*Good Decisions: A Monthly Webinar for Enterprise AI Governance Insights*

*Episode 9*

# The Explainable AI Dilemma

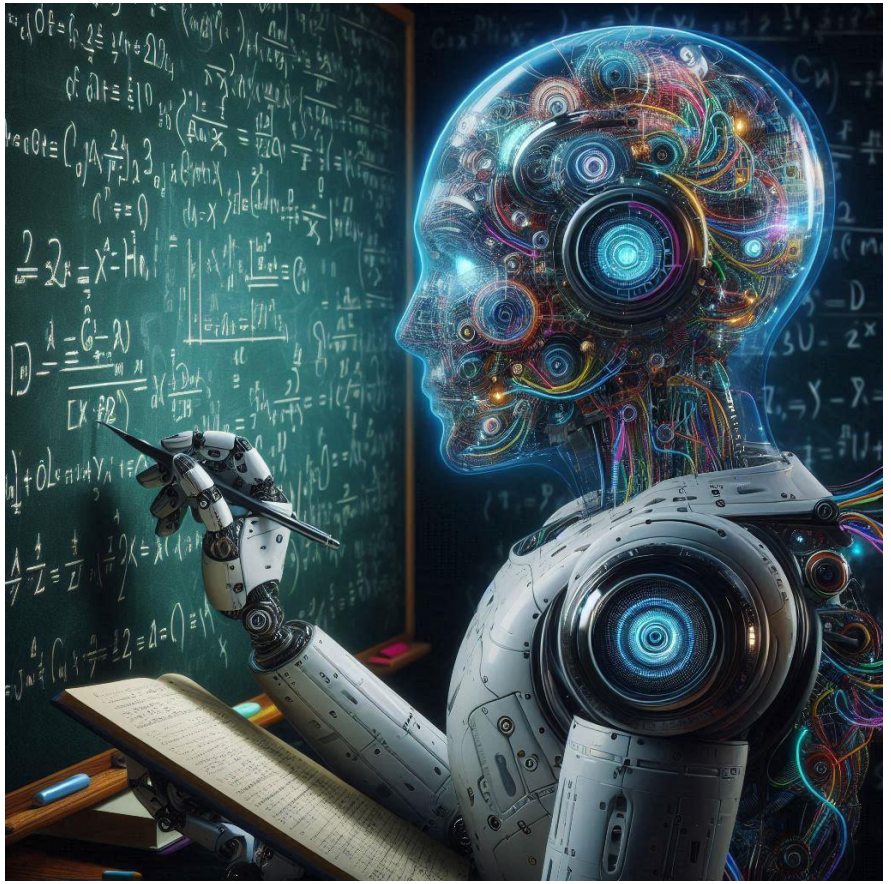
November 19, 2024

**Jim Olsen, CTO at ModelOp**

ModelOp

# The Explainable AI Dilemma

- What is explainable AI?
- Traditional ML model explainability techniques
- What can be done with Generative AI?
- Understand uses of AI in your organization



# What is Explainable AI?

| | |
|---|---|
| **Trustworthiness** | Provide faith in that the given model will behave towards the given problem in a consistent and predictable pattern |
| **Causality** | Finding which features develop a causal relationship to arrive at their conclusions |
| **Informativeness** | What is going on inside the model when it is reaching its decisions, and how is it arriving at its conclusions? |
| **Confidence** | How robust and stable is this model towards its intended task? |
| **Fairness** | Allow for the ethical analysis of if a model is biased in its predictions |
| **Privacy** | Is the model harboring any data that could disclose private information? |

Source: Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI https://arxiv.org/pdf/1910.10045

