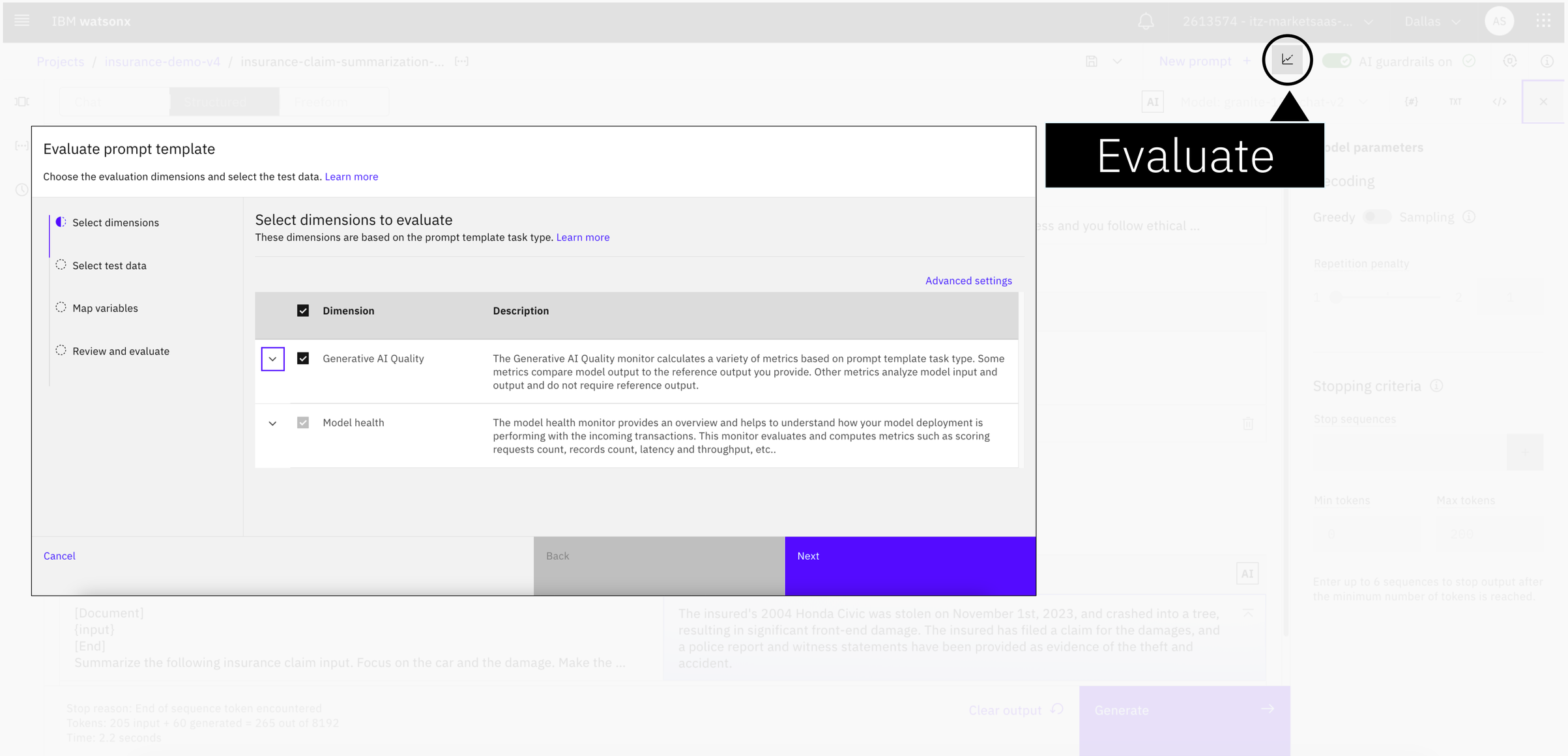
Data Scientist
Model selection, testing & evaluation



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



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
watsox.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

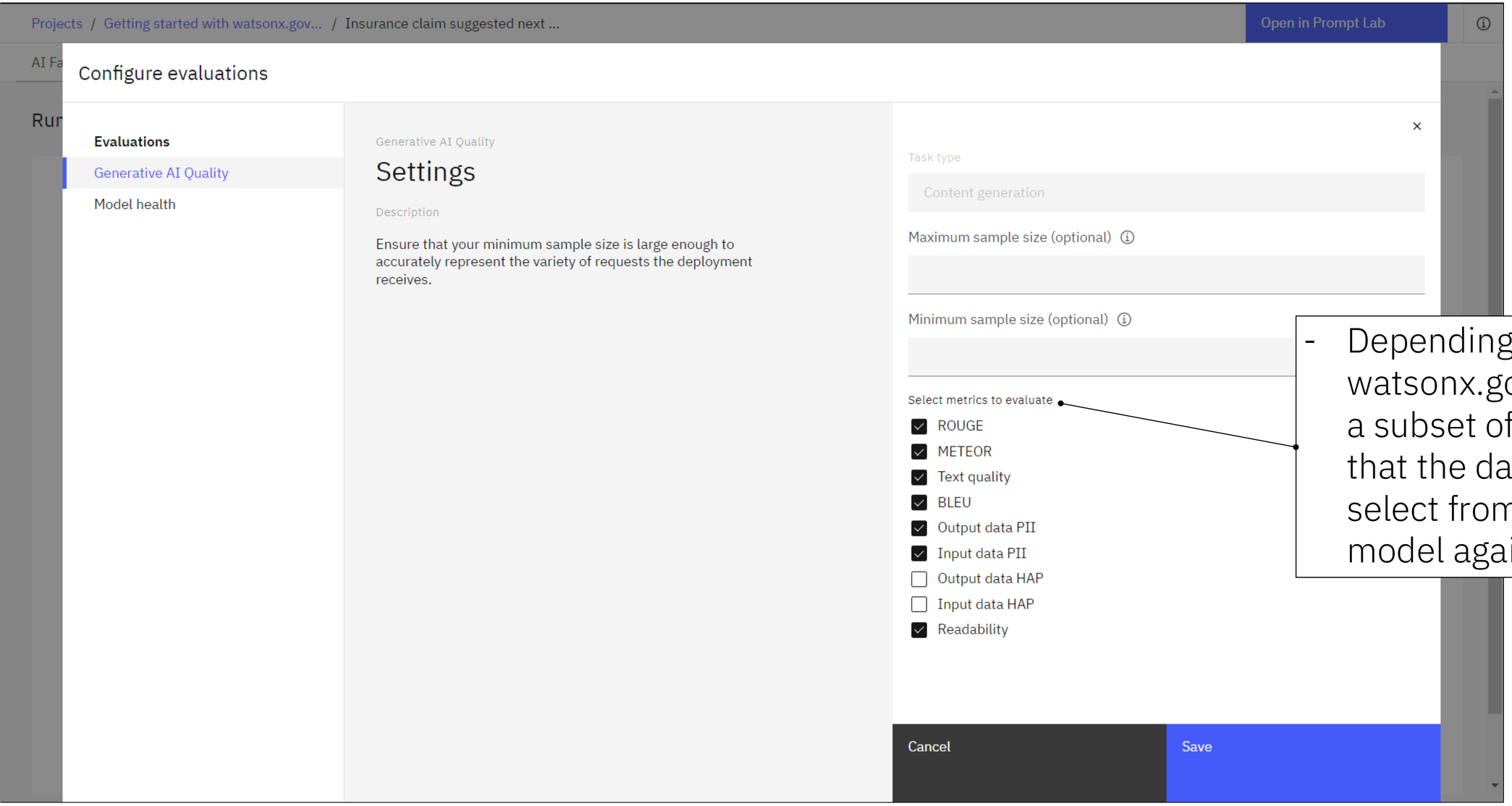
Model Health Metrics

(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

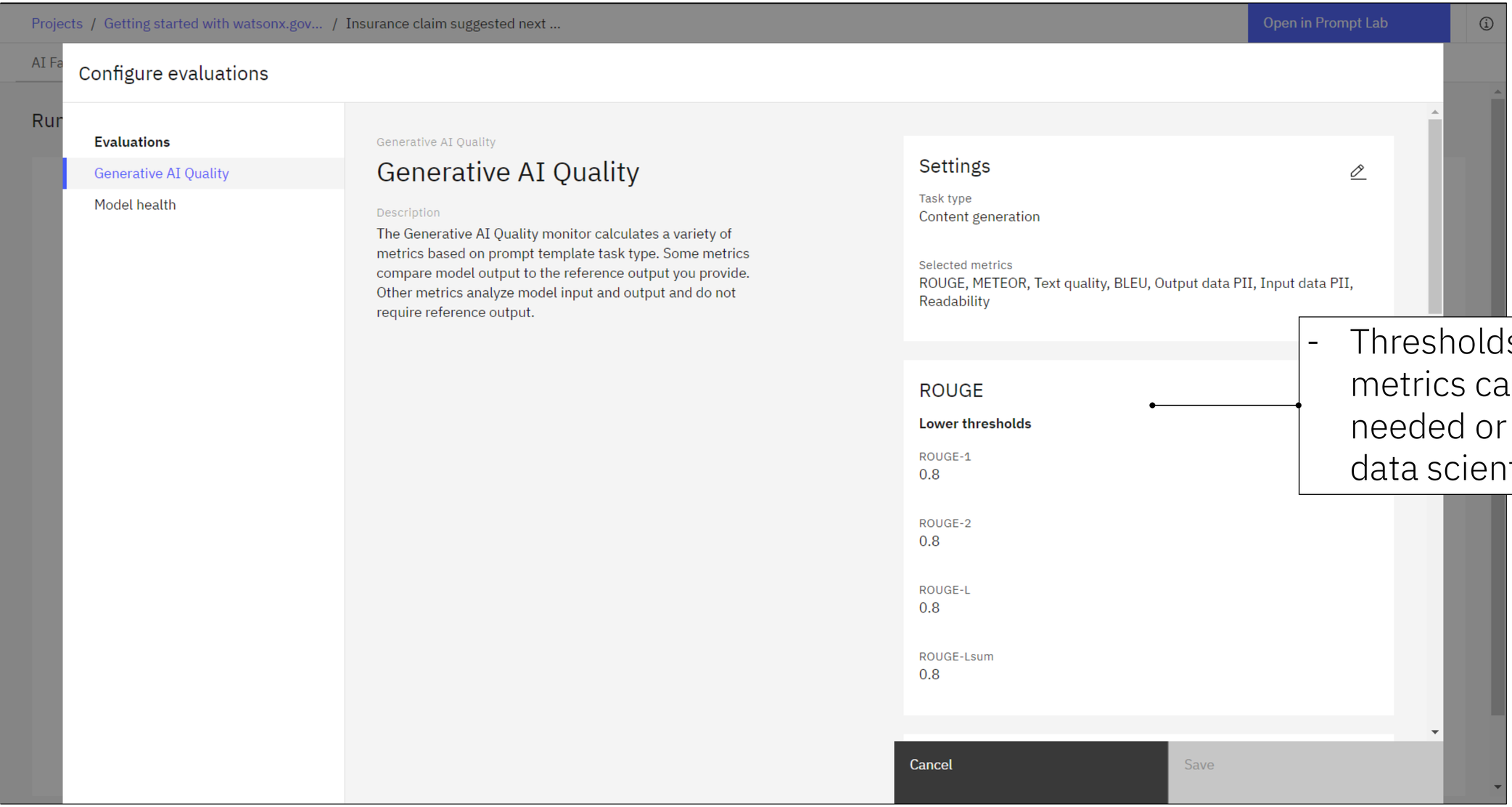
Data Scientist
Model selection, testing & evaluation



- 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.

