

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

*Episode 9*

# **The Explainable AI Dilemma**

November 19, 2024

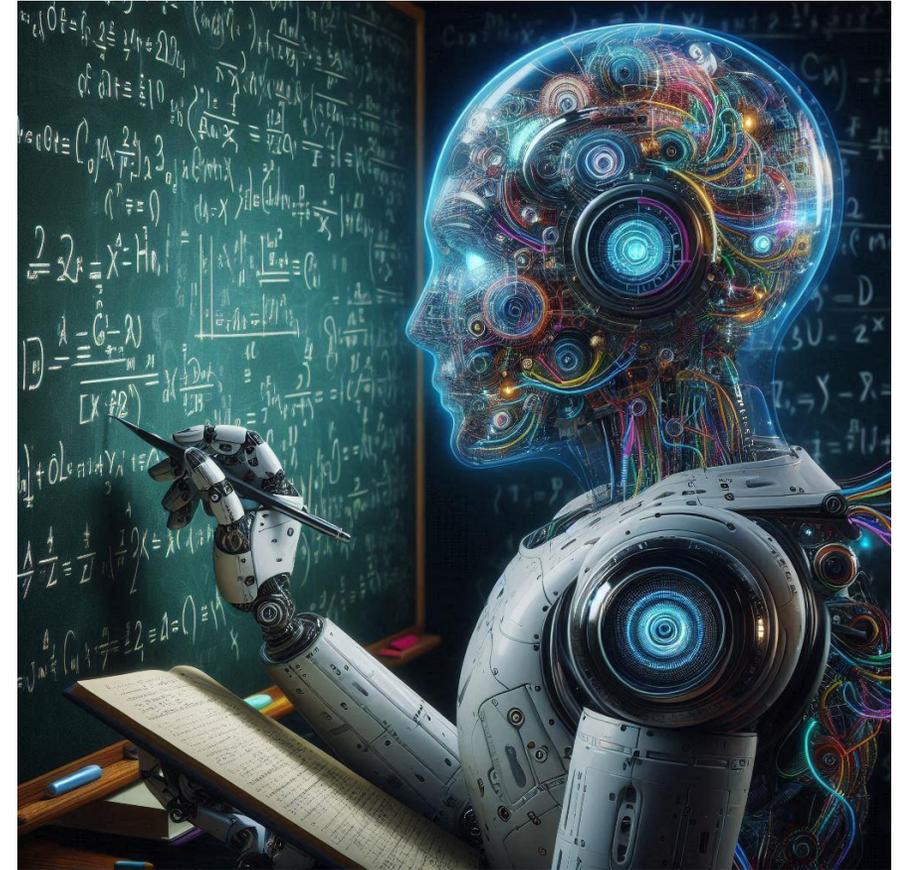
---

Jim Olsen, CTO at 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?