ModelOp

*Explainability Techniques for Traditional AI Models*

# Traditional AI (ML) Models May Already Be Inherently Interpretable In Nature

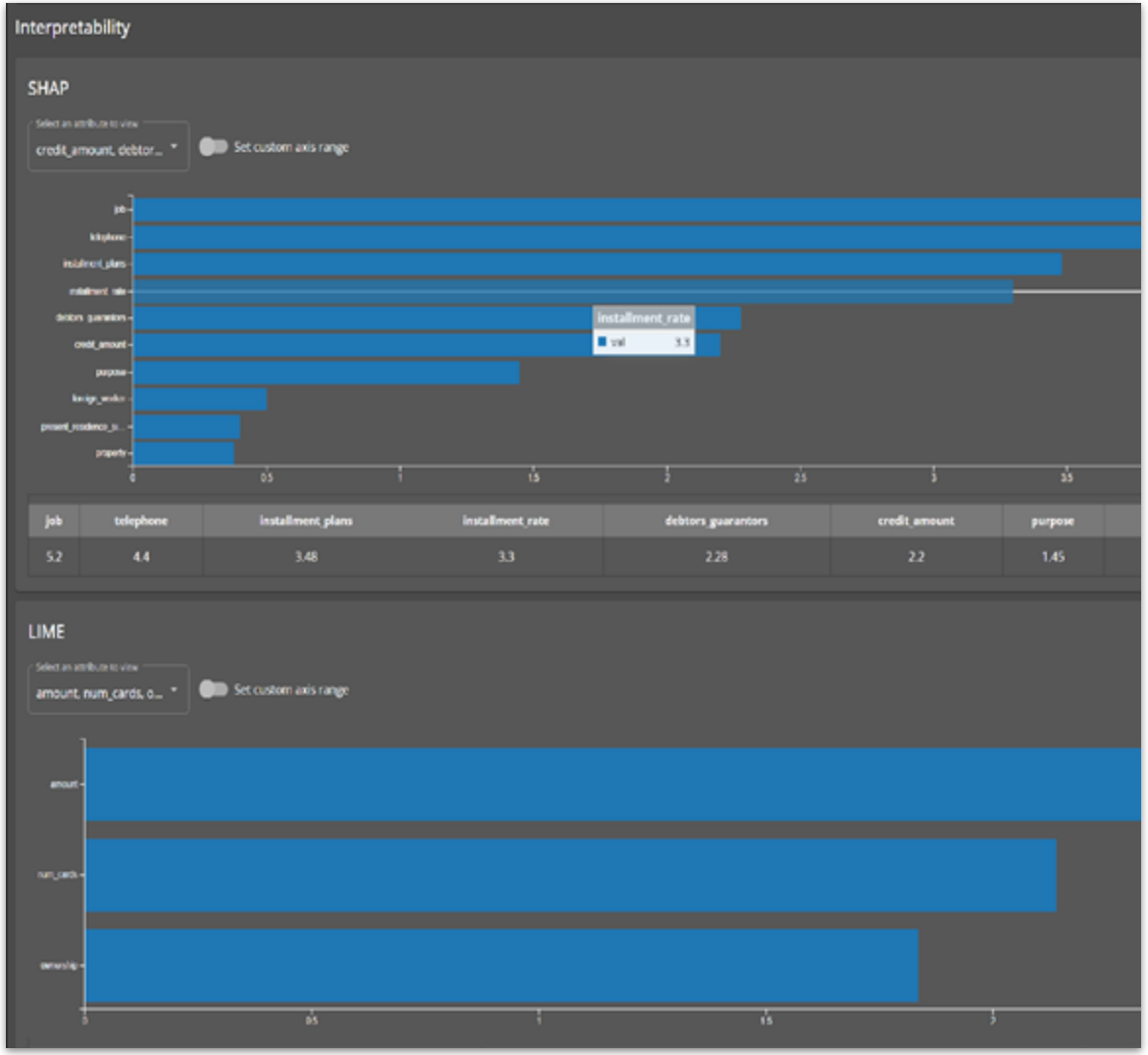
- Complexness, Decomposability, and Algorithmic Transparency come to play here

- Examples:
  - Linear/Logistic Regression
  - Decision Trees
  - K Nearest Neighbors (KNN)
  - Rules Definitions
  - Bayesian Models