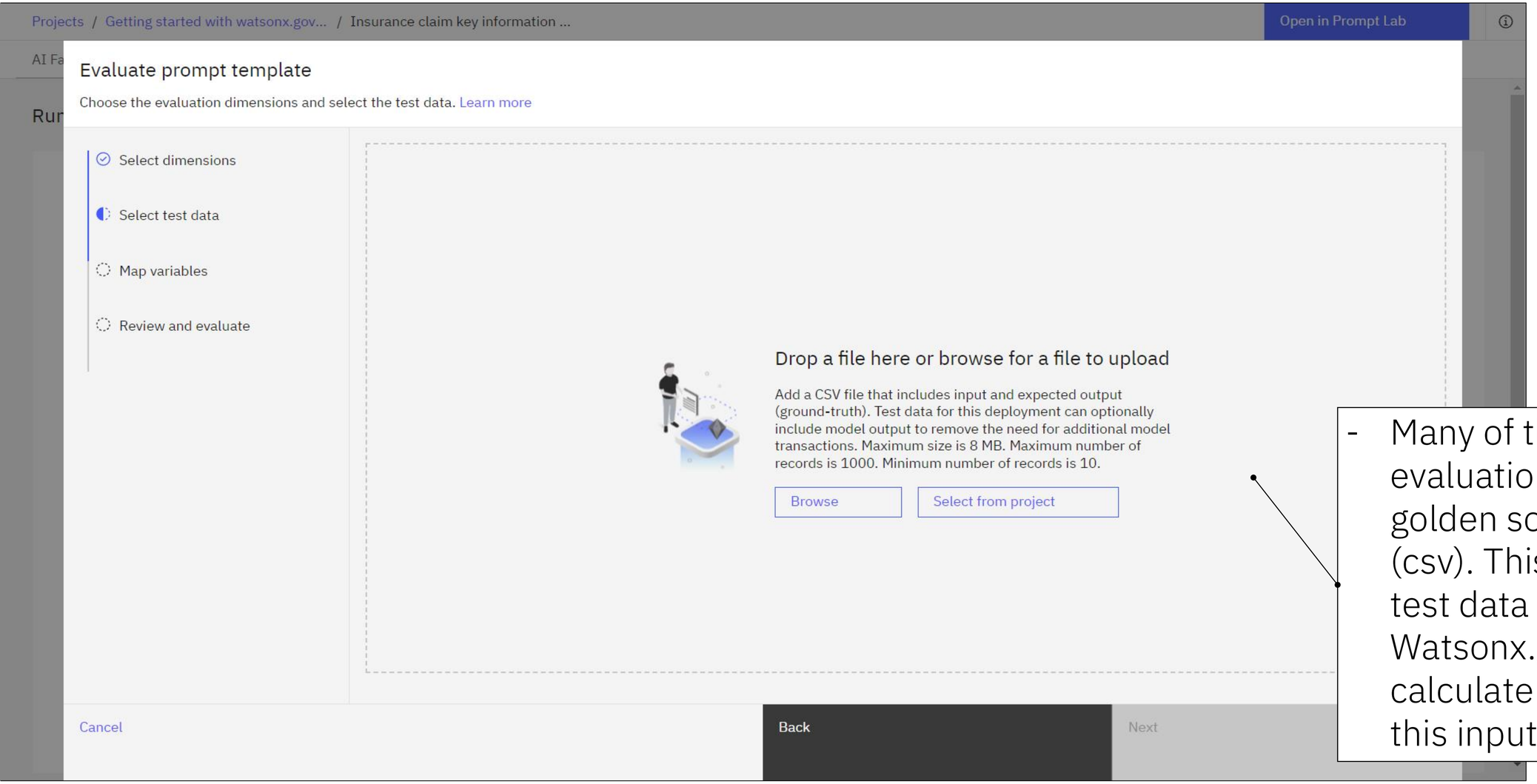
AI development & AI governance
Data Scientist evaluates prompt template and creates an AI Factsheet

Data Scientist
Model selection, testing & evaluation



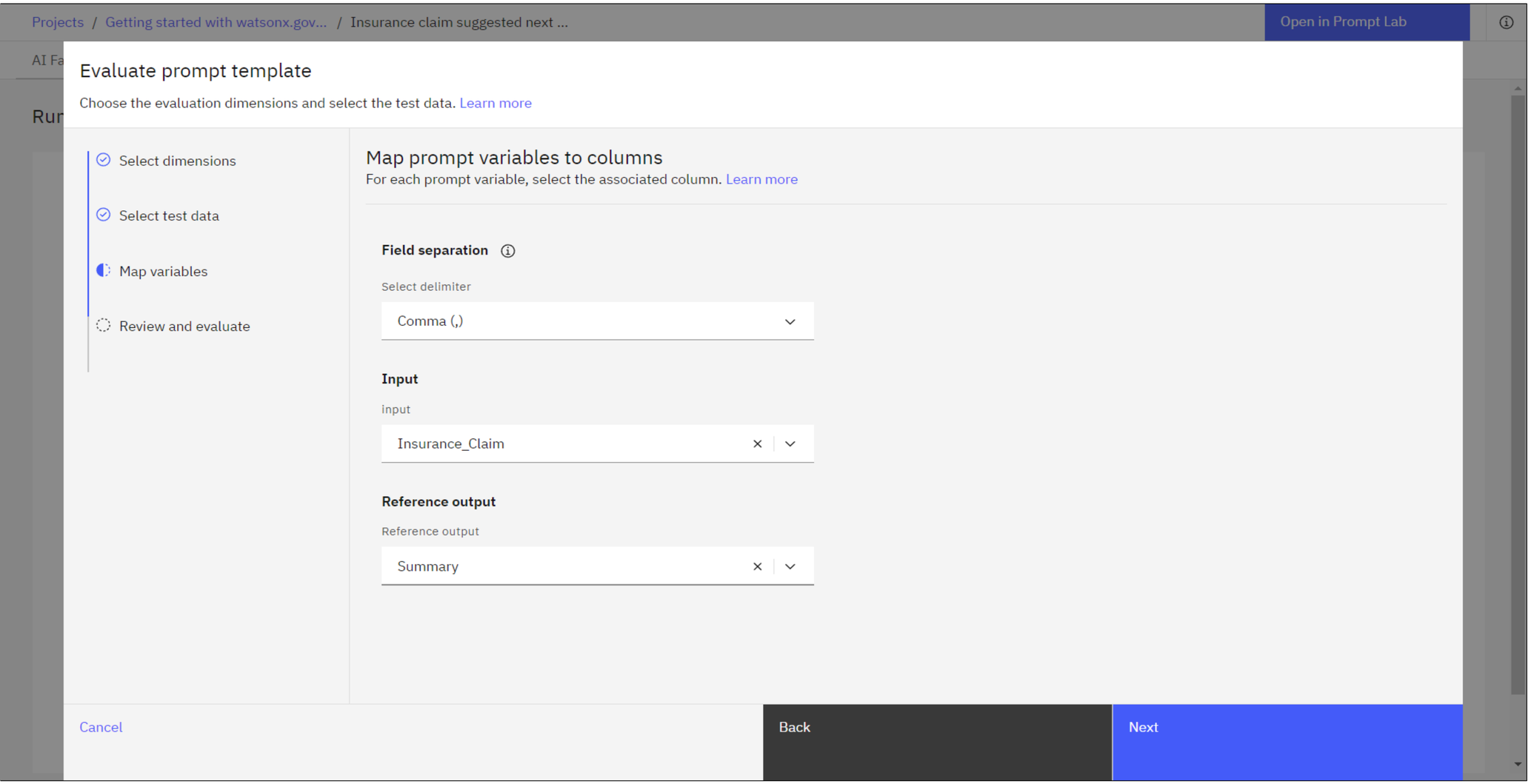
- Thresholds for the evaluated metrics can be defined as needed or approved by the data scientist.

AI development & AI governance
Data Scientist evaluates prompt template and creates an AI Factsheet

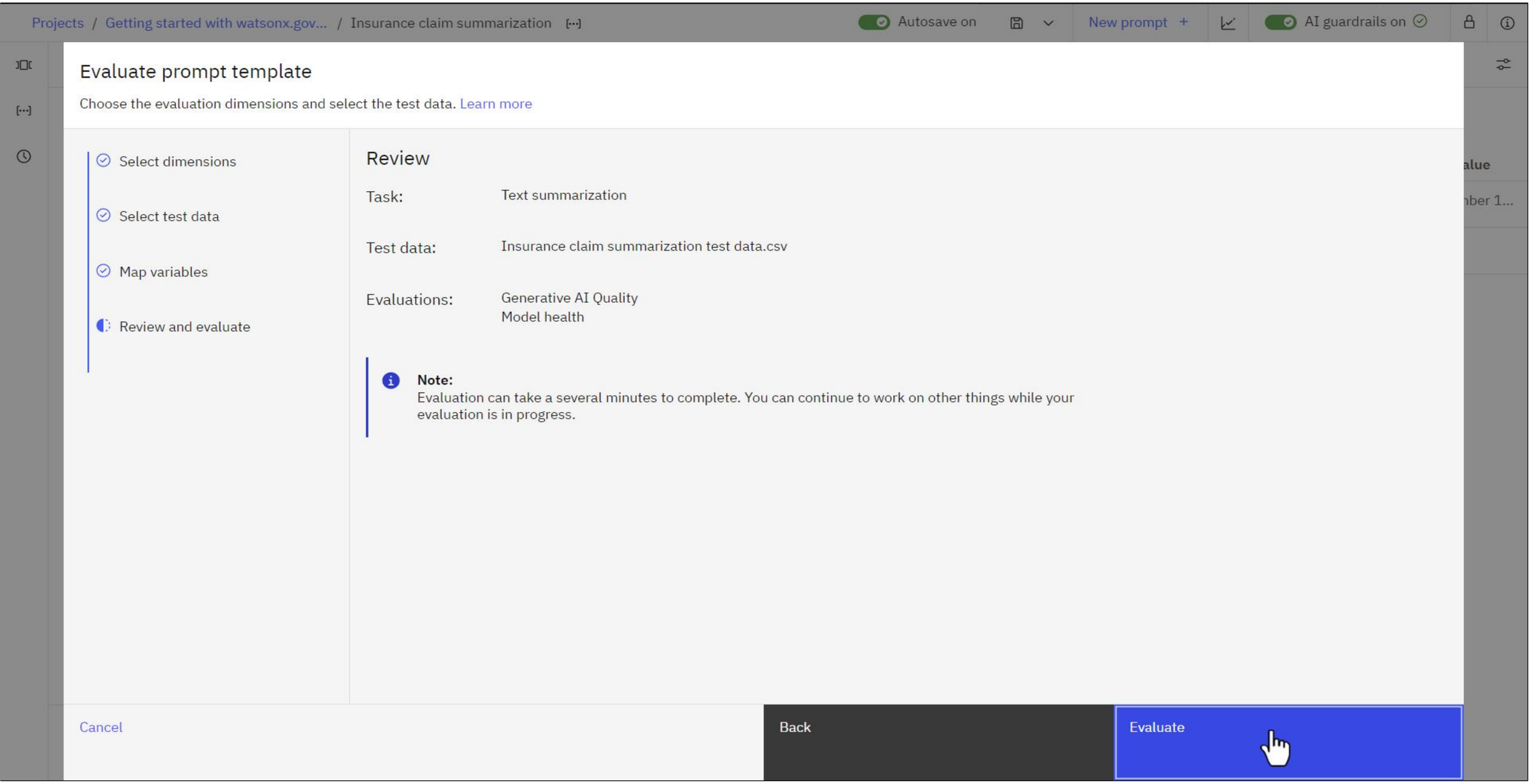


- Many of the provided evaluation metrics require a golden source training set (csv). This set needs to include test data of input and output. Watsonx.governance will calculate the metrics based on this input.