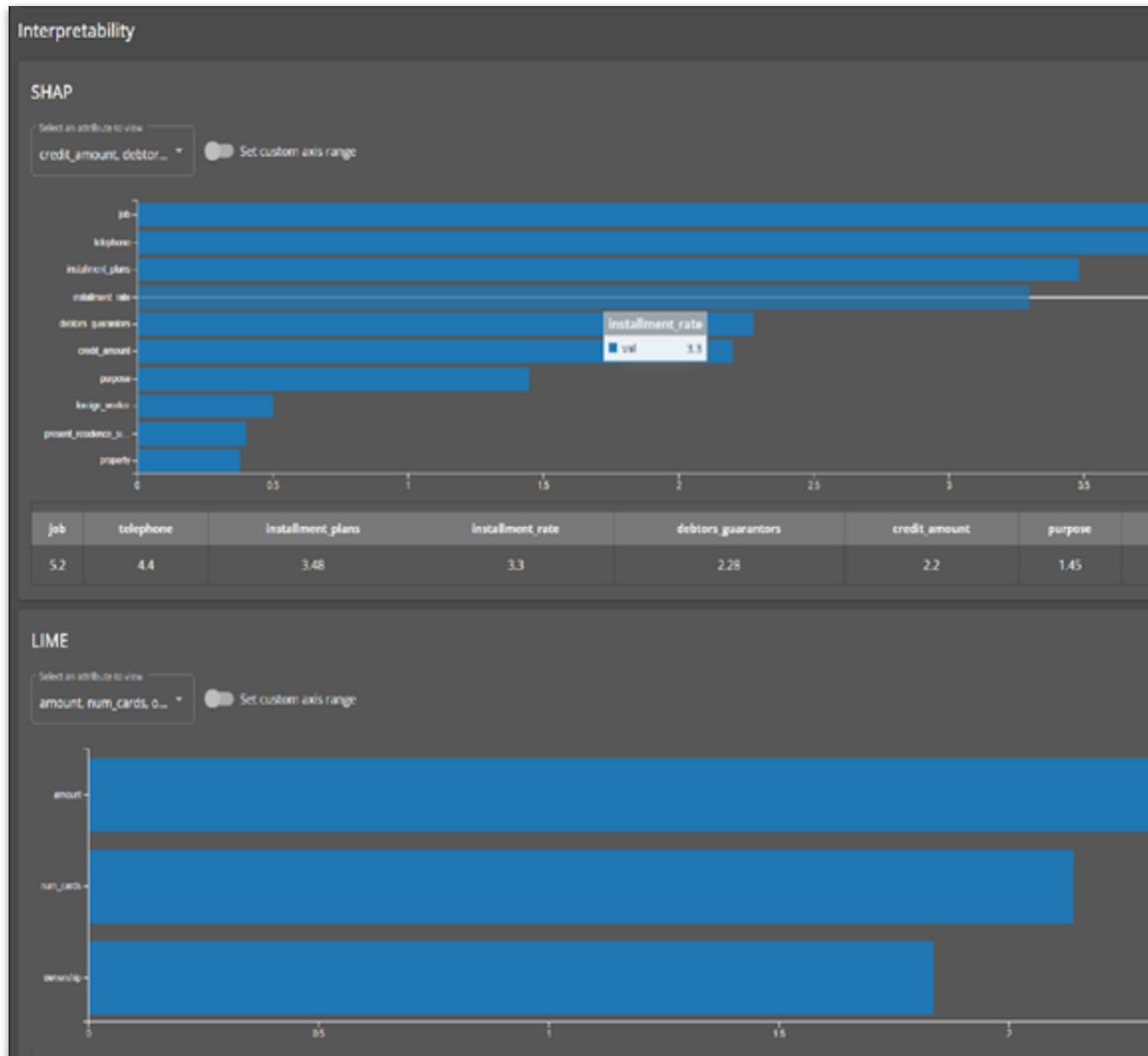
# Traditional AI (ML) Models May Already Be Inherently Interpretable In Nature



- Complexity, Decomposability, and Algorithmic Transparency come to play here
- Examples:
  - Linear/Logistic Regression
  - Decision Trees
  - K Nearest Neighbors (KNN)
  - Rules Definitions
  - Bayesian Models