ModelOp

*Explainability Techniques for Traditional AI Models*

# Post-hoc Interpretability Techniques



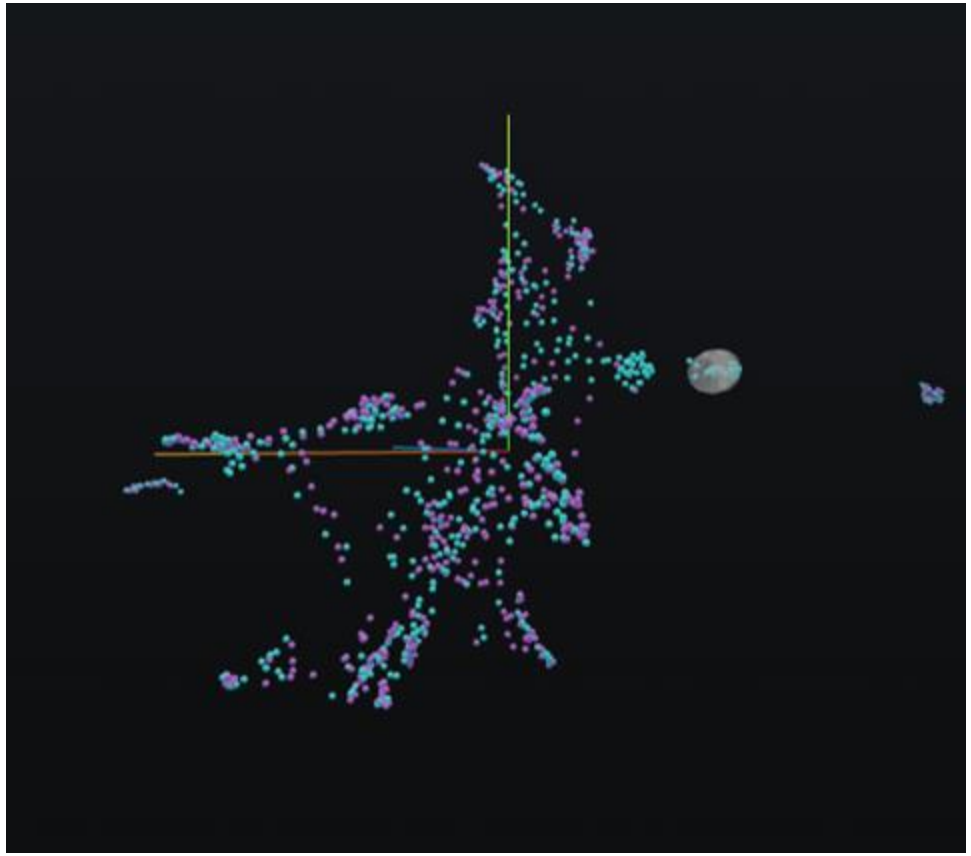

- **Feature importance techniques – SHAP (Shapley Additive exPlanations)**
  - Game Theory to determine individual feature's contribution to the prediction
- **Explainable Proxy Model – LIME (Local Interpretable Model-agnostic Explanations)**
  - Create dataset from permutations of prediction
  - Train new explainable model from that data
  - Use LIME library to interpret the new model

ModelOp

*How Can You Explain GenAI?*

# Generative AI and Explainability

- Generative AI models are "**Fluent, but not factual**"
- Generative AI models are generally considered "**not interpretable**"
- Early experimental techniques
  - Self explanation
    - Through prompts entice the LLM to explain its own decision-making process
    - Given models are 'Fluent, but not factual' results can not be relied upon
  - Visualization techniques
    - Look at vector/node clustering to determine factors in the outcome



ModelOp

*How Can You Explain GenAI?*

# Traceability vs Explainability

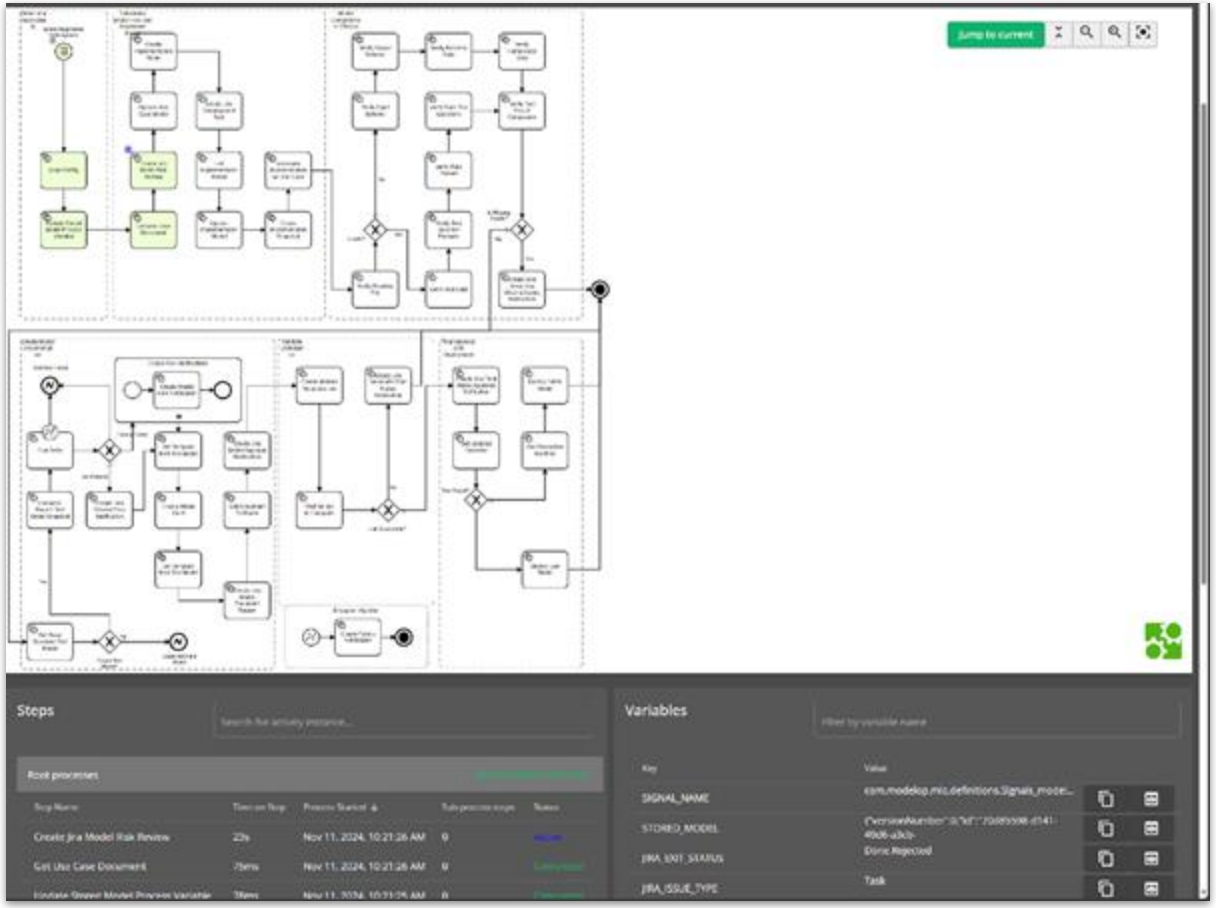
| | Explainability / Interpretability | Traceability |
|---|---|---|
| **Definition** | The ability to **describe how and why a model makes specific decisions** or generates certain outputs | The ability to **track and document the overall process the model followed** to be approved for usage in your organization<br>Includes details about the model, documentation, and other information during model approval |
| **Scope** | Focuses on making the model's behavior understandable to humans | Focuses on the steps followed during the model's approval and eventual deployment |
| **Objective** | Helps users and stakeholder comprehend the model's decision-making process, increasing trust in the model | Ensures accountability and compliance by providing a clear record of the approval process |