AI development & AI governance
Data Scientist evaluates prompt template and creates an AI Factsheet



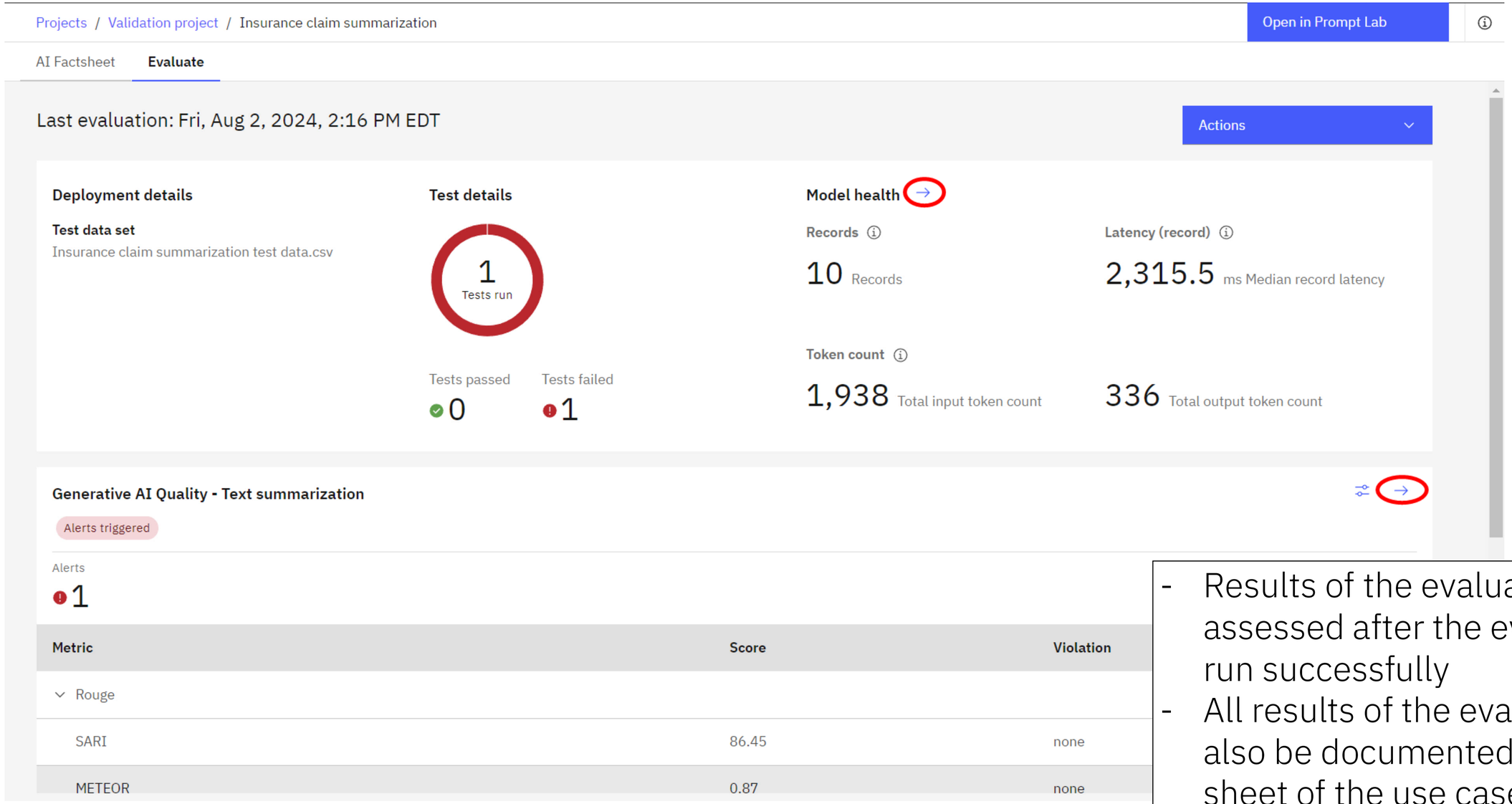
AI development & AI governance
Data Scientist evaluates prompt template and creates an AI Factsheet



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



- 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

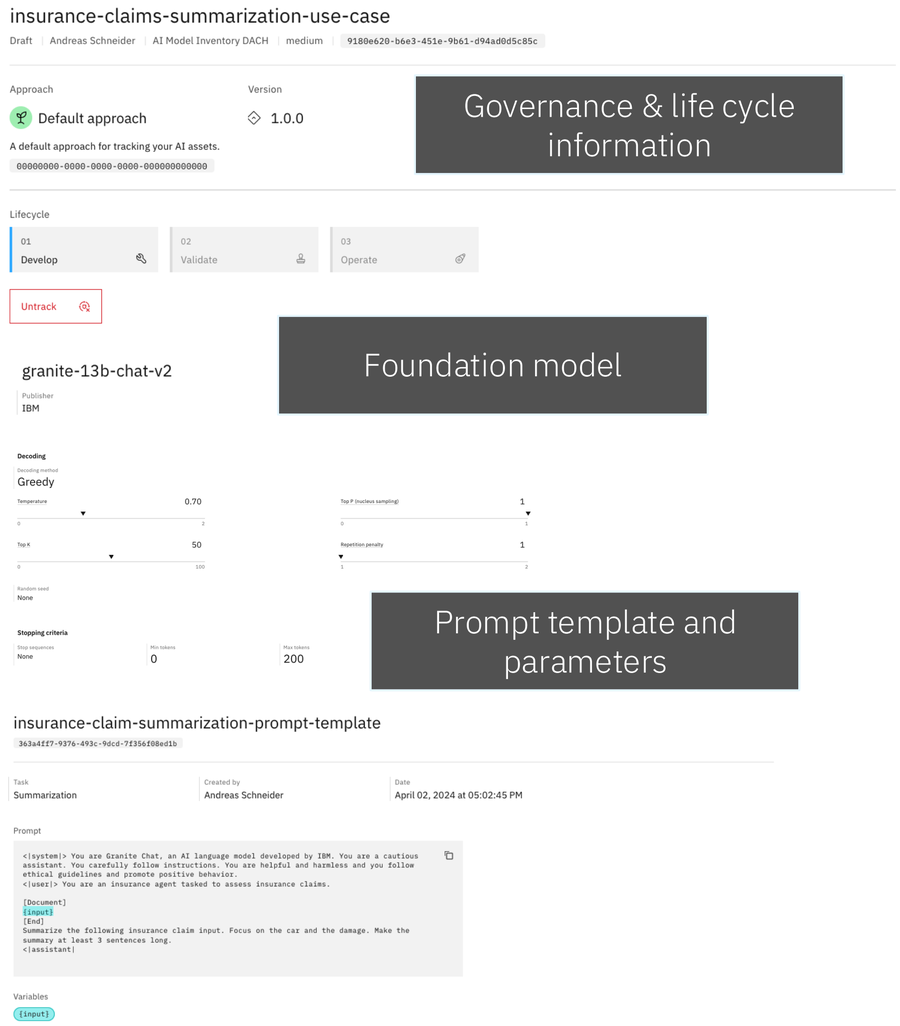
Data Scientist
Model selection, testing & evaluation

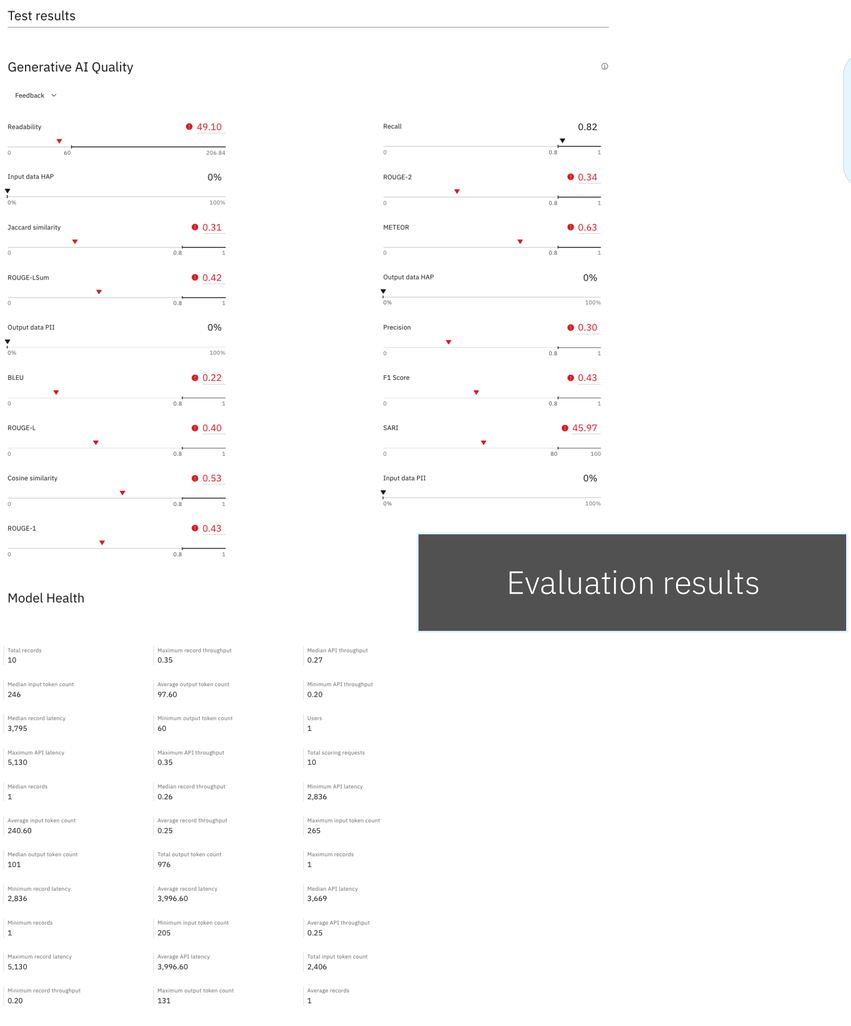
Compliance / Risk Officer
Governance & risk assessment

Business User
Use case definition & request

AI Factsheet

- Governance
- Foundation model
- Prompt template
- Prompt parameters
- Evaluation
  - Develop
    - insurance-demo-v4
      - Test results
- Attachments
  - Other attachments



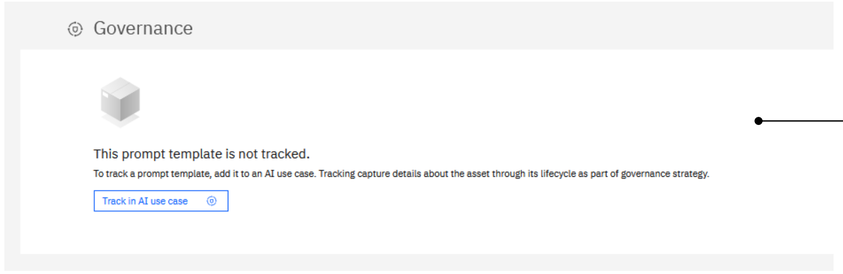


AI governance
Data Scientist evaluates prompt template and creates an AI Factsheet

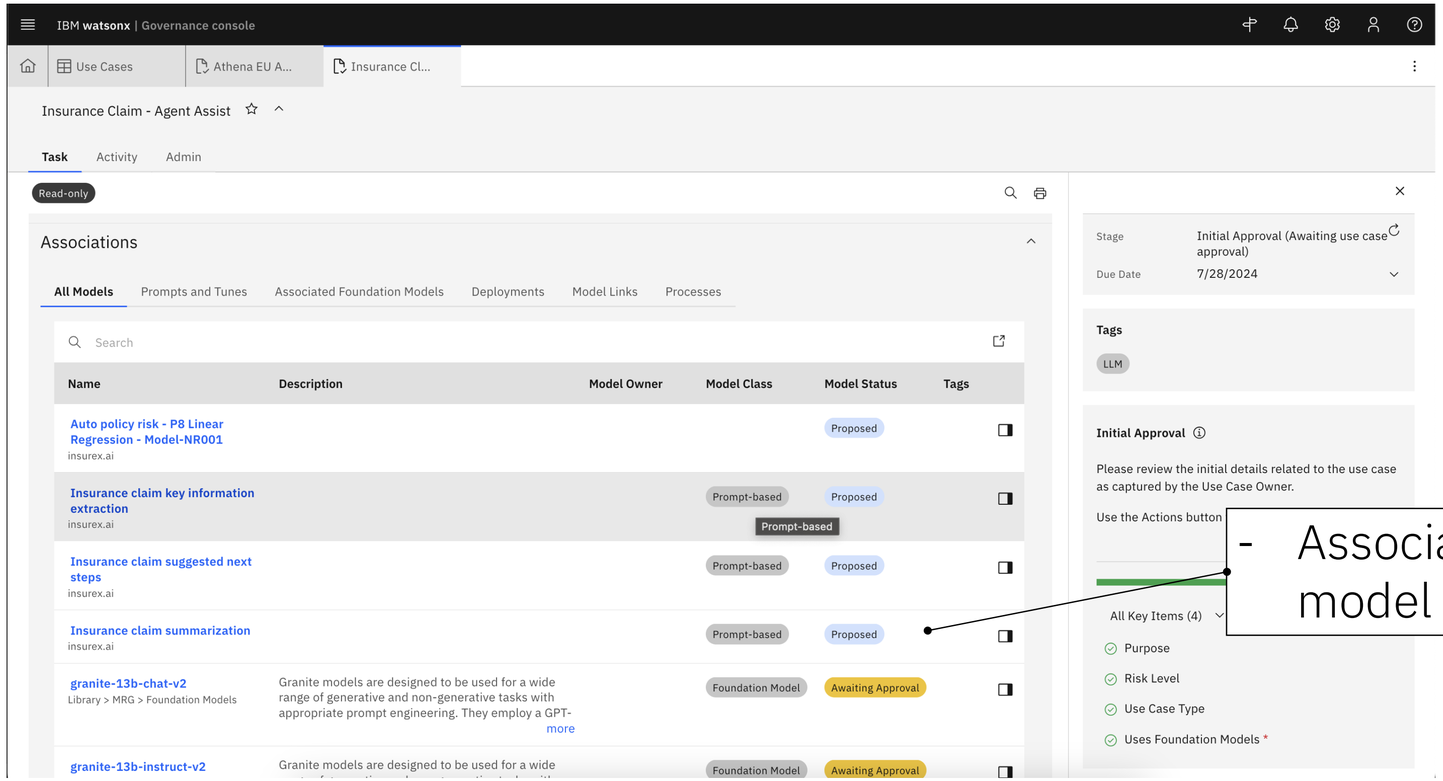
Data Scientist
Model selection, testing & evaluation

Compliance / Risk Officer
Governance & risk assessment

Business User
Use case definition & request



- 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.