# Explainability Techniques for Traditional AI Models

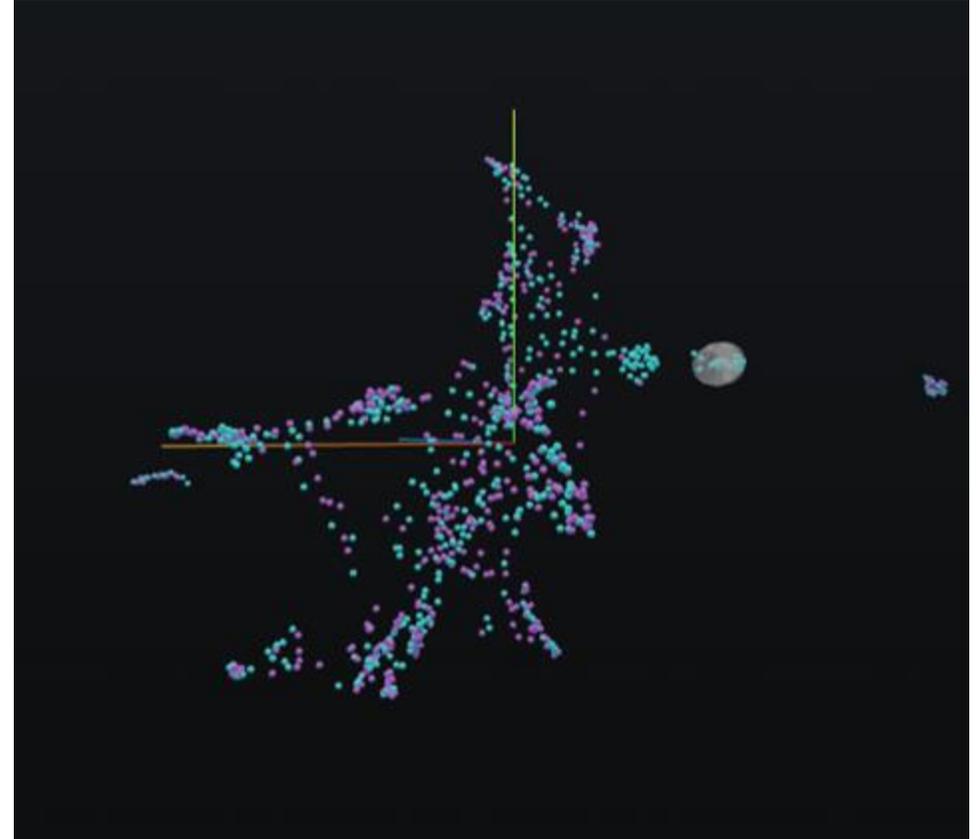
## Post-hoc Interpretability Techniques



- **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

# 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

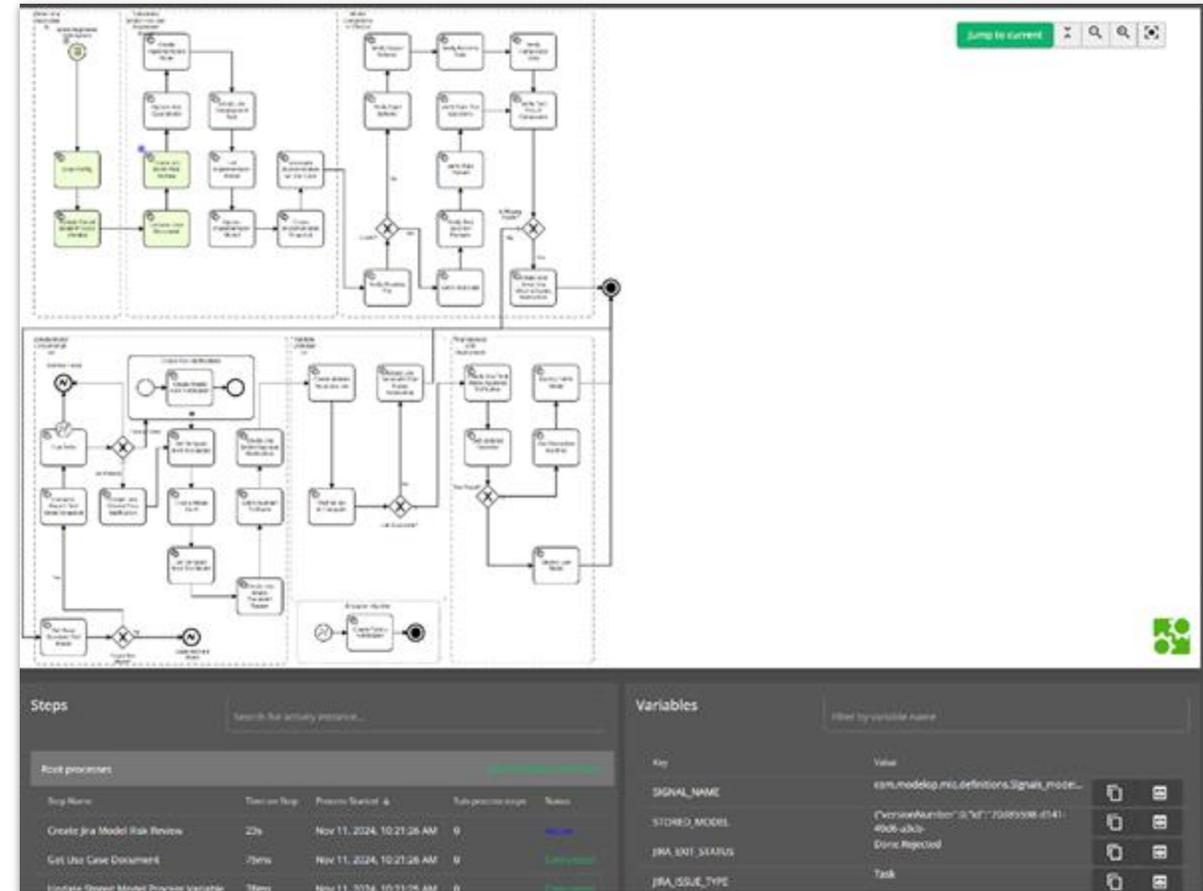


# Traceability vs Explainability

	Explainability / Interpretability	Traceability
Definition	The ability to <b>describe how and why a model makes specific decisions</b> or generates certain outputs	The ability to <b>track and document the overall process the model followed</b> to be approved for usage in your organization 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