ModelOp

*How Can You Explain GenAI?*

# Establish Trust in the Model: **Process**

- Follow a **well documented** and **consistent process**
  - Gather all information available about the model
  - Automate standard tests that are run on all models for consistent comparison
  - Promote compliance with any applicable regulations
- **Visibility** for users into the results of this process
  - All models following the same process increases trust
  - Transparency in where your models are at within this process




ModelOp

*How Can You Explain GenAI?*

# Establish Trust in the Model: **Documentation**

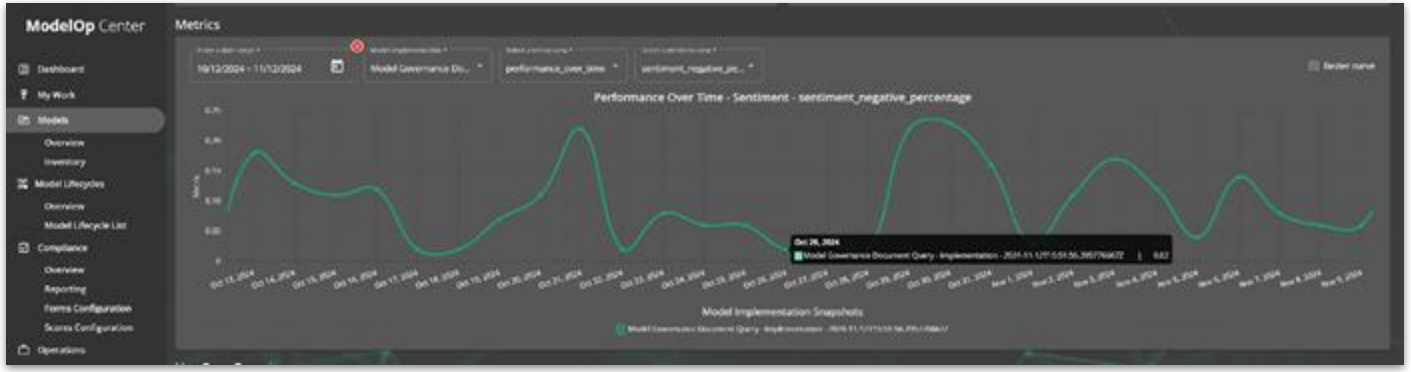
## Automate Consistent Documentation

- Manually creating model documentation is time consuming, prone to error, and tends to grow stale
- Automation leveraging your model inventory eases the pain of documentation
- Multiple forms of documentation required to help establish trust
  - Formal detailed review documents for deep dives
  - Model Cards for quick understanding of a model


*How Can You Explain GenAI?*

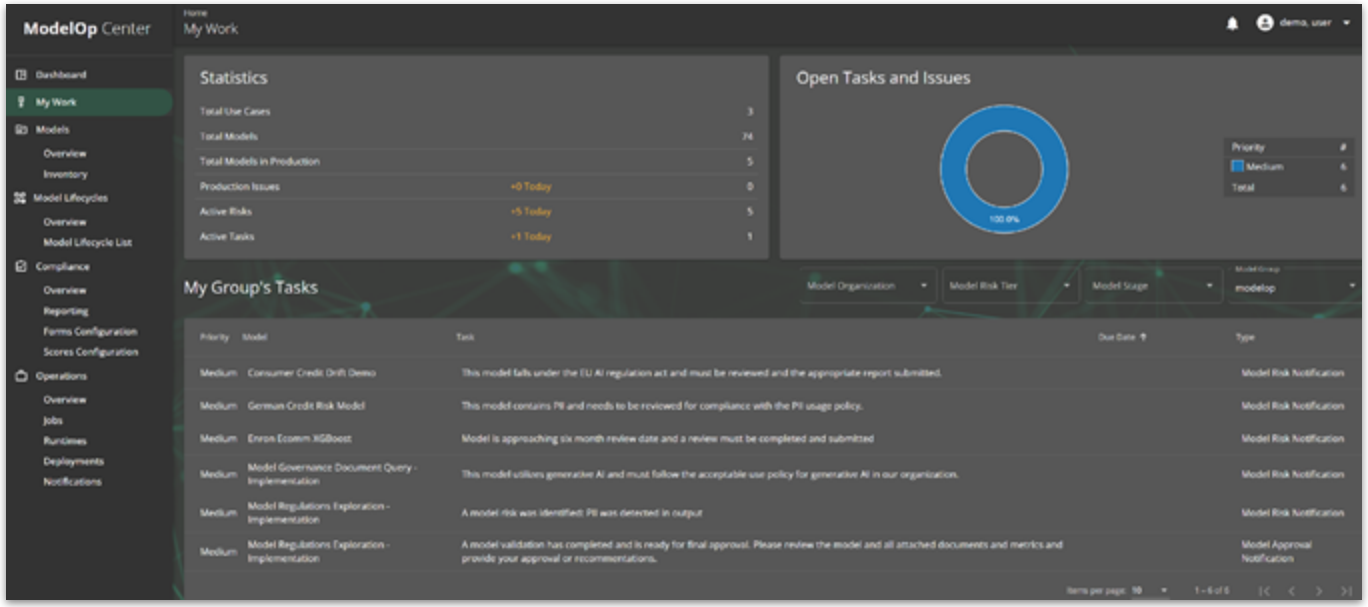
# Establish Trust in the Model: **Baselines**