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

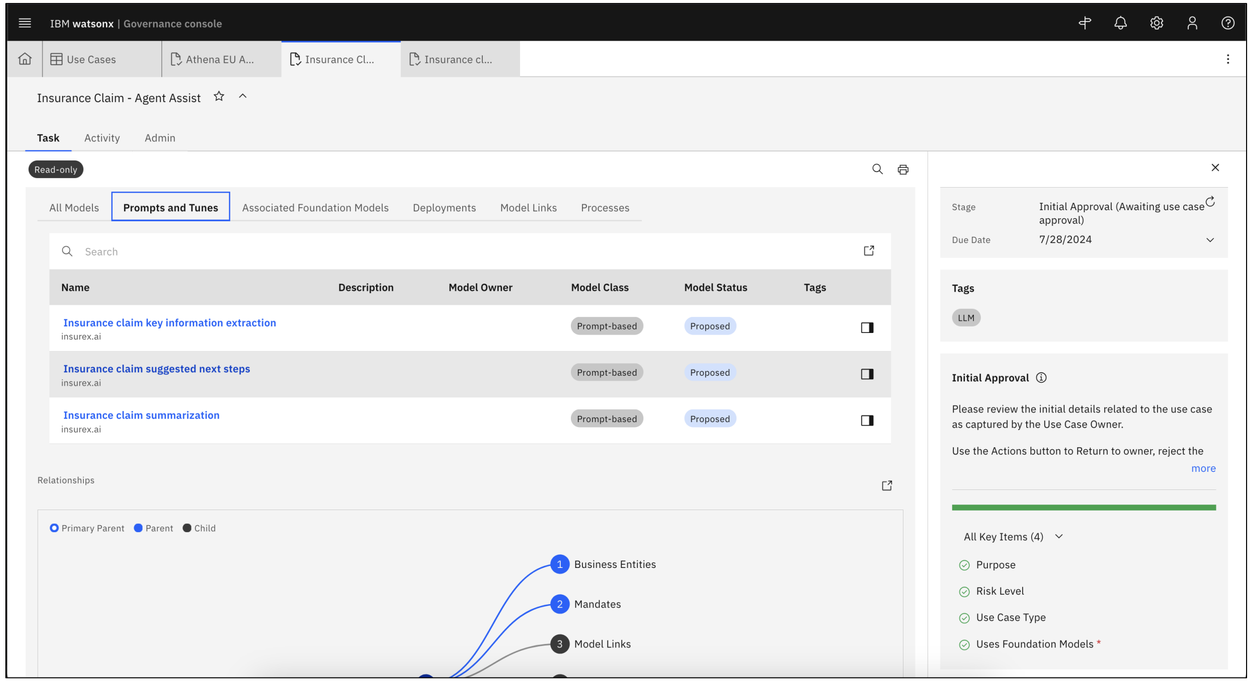
AI governance
Data Scientist evaluates prompt template and creates an AI Factsheet

Data Scientist
Model selection, testing & evaluation

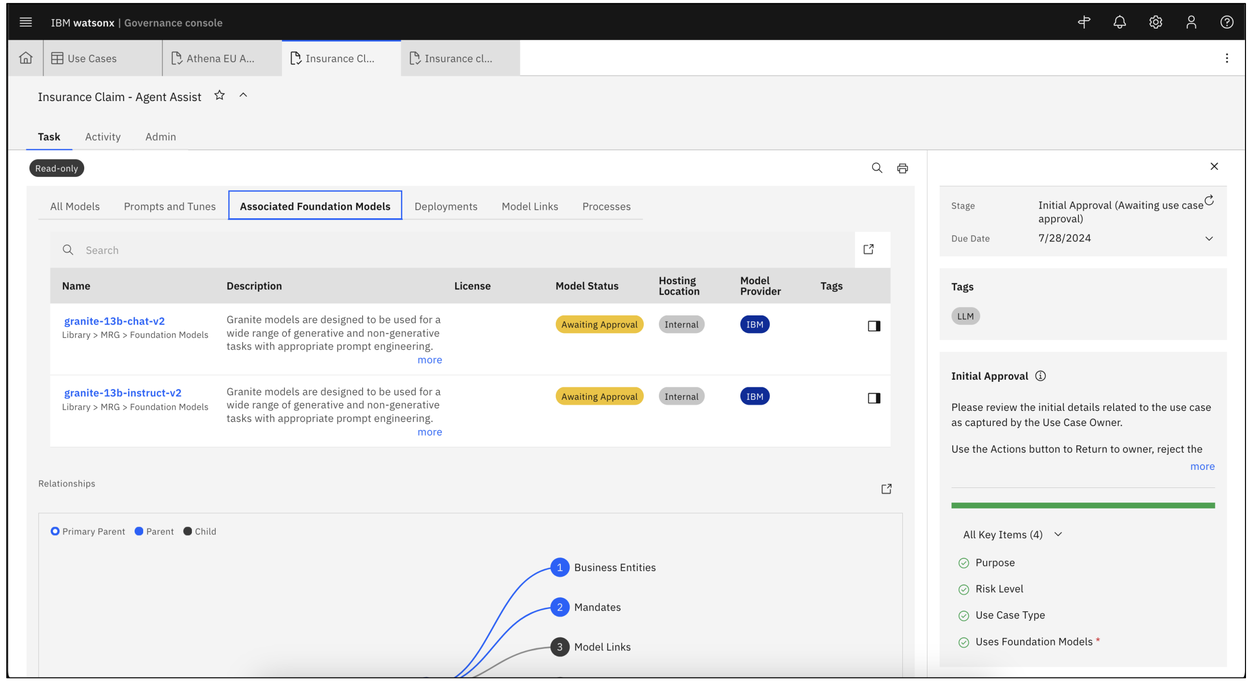
Compliance / Risk Officer
Governance & risk assessment

Business User
Use case definition & request

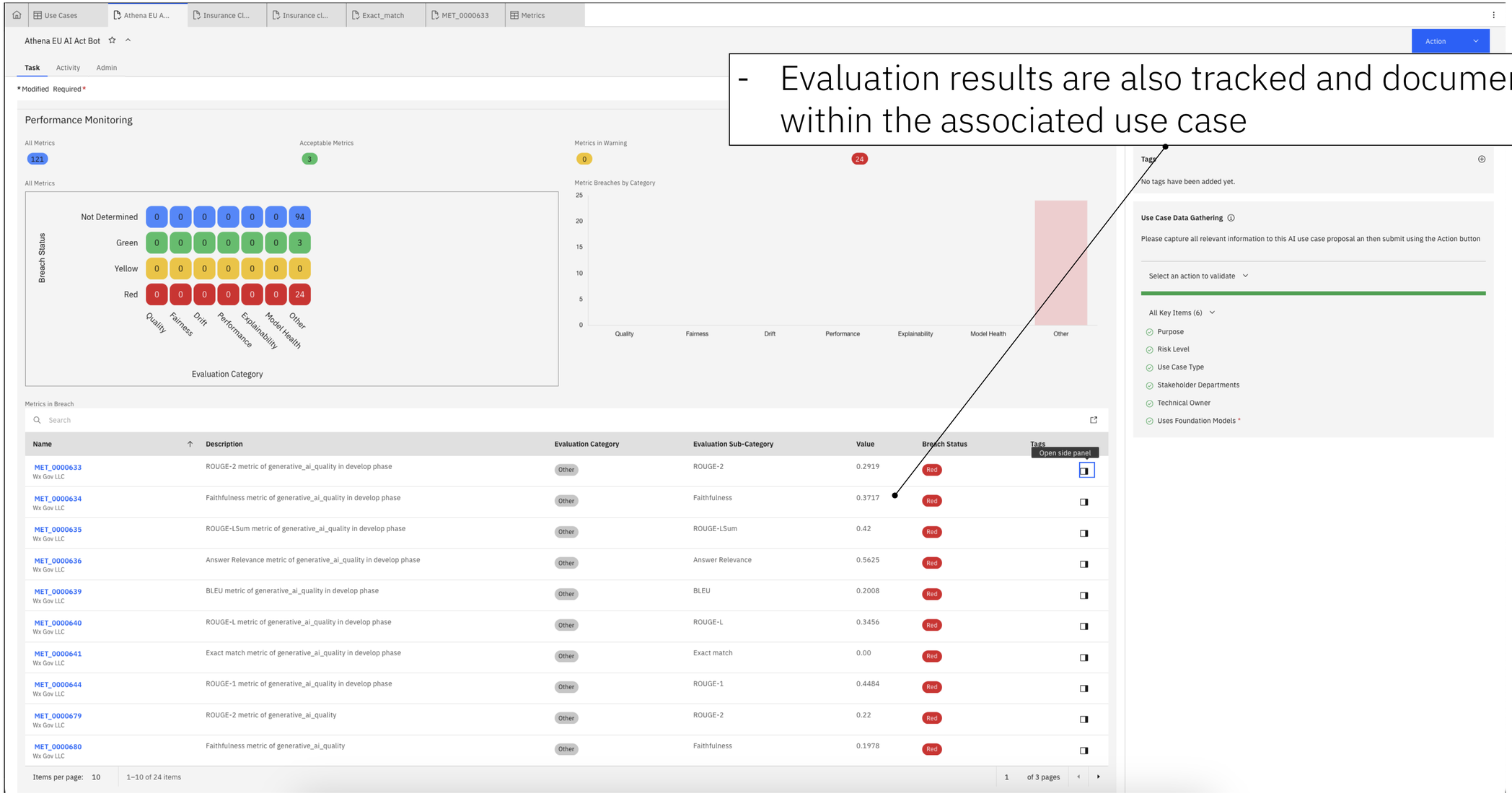
Associated summarization Prompt with the use case