# Establish Trust in the Model: Process

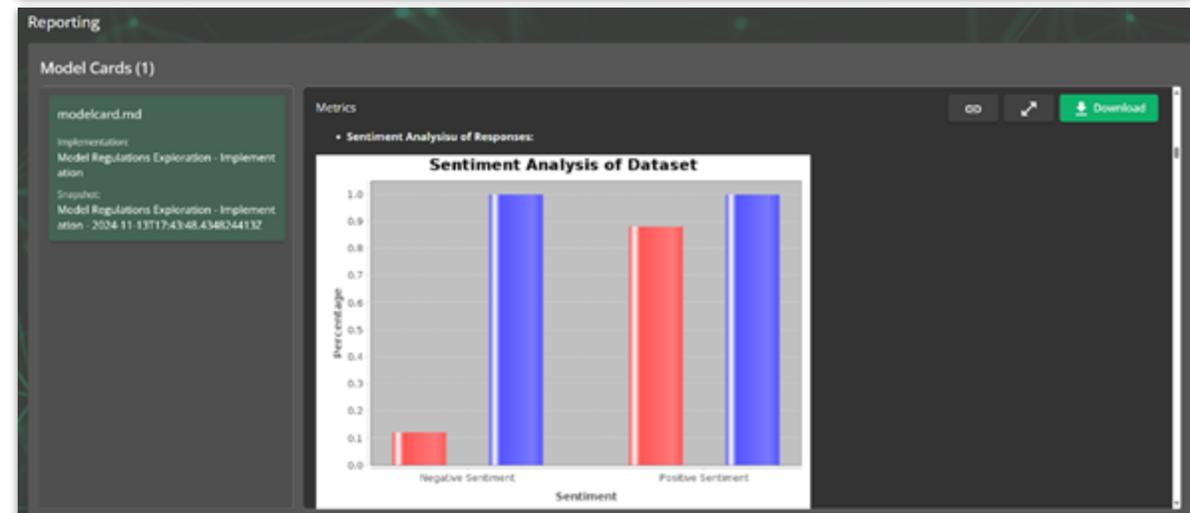
- Follow a **well documented** and **consistent process**
  - Gather all information available about the model
  - Automate standard tests that are run on all models for consistent comparison
  - Promote compliance with any applicable regulations
- **Visibility** for users into the results of this process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