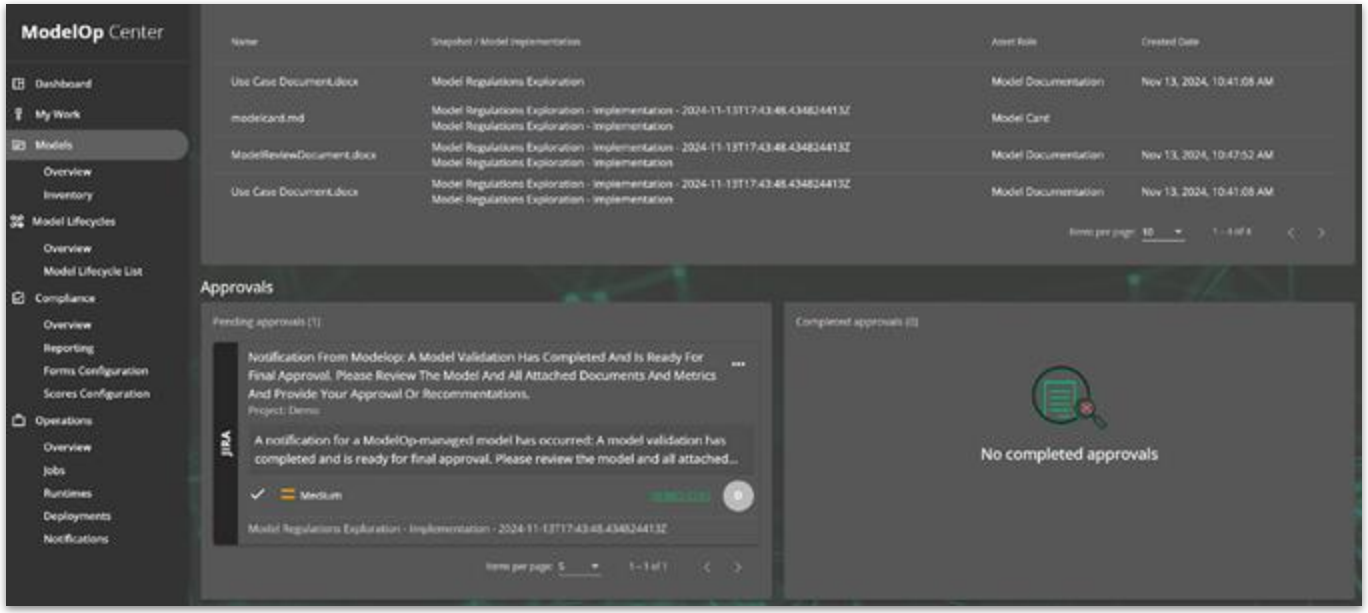
- **Automate consistent baseline metrics**
  - All models require baseline metrics before deployment
- **Automate ongoing runs of more recent data**
  - Detection of anomalous drift
  - When detected, document any findings and approvals
- **Access to results for consumers of the model**
  - Provide insight on the model's performance over time.