Associated Foundation Model with the use case

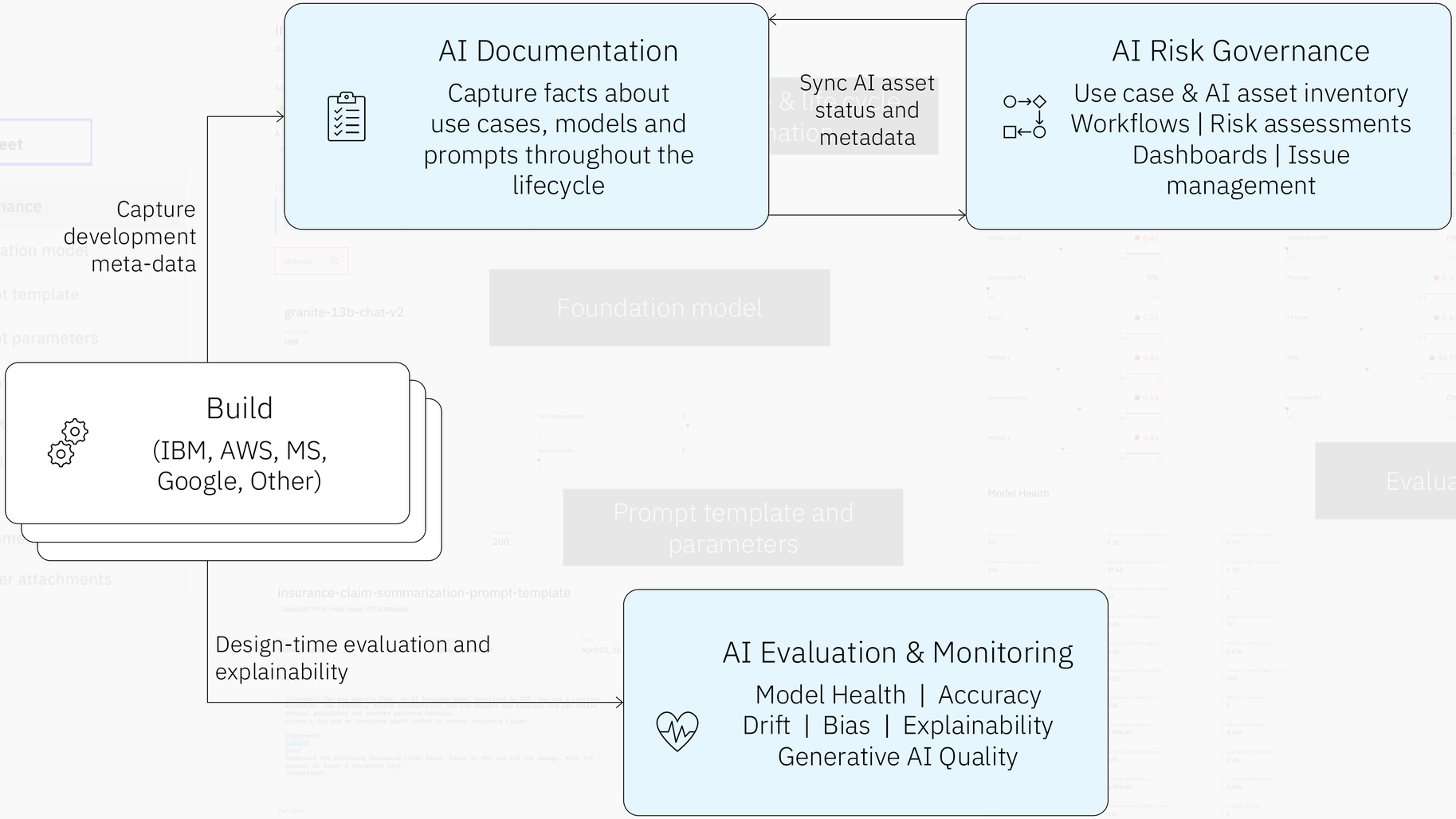


AI governance
Data Scientist evaluates prompt template and creates an AI Factsheet



- 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

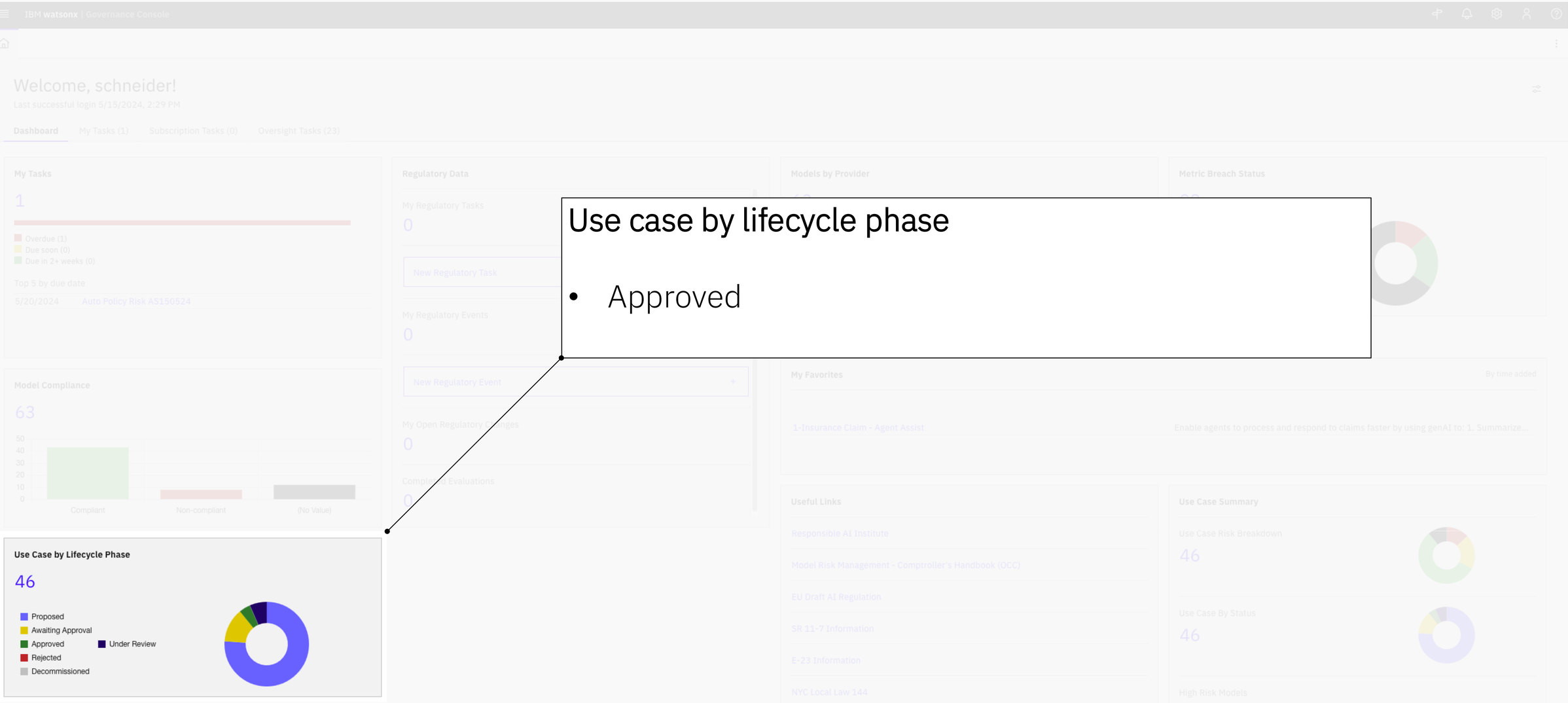
Compliance / Risk Officer
Governance & risk assessment

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

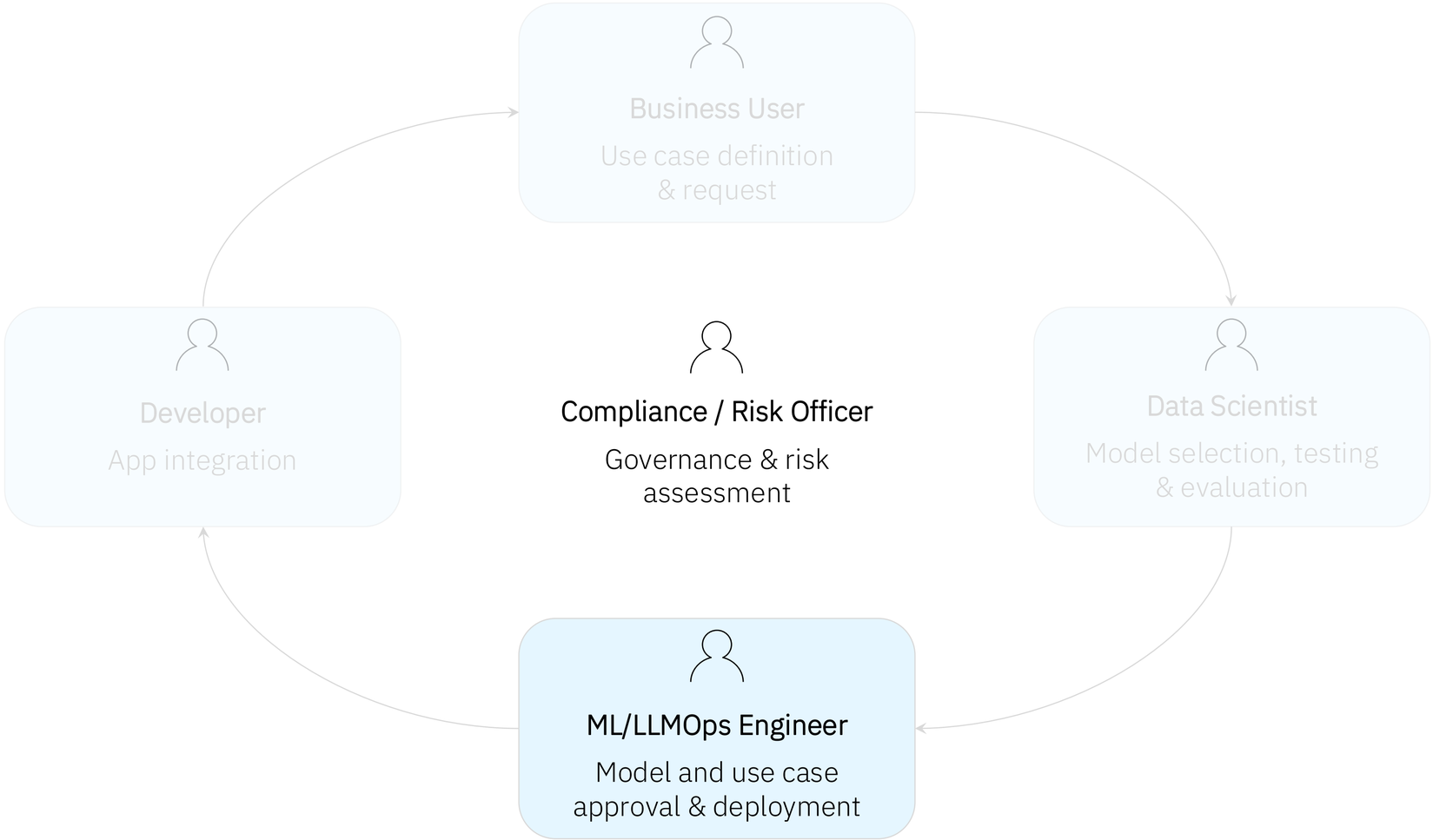
Compliance / Risk Officer
Governance & risk assessment