# Establish Trust in the Model: Documentation

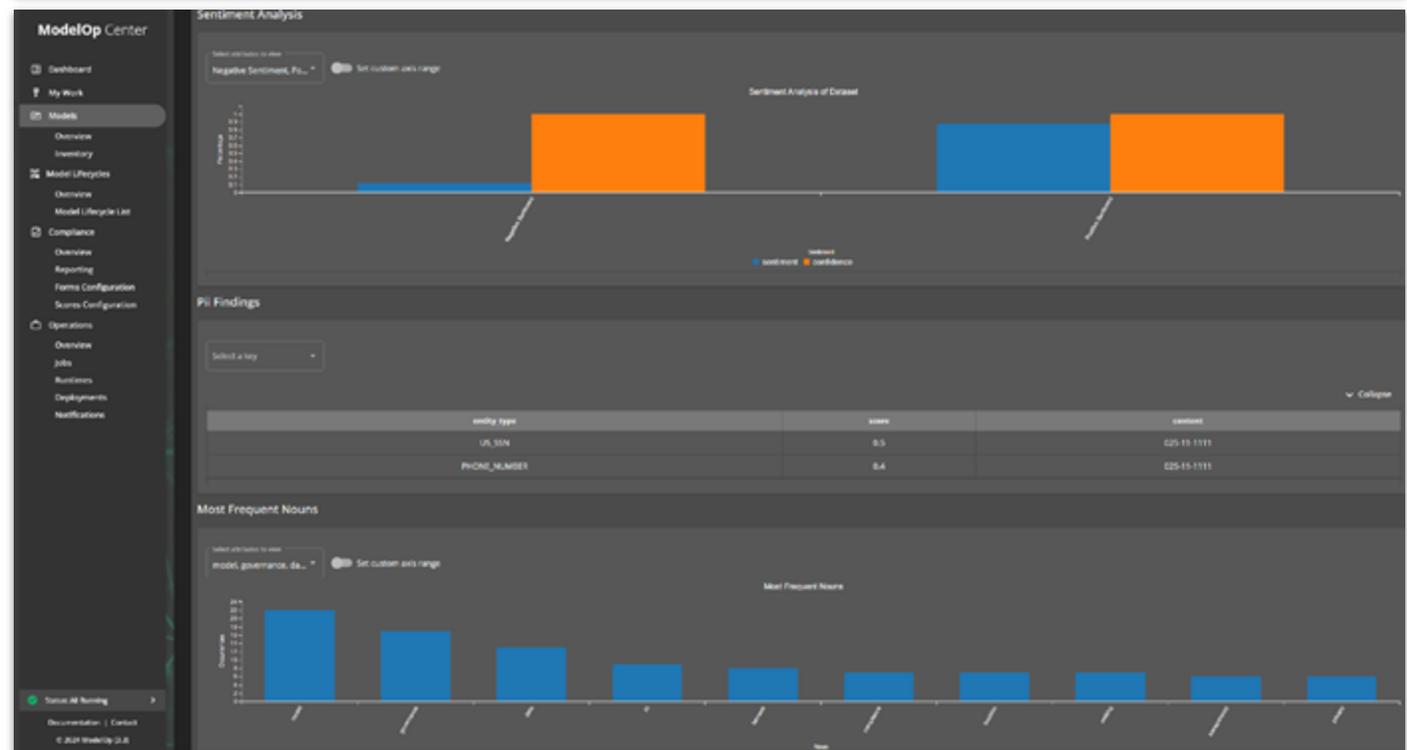
## Automate Consistent Documentation

- 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



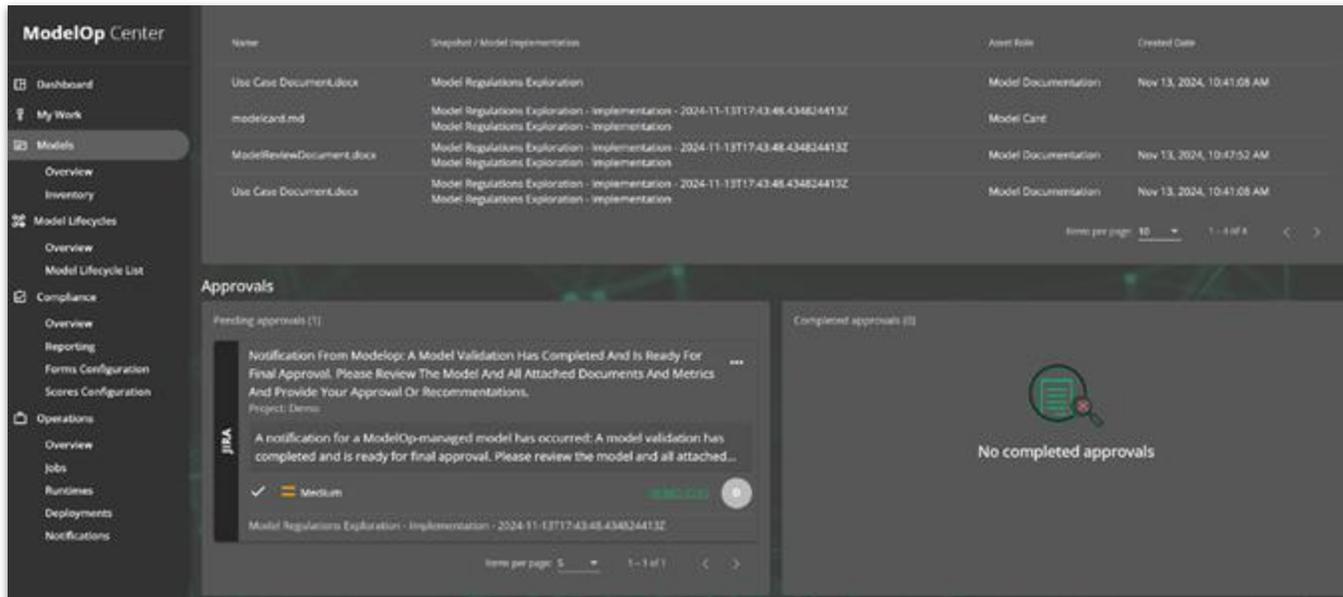
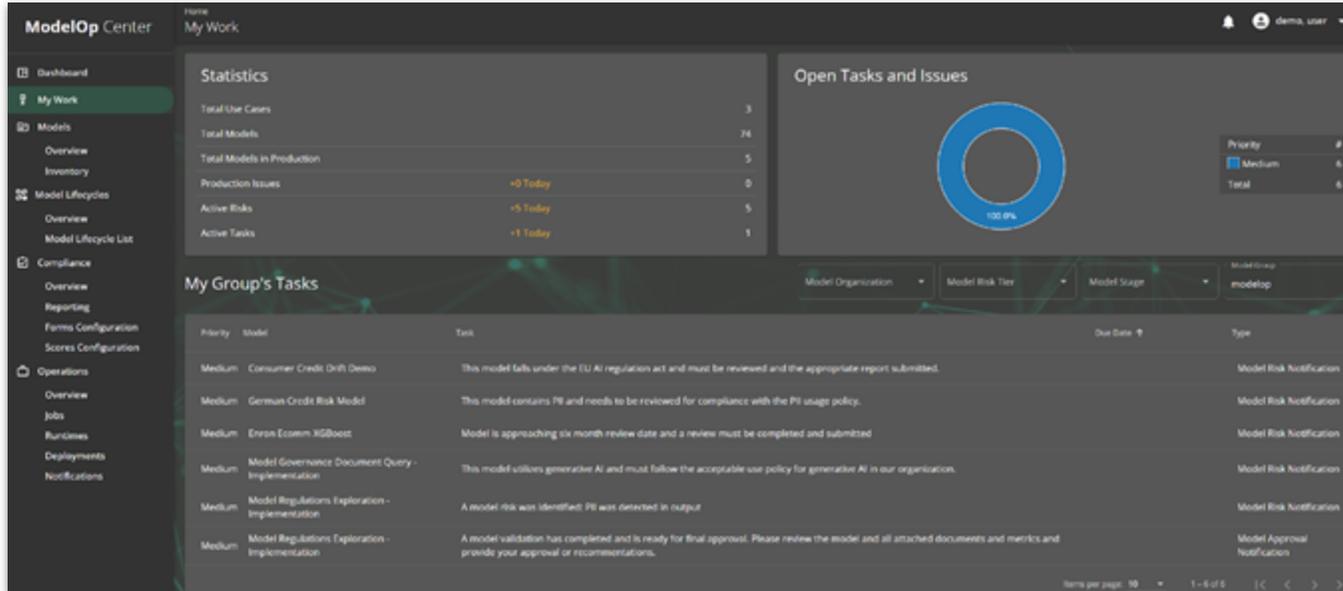
# Establish Trust in the Model: Baselines