*How Can You Explain GenAI?*

# Establish Trust in the Model: **Attestations**



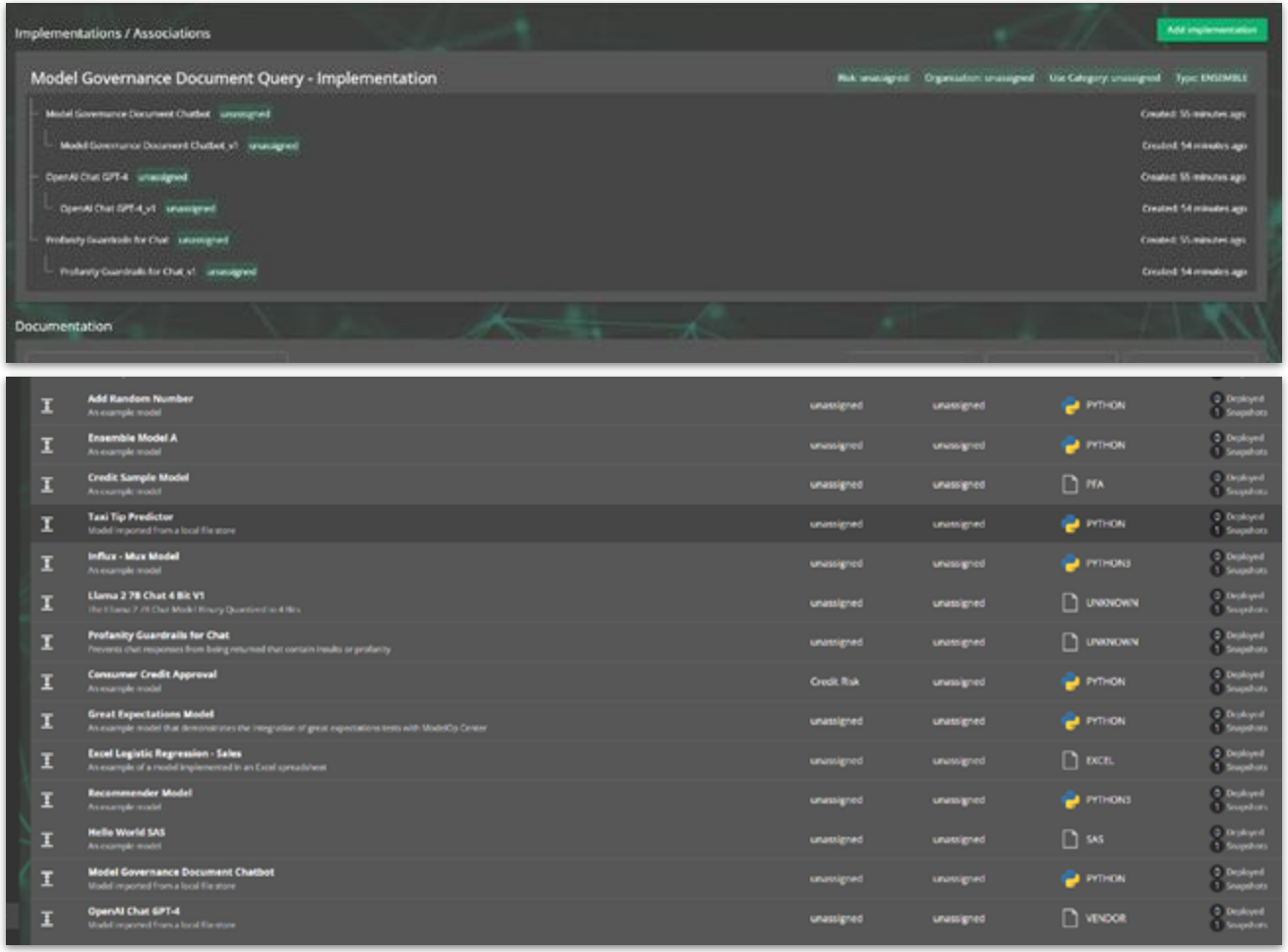




- **Collect attestations at appropriate points in the model lifecycle**
- **Approvals at various stages from humans are required**
  - Initial model development
  - Annual reviews
  - Onboarding of new vendor models
  - Detected exceptional conditions
- **Attestations contain information about who, what, when, and why**