Use case by lifecycle phase

- Approved

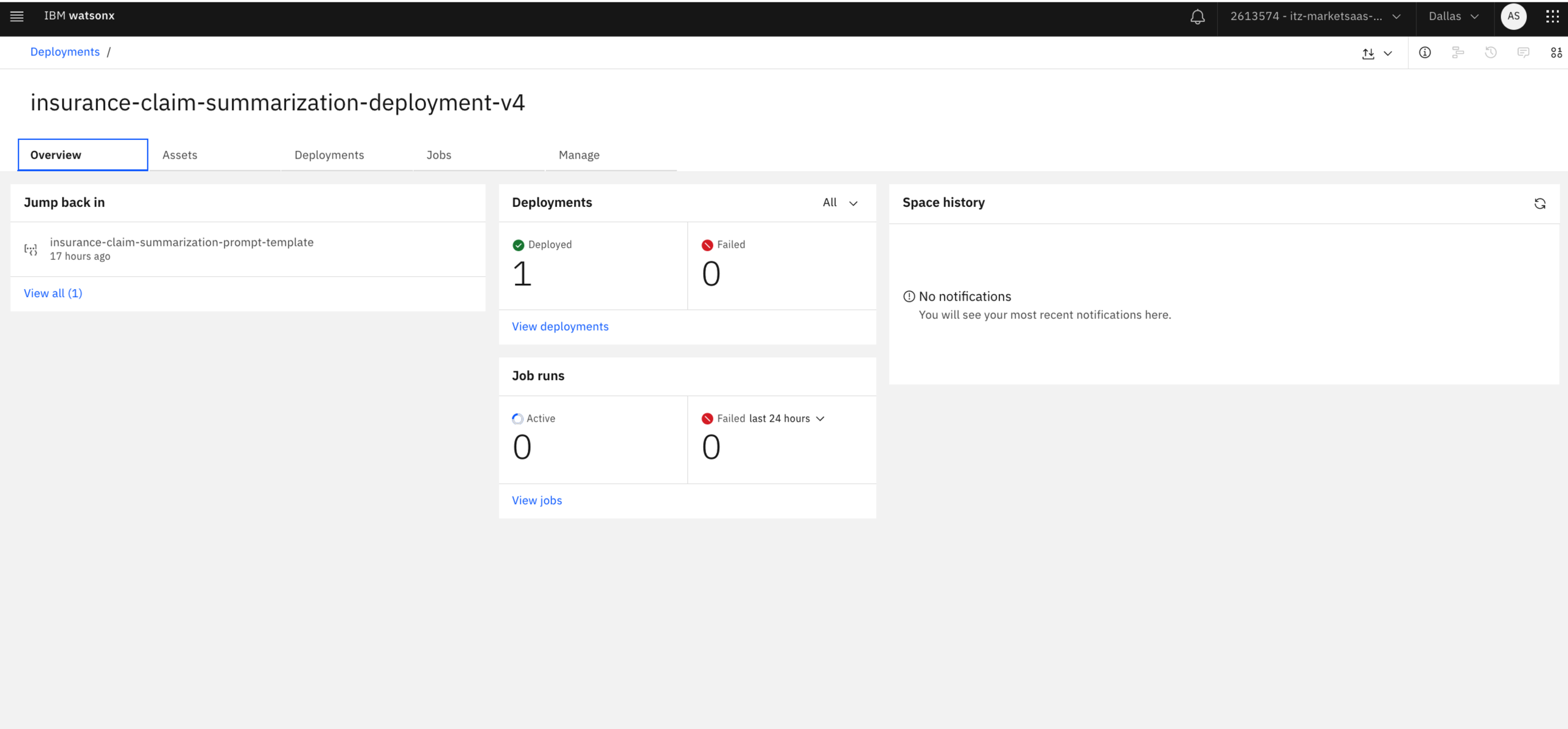


AI development & AI governance
Summarization of car insurance claims



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

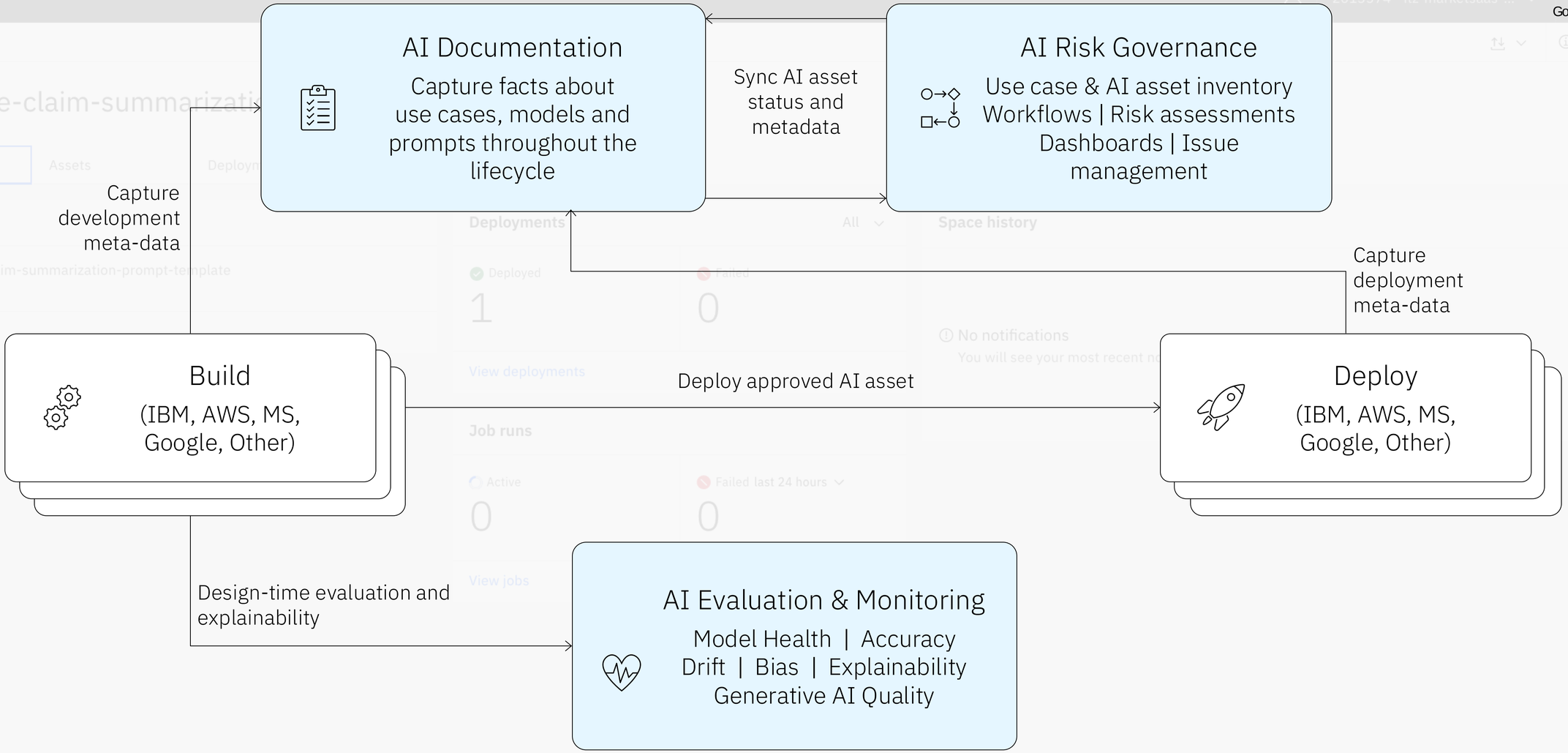
ML/LLMOps Engineer
Model and use case approval & deployment



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

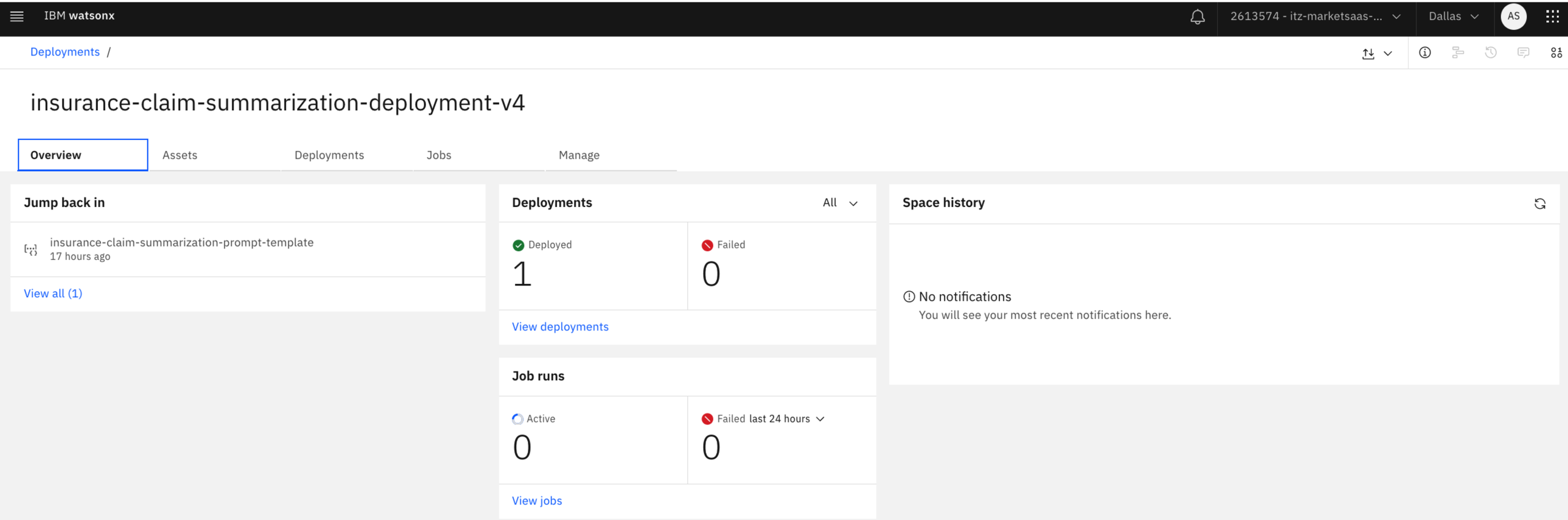
ML/LLMOps Engineer
Model and use case approval & deployment