- **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?

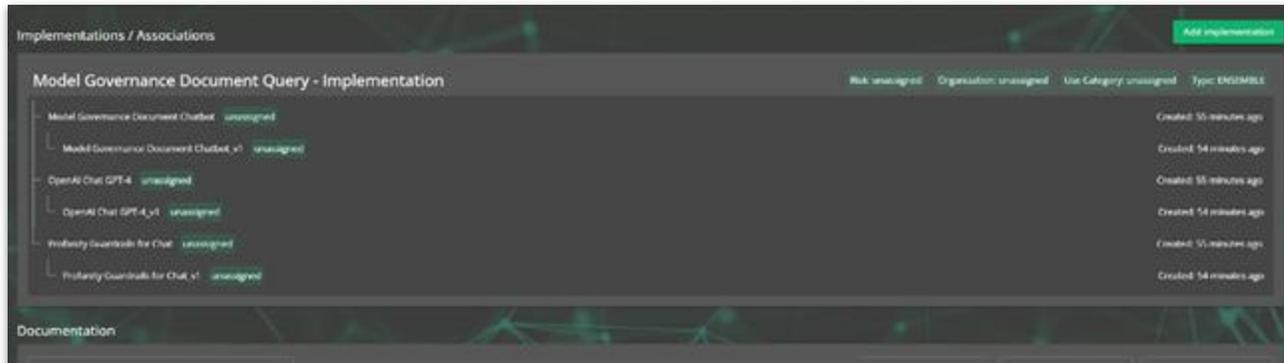
# Establish Trust in the Model: Attestations