ModelOp

*How is AI Being Used in Your Organization?*

# Looking Back – Uses of AI in Your Enterprise

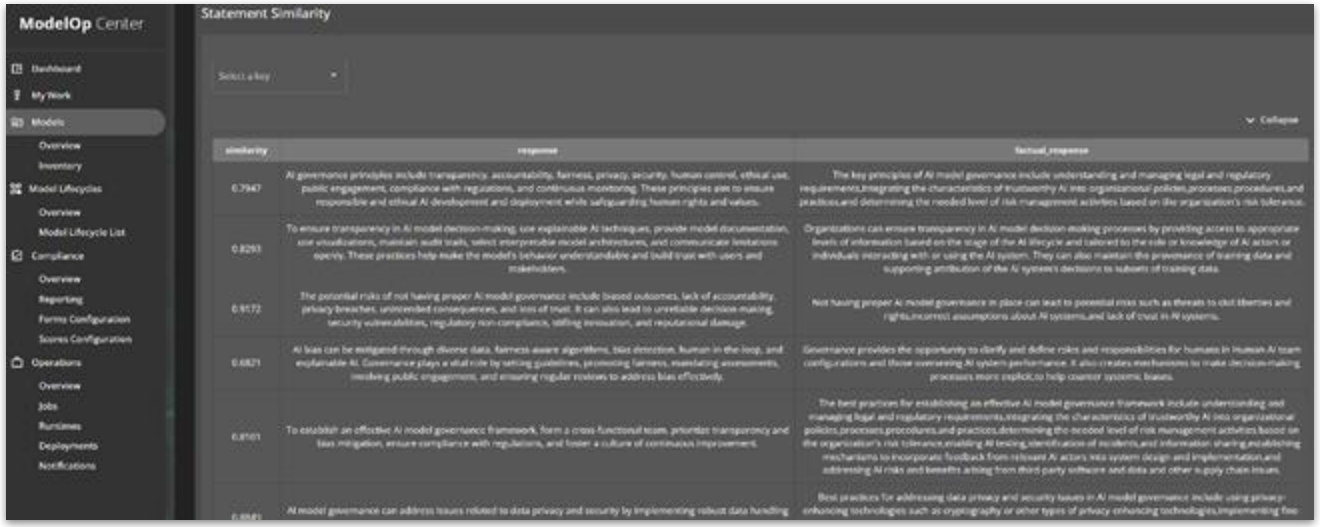


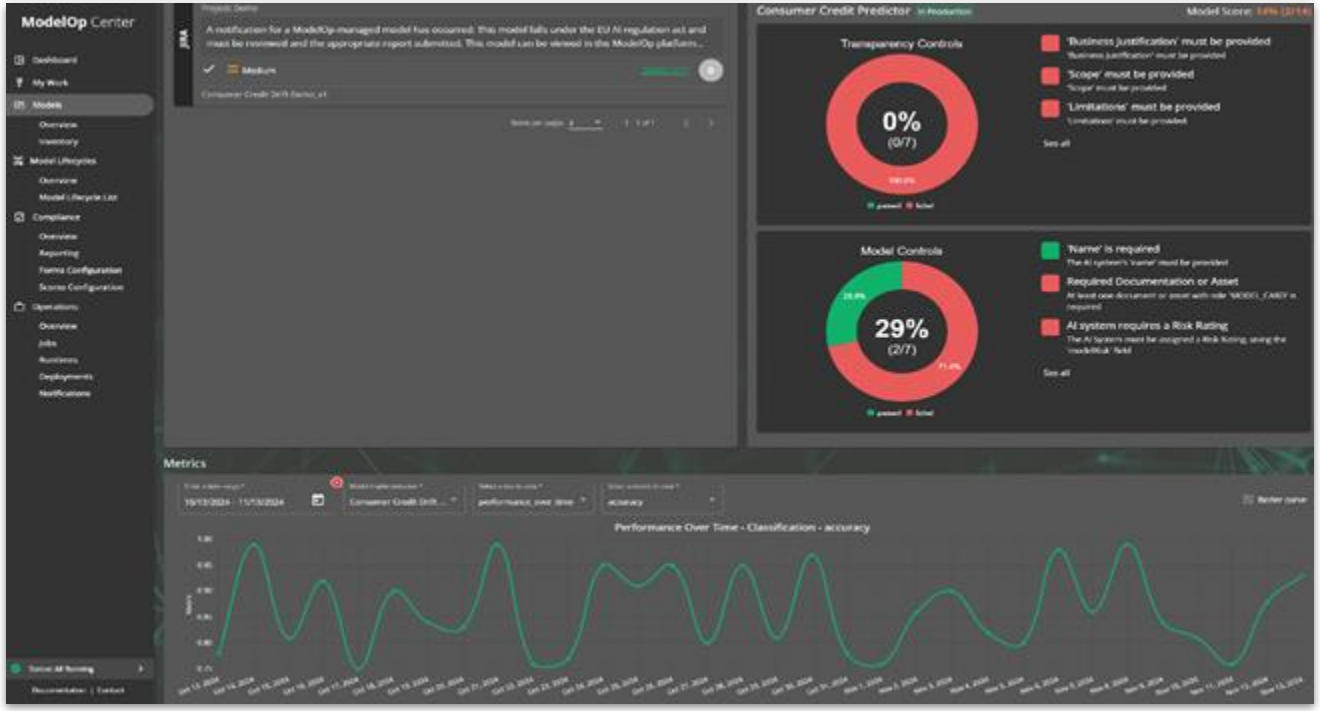

- **What vendor models are in your organization and what are they used for?**
  - Has the use case been approved specifically for generative AI?
  - Are there processes in place to ensure privacy?
- **What foundation models are used locally / internally in your organization?**
  - Each new version of models, such as Llama2 and others should be specifically reviewed
  - Understand affected applications when an instance is upgraded