Compliance / Risk Officer
Governance & risk assessment

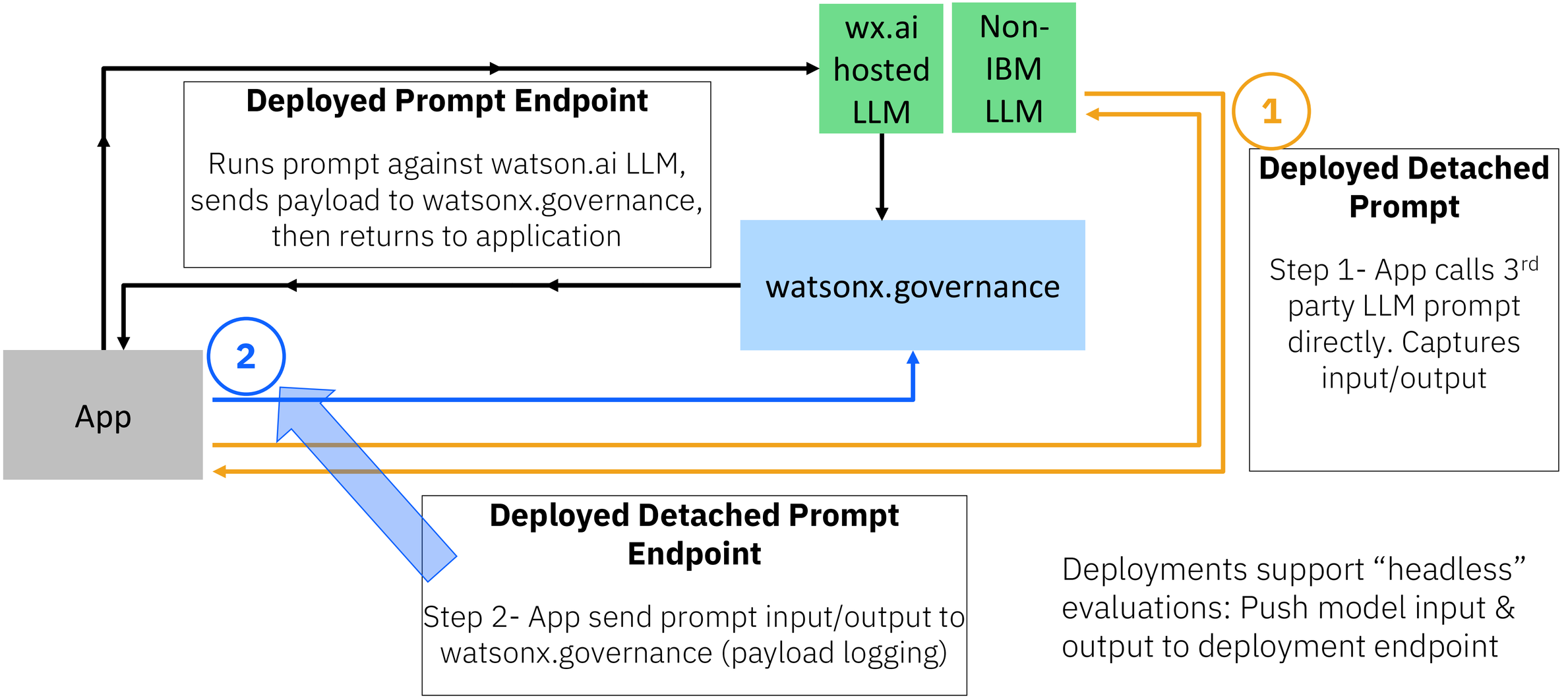


AI development
ML/LLMOps Engineer monitors deployed prompt template

ML/LLMOps Engineer
Model and use case approval & deployment



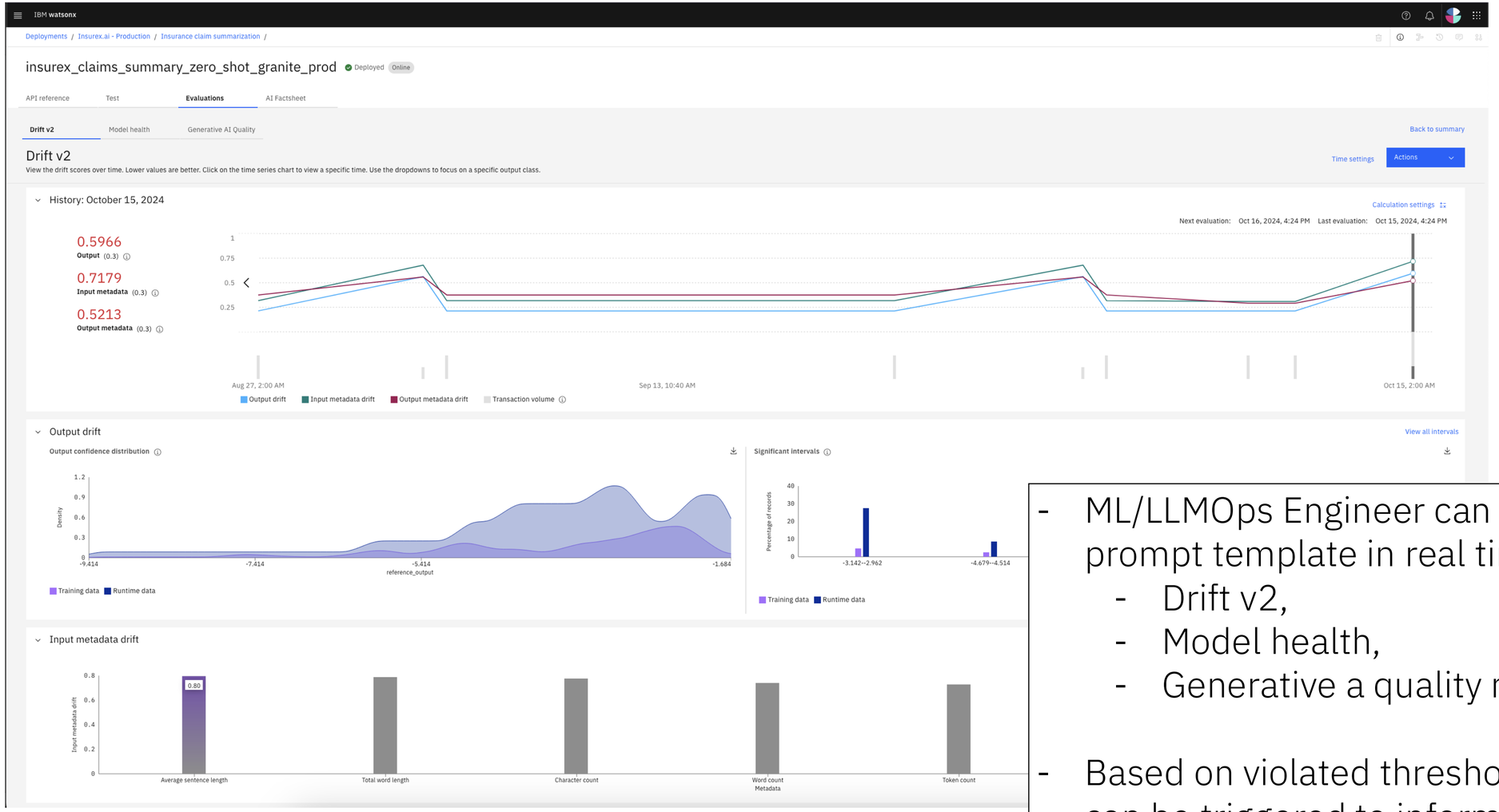
AI development & AI governance
3rd party / external Foundation Model support: watsonx.governance 2.0



Deployments support “headless” evaluations: Push model input & output to deployment endpoint

AI governance
ML/LLMOps Engineer monitors deployed prompt template

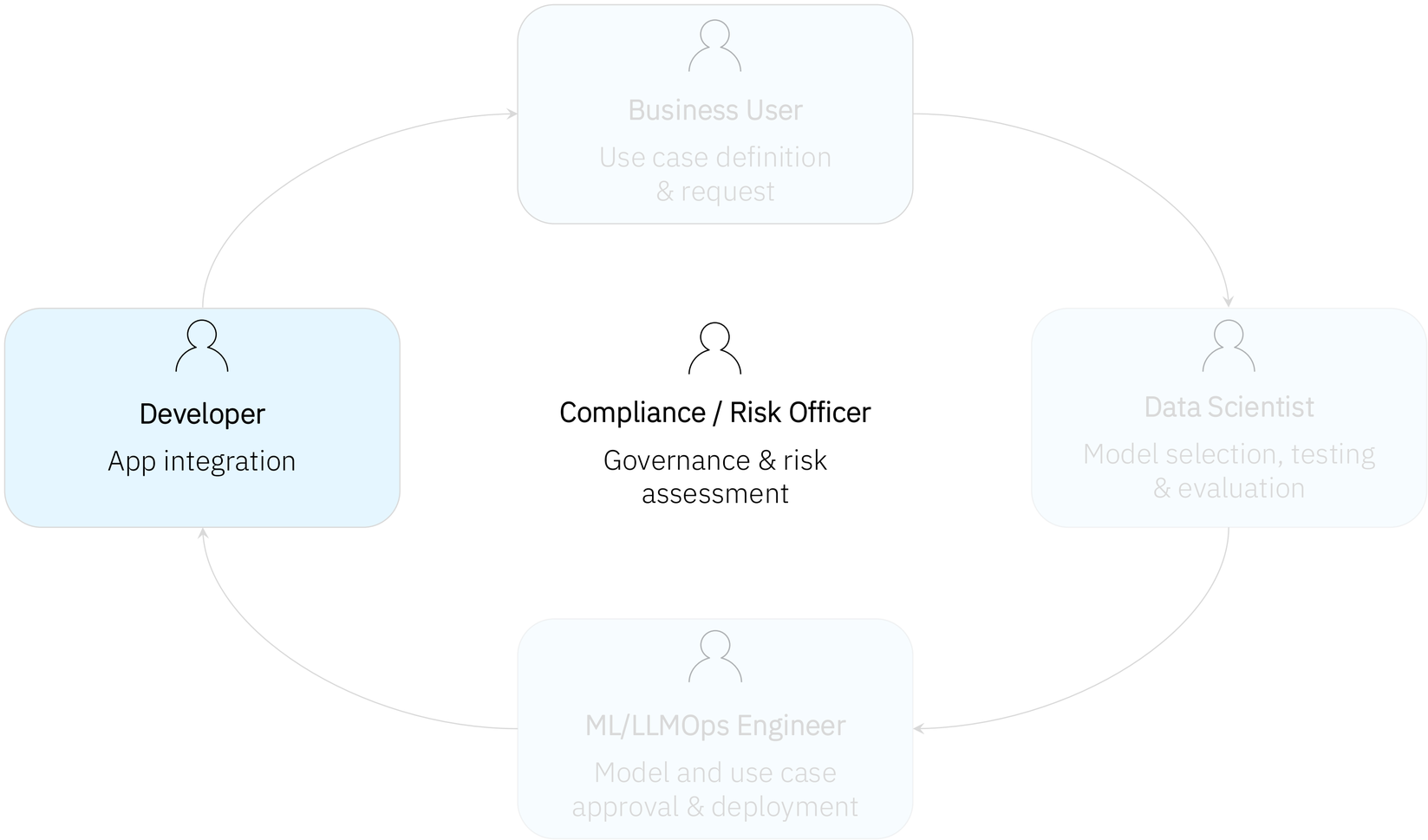
ML/LLMOps Engineer
Model and use case approval & deployment



- 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

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

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-f2b0dafe3dc2/text/generation?version=2021-05-01`

`https://private.us-south.ml.cloud.ibm.com/ml/v1/deployments/a1667036-1899-400c-8e1d-f2b0dafe3dc2/text/generation_stream?version=2021-05-01`

Public endpoint

`https://us-south.ml.cloud.ibm.com/ml/v1/deployments/a1667036-1899-400c-8e1d-f2b0dafe3dc2/text/generation?version=2021-05-01`

`https://us-south.ml.cloud.ibm.com/ml/v1/deployments/a1667036-1899-400c-8e1d-f2b0dafe3dc2/text/generation_stream?version=2021-05-01`

Bearer <token>

`IAM`

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://dataplatform.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: Programmtic access
Use the development environment of your choice to interact with the watsonx platform (see some samples below)

Developer
App integration

Model inference

```python
from ibm_watsonx_ai.foundation_models import ModelInference
from ibm_watsonx_ai.metanames import GenTextParamsMetaNames as GenParams
from ibm_watsonx_ai.foundation_models.utils.enums import ModelTypes, DecodingMethods

# To display example params enter
GenParams().get_example_values()

generate_params = {
    GenParams.MAX_NEW_TOKENS: 200
}

model_inference = ModelInference(
    model_id=ModelTypes.GRANITE_13B_CHAT_V2,
    params=generate_params,
    credentials={
        "apikey": "DHl4To9OTGR_ICEWybY0g9mEBfzSxzqJ8Iu-nYwREPI0",
        "url": "https://us-south.ml.cloud.ibm.com"
    },
    project_id="cb1425f8-e2e3-414a-b469-0182932e7473"
    )

q = "<|system|> You are Granite Chat, an AI language model developed by IBM. You are
    <|user|> You are an insurance agent tasked to assess insurance claims. \
    [Document] \
    On November 1st, 2023, at 11:00 am, my vehicle, a 2004 Honda Civic, was stolen fr
    [End] \
    Summarize the following insurance claim input. Focus on the car and the damage. M
    <|assistant|"

generated_response = model_inference.generate(prompt=q)
print (generated_response)
```

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

LangChain integration

```python
# LangChain integration
from langchain_ibm import WatsonxLLM

flan_ul2_llm = WatsonxLLM(
    model_id=model_id_1,
    url=credentials["url"],
    apikey=credentials["apikey"],
    project_id=credentials["project_id"],
    params=parameters
    )

flan_t5_llm = WatsonxLLM(
    model_id=model_id_2,
    url=credentials["url"],
    apikey=credentials["apikey"],
    project_id=credentials["project_id"],
    params=parameters
    )

#print(flan_ul2_llm.dict())
#print(flan_t5_llm.dicht())

# Sequential Chain experiment

from langchain_core.prompts import PromptTemplate
prompt_1 = PromptTemplate(
    input_variables=["topic"],
    template="Generate a random question about {topic}: Question: "
)
prompt_2 = PromptTemplate(
    input_variables=["question"],
    template="Answer the following question: {question}",
)

from langchain.chains import LLMChain
prompt_to_flan_ul2 = LLMChain(llm=flan_ul2_llm, prompt=prompt_1, outpu
flan_to_t5 = LLMChain(llm=flan_t5_llm, prompt=prompt_2, output_key='ar

from langchain.chains import SequentialChain
qa = SequentialChain(chains=[prompt_to_flan_ul2, flan_to_t5], input_va

print(qa.invoke({"topic": "life"}))
```

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/llms/ibm_watsonx/

Governance metrics toolkit

Azure OpenAI related

```python
import os
import openai
```

Get the Azure OpenAI deployment details from Azure OpenAI Studio

```python
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>'
```

```python
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

```python
# 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
```

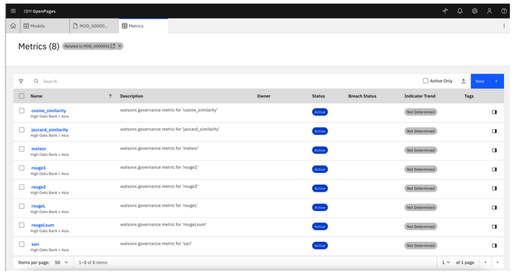
Evaluate Metrics

IBM watsonx.governance authentication