- 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**

## How is AI Being Used in Your Organization?

# Looking Back – Uses of AI in Your Enterprise



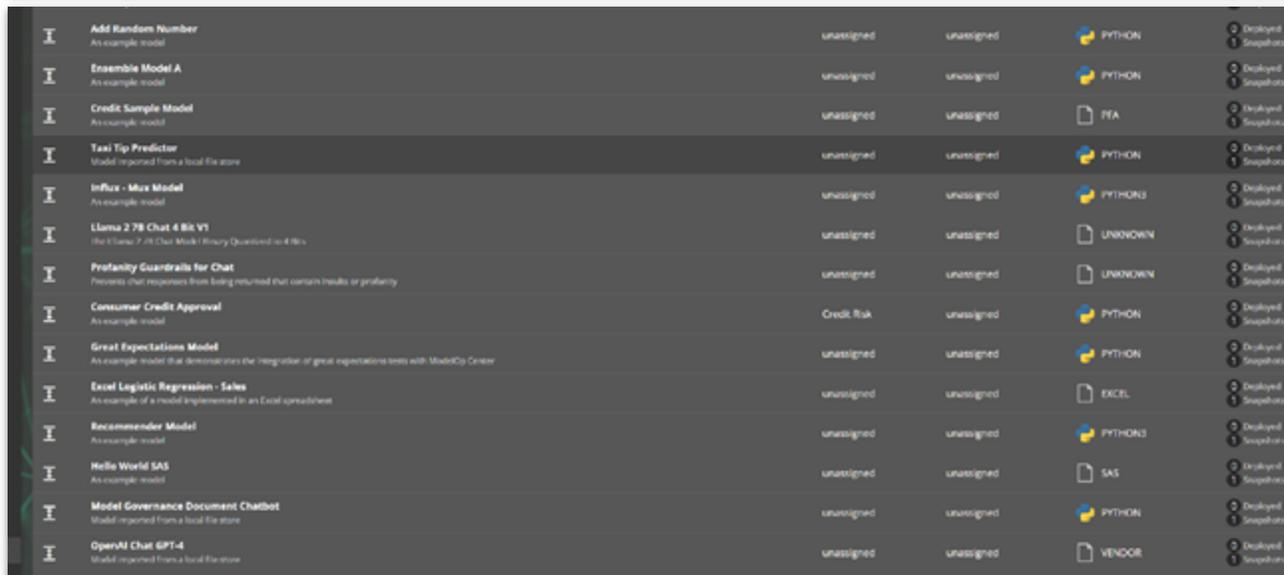
Implementations / Associations

Model Governance Document Query - Implementation

Risk: unassigned Organization: unassigned Use Category: unassigned Type: ENSEMBLE

Model Name	Status	Created
Model Governance Document Chatbot	unassigned	Created 35 minutes ago
Model Governance Document Chatbot v1	unassigned	Created 34 minutes ago
OpenAI Chat GPT-4	unassigned	Created 35 minutes ago
OpenAI Chat GPT-4 v1	unassigned	Created 34 minutes ago
Profanity Guardrails for Chat	unassigned	Created 35 minutes ago
Profanity Guardrails for Chat v1	unassigned	Created 34 minutes ago

Documentation



Model Name	Description	Risk	Organization	Use Category	Type	Deployment Status
Add Random Number	An example model	unassigned	unassigned	PYTHON	Deployed Snapshots	
Ensemble Model A	An example model	unassigned	unassigned	PYTHON	Deployed Snapshots	
Credit Sample Model	An example model	unassigned	unassigned	PTA	Deployed Snapshots	
Taxi Tip Predictor	Model imported from a local file store	unassigned	unassigned	PYTHON	Deployed Snapshots	
Influx - Mux Model	An example model	unassigned	unassigned	PYTHON	Deployed Snapshots	
Llama 2 7B Chat 4 Bit V1	The Llama 2 7B Chat Model Binary Quantized in 4 bits	unassigned	unassigned	UNKNOWN	Deployed Snapshots	
Profanity Guardrails for Chat	Prevents chat responses from being returned that contain insults or profanity	unassigned	unassigned	UNKNOWN	Deployed Snapshots	
Consumer Credit Approval	An example model	Credit Risk	unassigned	PYTHON	Deployed Snapshots	
Great Expectations Model	An example model that demonstrates the integration of great expectations tests with ModelOps Center	unassigned	unassigned	PYTHON	Deployed Snapshots	
Excel Logistic Regression - Sales	An example of a model implemented in an Excel spreadsheet	unassigned	unassigned	EXCEL	Deployed Snapshots	
Recommender Model	An example model	unassigned	unassigned	PYTHON	Deployed Snapshots	
Hello World SAS	An example model	unassigned	unassigned	SAS	Deployed Snapshots	
Model Governance Document Chatbot	Model imported from a local file store	unassigned	unassigned	PYTHON	Deployed Snapshots	
OpenAI Chat GPT-4	Model imported from a local file store	unassigned	unassigned	VENDOR	Deployed Snapshots	