ModelOp

*How is AI Being Used in Your Organization?*

# Monitors, Use Cases, and Results



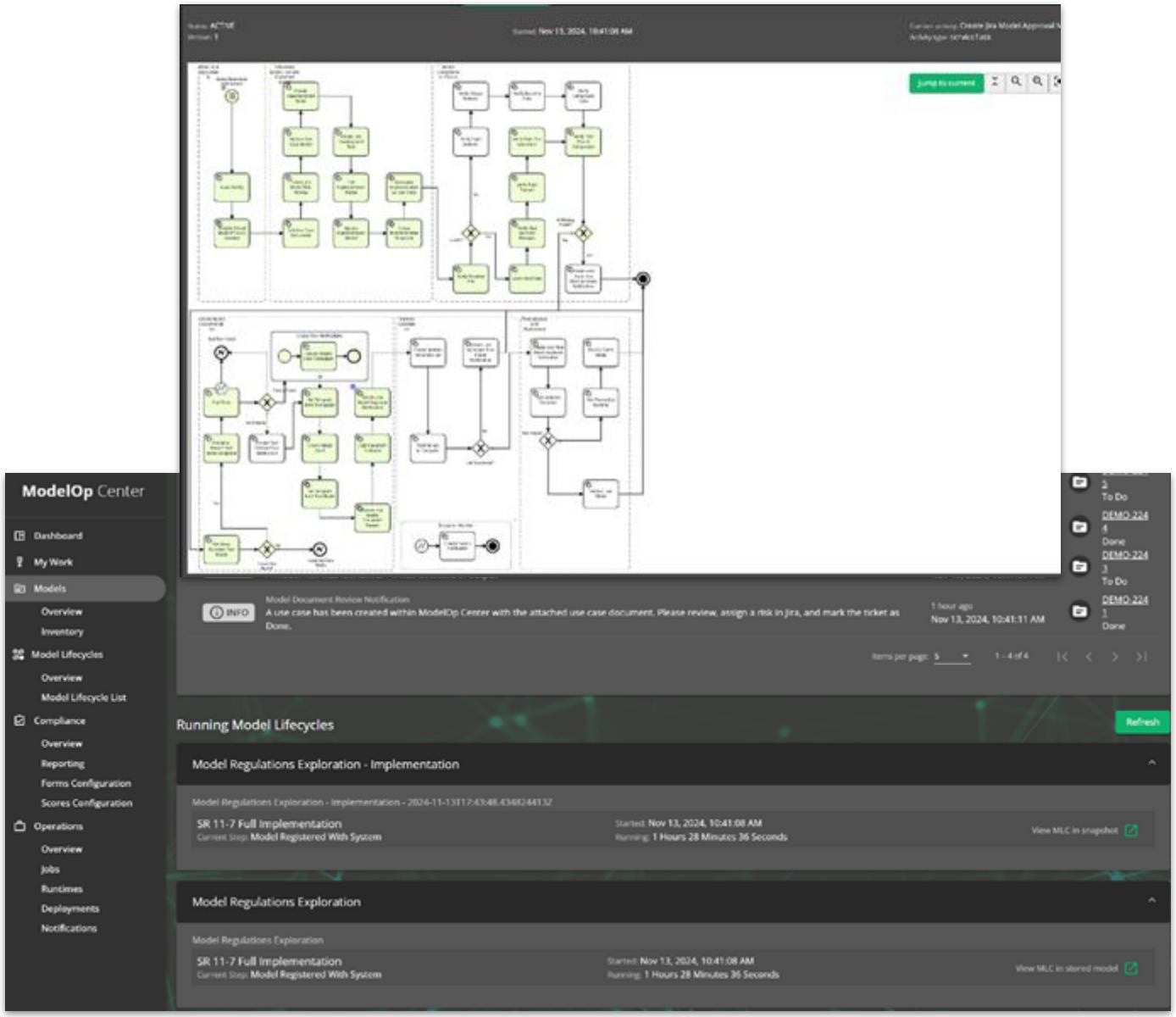

## Monitors should consider the business use case in their choice and execution

- ModelOp Center provides necessary information for this
- Monitors should be diverse and extensible
- Monitor results should drive process
  - Auto generate exception reports on violations
  - Detect the requirement to conduct annual reviews of models and automate the process

ModelOp

*How is AI Being Used in Your Organization?*

# Prove That You Are Adhering to Regulations




- Track the path models take through the entire journey
- On each stage of the journey, make sure you can capture the information about what was done and by who
- ModelOp Center provides a comprehensive view of every step within the process
- Allows instant auditability of adherence to each required step of a regulation

*Trust is Lost Quickly, Get It Right The First Time!*

# The Importance of Trust and Getting it Right: **Dollars and Sense**

- **McDonald's ends AI experiment with IBM**
  On TikTok a video appears of the drive thru repeatedly adding Chicken McNuggets to their order despite their pleas, reaching a total of 260 orders

- **iTutor Group's recruiting AI model rejects older applicants**
  Company had to settle in a $365,0000 payment to the EEOC for rejecting applicants older than 55

- **Widely used healthcare algorithm ignores black patients in need**
  Algorithm used by healthcare companies and insurance were far less likely to flag black patients in need of high-risk care

- **Lawyer used court cases hallucinated by Chat GPT**
  Chat GPT used to look up prior case law by lawyer, and six of the cases submitted in the brief were made up.  The lawyer was fined $5000

ModelOp

*Trust is lost quickly, get it right the first time!*

# How To Build Trust With GenAI and Third-party Vendor Models

- **Get visibility** into all your AI initiatives — including GenAI and Third-party vendor — across your enterprise today
- Enable your teams with **explainability, interpretability, and traceability** capabilities
- Establish trust in your models with **process, documentation, baselines, and attestations**
- Partner with a **proven AI Governance software** provider, like **ModelOp,** and establish Minimum Viable Governance (MVG)

ModelOp

# Thank You

## Request a personalized demo of MVG for your enterprise

https://www.modelop.com/demo-request

www.modelop.com

ModelOp is the leading AI Governance software for enterprises and helps safeguard all AI initiatives — including generative AI, Large Language Models (LLMs), in-house, third-party, and embedded systems — without stifling innovation. Through automation and integrations, ModelOp empowers enterprises to quickly address the critical governance and scale challenges necessary to protect and fully unlock the transformational value of enterprise AI — resulting in effective and responsible AI systems.