```python
from ibm_cloud_sdk_core.authenticators import IAMAuthenticator,BearerTokenAuthenticator,CloudPakForDataAuthentica

from ibm_watson_openscale import *
from ibm_watson_openscale.supporting_classes.enums import *
from ibm_watson_openscale.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
        )

    client = APIClient(service_url=WOS_CREDENTIALS['url'],authenticator=authenticator)
    print(client.version)
else:
    authenticator = IAMAuthenticator(apikey=CLOUD_API_KEY)
    client = APIClient(authenticator=authenticator)
    print(client.version)
```

Metrics configuration for evaluation

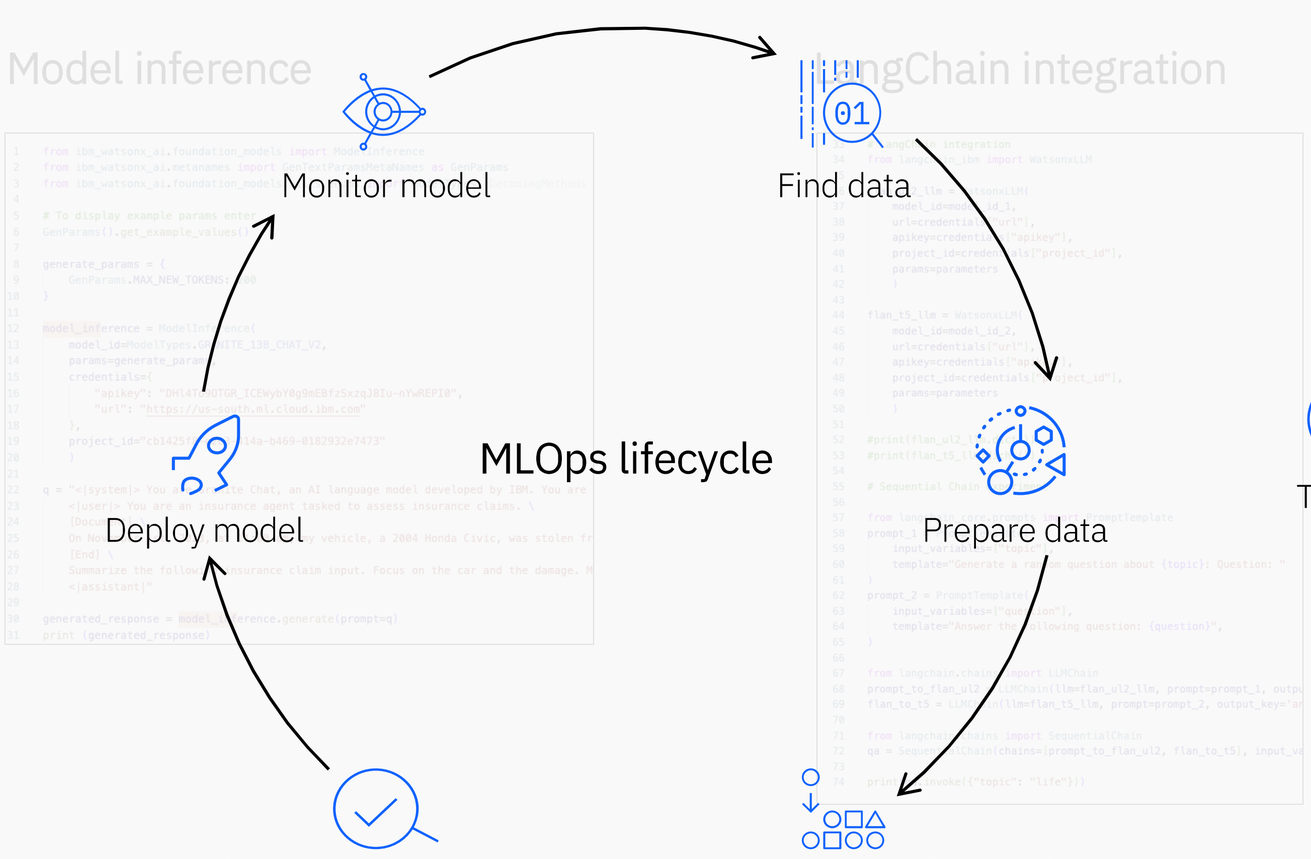
```python
metric_config = {
    "configuration": {
        LLMTextMetricGroup.SUMMARIZATION.value: {
            LLMSummarizationMetrics.ROUGE_SCORE.value: {},
            LLMSummarizationMetrics.SARI.value: {},
            LLMSummarizationMetrics.METEOR.value: {},
            LLMSummarizationMetrics.NORMALIZED_RECALL.value: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

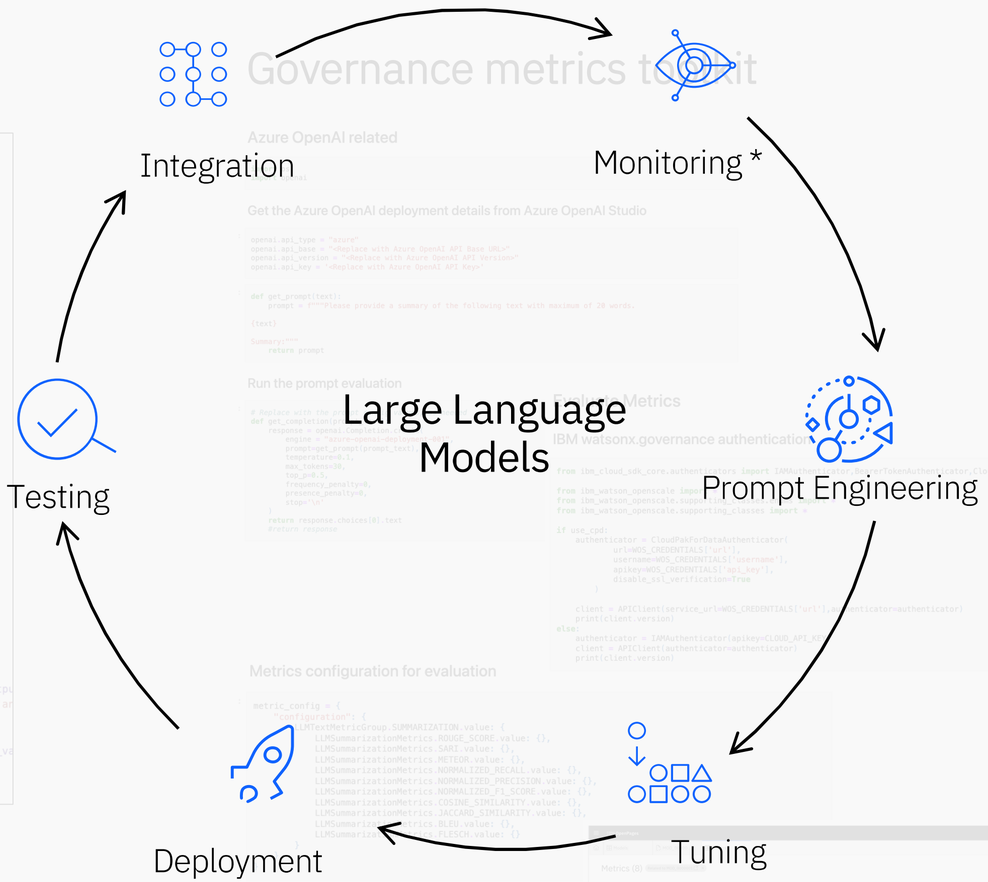


AI development: Programmtic access

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

Developer
App integration





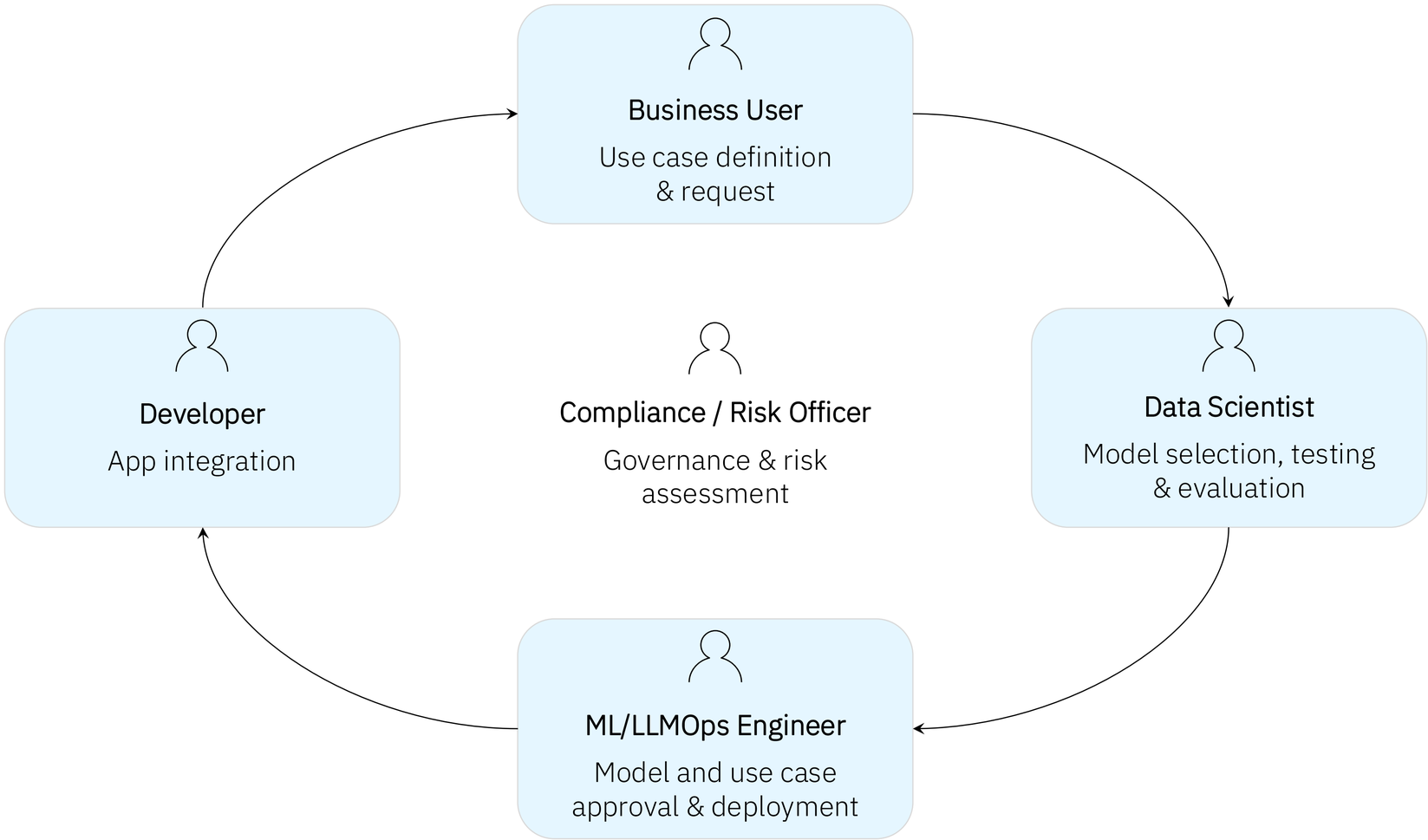
Product documentation
https://dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/fm-python-lib.html?context=wx

ibm-watsonx-ai Python library
https://ibm.github.io/watsonx-ai-python-sdk/foundation_models.html

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/llms/ibm_watsonx/

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 ...	Work on this claim
CIC202105895		1	On August 10th, 2023, the insured vehicl...	Work on this claim
CIC202108708		1	The insured vehicle, a Chevrolet Silvera...	Work on this claim
CIC202109875		2	On May 30th, 2023, heavy rainfall caused...	Work on this claim
CIC202105995		2	While driving on September 8th, 2023, th...	Work on this claim
CIC202109635		3	On January 15th, 2023, the insured vehic...	Work on this claim
CIC202123405		3	The insured vehicle, a Honda Accord, was...	Work on this claim
CIC202103405		3	The insured vehicle, a BMW 3 Series, was...	Work on this claim
CIC202105005		3	The insured vehicle, a Toyota Camry, was...	Work on this claim
CIC202105505		3	The insured vehicle, a Tesla Model S, wa...	Work on this claim

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.

Generated at: 10:50 04/04/2024

Case Number: CIC202107863

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	☐ Verify the insured's policy coverage to ensure that it includes coverage for fire damage and the necessary repairs. ☐ 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. ☐ 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. ☐ 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. ☐ 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. ☐ 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

Finalize Claim

AI development & AI governance
Example application from the business user perspective

Compliance / Risk Officer

Governance & risk assessment