- **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

# How is AI Being Used in Your Organization?

## Monitors, Use Cases, and Results

The screenshot displays the 'Statement Similarity' feature in ModelOp Center. It shows a table with columns for 'similarity' and 'response'. The table contains several rows of data, each with a similarity score and a corresponding response text. The interface includes a sidebar with navigation options like 'Dashboard', 'My Work', 'Models', 'Overview', 'Inventory', 'Model Lifecycle', 'Compliance', 'Operations', and 'Notifications'.

similarity	response
0.7947	AI governance principles include transparency, accountability, fairness, privacy, security, human control, ethical use, public engagement, compliance with regulations, and continuous monitoring. These principles aim to ensure responsible and ethical AI development and deployment while safeguarding human rights and values.
0.8293	To ensure transparency in AI model decision-making, use explainable AI techniques, provide model documentation, use visualizations, maintain audit trails, select interpretable model architectures, and communicate limitations openly. These practices help make the model's behavior understandable and build trust with users and stakeholders.
0.9172	The potential risks of not having proper AI model governance include biased outcomes, lack of accountability, privacy breaches, unintended consequences, and loss of trust. It can also lead to unreliable decision-making, security vulnerabilities, regulatory non-compliance, reputational damage, and financial loss.
0.8821	AI bias can be mitigated through diverse data, fairness-aware algorithms, bias detection, human-in-the-loop, and explainable AI. Governance plays a vital role by setting guidelines, promoting fairness, maintaining assessments, involving public engagement, and ensuring regular reviews to address bias effectively.
0.8761	To establish an effective AI model governance framework, form a cross-functional team, prioritize transparency and bias mitigation, ensure compliance with regulations, and foster a culture of continuous improvement.
0.8581	AI model governance can address issues related to data privacy and security by implementing robust data handling

## 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

The screenshot displays the 'Consumer Credit Predictor' model evaluation page in ModelOp Center. It features a 'Model Score: 89% (27/14)' and two donut charts. The 'Transparency Controls' chart shows 0% (0/7) compliance, and the 'Model Controls' chart shows 29% (2/7) compliance. A list of requirements is shown on the right, including 'Business Justification', 'Scope', 'Limitations', 'Name', 'Required Documentation or Asset', and 'AI system requires a Risk Rating'. Below the charts is a 'Metrics' section with a line graph titled 'Performance Over Time - Classification - accuracy' showing accuracy fluctuating between 0.0 and 1.0 from 19/12/2023 to 15/03/2024.

How is AI Being Used in Your Organization?

# Prove That You Are Adhering to Regulations

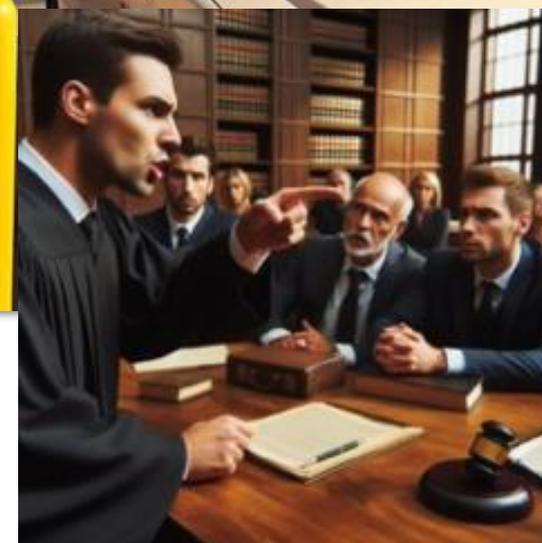
The screenshot displays the ModelOp Center interface. On the left is a navigation sidebar with categories: Dashboard, My Work, Models (Overview, Inventory), Model Lifecycles (Overview, Model Lifecycle List), Compliance (Overview, Reporting, Forms Configuration, Scores Configuration), and Operations (Overview, Jobs, Runtimes, Deployments, Notifications). The main area shows a detailed flowchart of the model lifecycle process, with various stages and decision points. Below the flowchart, there is a notification: "Model Document Review Notification: A use case has been created within ModelOp Center with the attached use case document. Please review, assign a risk in Jira, and mark the ticket as Done." Below the notification is a "Running Model Lifecycles" section with a "Refresh" button. It lists two model lifecycles: "Model Regulations Exploration - Implementation" and "Model Regulations Exploration". Each entry shows the model name, ID, start time (Nov 13, 2024, 10:41:08 AM), and running time (1 Hour 28 Minutes 36 Seconds). There are also links to "View MLC in snapshot" and "View MLC in stored model".

- 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 \$365,000 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



*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)



# Thank You

**Request a personalized demo of MVG for your enterprise**

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

[www.modelop.com](https://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.