



Guide

# How to Track Fairness and Bias In Predictive and Generative AI

Monitor Fairness Metrics for Predictive and Generative Models ..... 3

Monitoring Fairness in Predictive Models ..... 4

Monitoring Fairness in Generative AI Models ..... 6

Customized Dashboards with Fairness Reports ..... 7

Monitoring Fairness Metrics is Essential for responsible AI and GRC Compliance ..... 8

As predictive and generative AI (GenAI) models become more embedded in our daily applications and services, practicing responsible AI and implementing strong AI governance, risk, and compliance management (GRC) is crucial for oversight. A critical part of practicing responsible AI is ensuring fairness in both training data and model deployment so that all users, and organizations affected experience transparent, trustworthy, and equitable outcomes.

The Fiddler AI Observability platform enables enterprises to:



#### **Monitor for Bias and Fairness**

- Prevent models from favoring specific entities or users
- Reduce compliance challenges, legal risks, and reputational damage



#### **Track Intersectional Fairness**

- Ensure ML models align with business ethics
- Enable teams to identify and address fairness issues throughout the model development lifecycle



#### **Embrace Responsible AI**

- Analyze outcomes across intersections of various protected attributes
- Enable accurate model monitoring and performance comparison

# Monitor Fairness Metrics for Predictive and Generative Models

Fiddler empowers enterprises to track and visualize fairness and bias metrics for both predictive and generative AI models. In this blog, we'll explore key examples, including:

#### **For Predictive Models:**

- Defining intersection of various protected attributes using segments
- Creating industry-specific metrics to track fairness and bias using custom metrics

#### **For Generative Models:**

- Detect LLM responses that consist of racist or sexist content using the Fiddler Trust Service

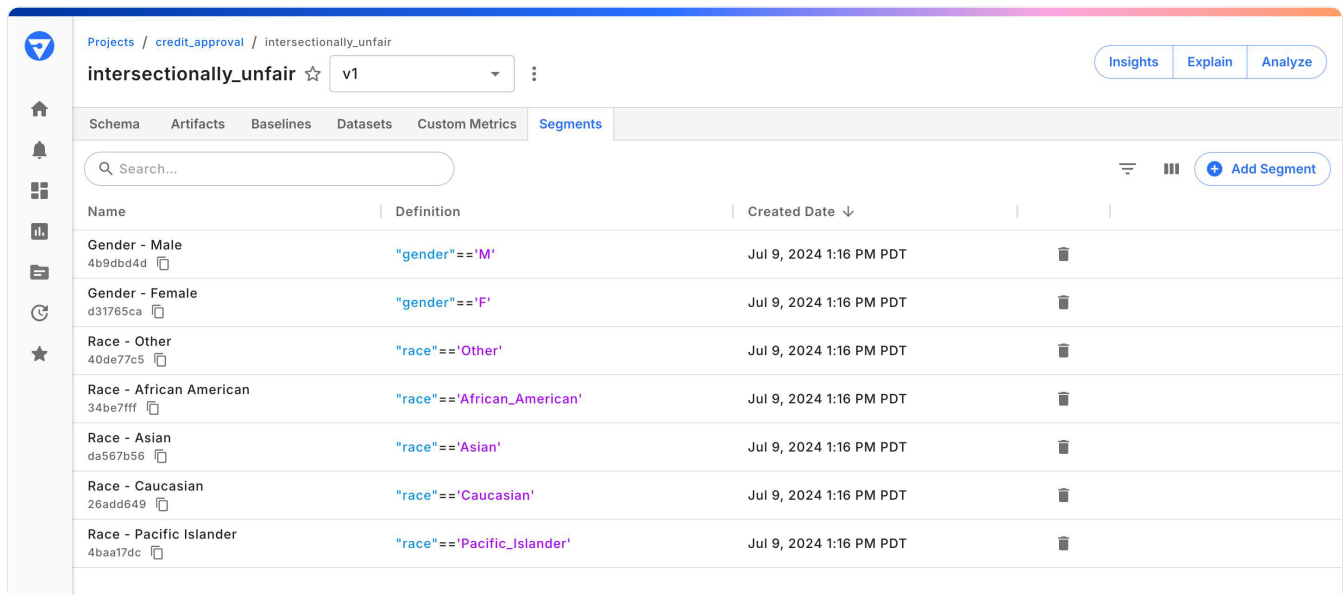
# Monitoring Fairness in Predictive Models

With Fiddler you can use model metadata and protected attributes from your datasets to track different aspects of fairness across the model lifecycle and on datasets the model is trained on.

The segments and metrics defined in the platform use the flexible Fiddler Query language (FQL) interface to allow your team to define Pythonic conditions and calculations via Fiddler's UI or API.

## Define Intersectionality Using Segments

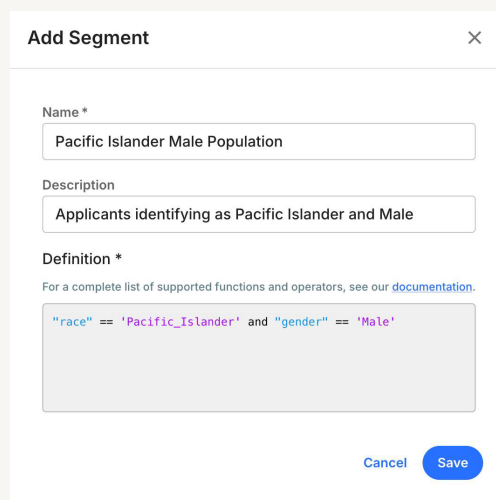
Segments let you define the intersection of several protected attributes like race, gender, sexual orientation, and any other attributes to ensure your models are fair and not biased towards any specific group.



The screenshot shows the Fiddler Segments interface. At the top, there's a breadcrumb trail: Projects / credit\_approval / intersectionally\_unfair. Below this, the segment name 'intersectionally\_unfair' is displayed with a star icon and a version dropdown set to 'v1'. On the right, there are buttons for 'Insights', 'Explain', and 'Analyze'. A navigation bar below the header includes tabs for Schema, Artifacts, Baselines, Datasets, Custom Metrics, and Segments (which is active). A search bar is present with the placeholder 'Search...'. On the right of the search bar are icons for a list, a menu, and an 'Add Segment' button. The main table lists segments with columns: Name, Definition, and Created Date. The segments listed are:

Name	Definition	Created Date
Gender - Male 4b9dbd4d	"gender" == 'M'	Jul 9, 2024 1:16 PM PDT
Gender - Female d31765ca	"gender" == 'F'	Jul 9, 2024 1:16 PM PDT
Race - Other 40de77c5	"race" == 'Other'	Jul 9, 2024 1:16 PM PDT
Race - African American 34be7fff	"race" == 'African_American'	Jul 9, 2024 1:16 PM PDT
Race - Asian da567b56	"race" == 'Asian'	Jul 9, 2024 1:16 PM PDT
Race - Caucasian 26add649	"race" == 'Caucasian'	Jul 9, 2024 1:16 PM PDT
Race - Pacific Islander 4baa17dc	"race" == 'Pacific_Islander'	Jul 9, 2024 1:16 PM PDT

Convert your model metadata into trackable identities



The 'Add Segment' dialog box is shown. It has a close button (X) in the top right corner. The form contains the following fields:

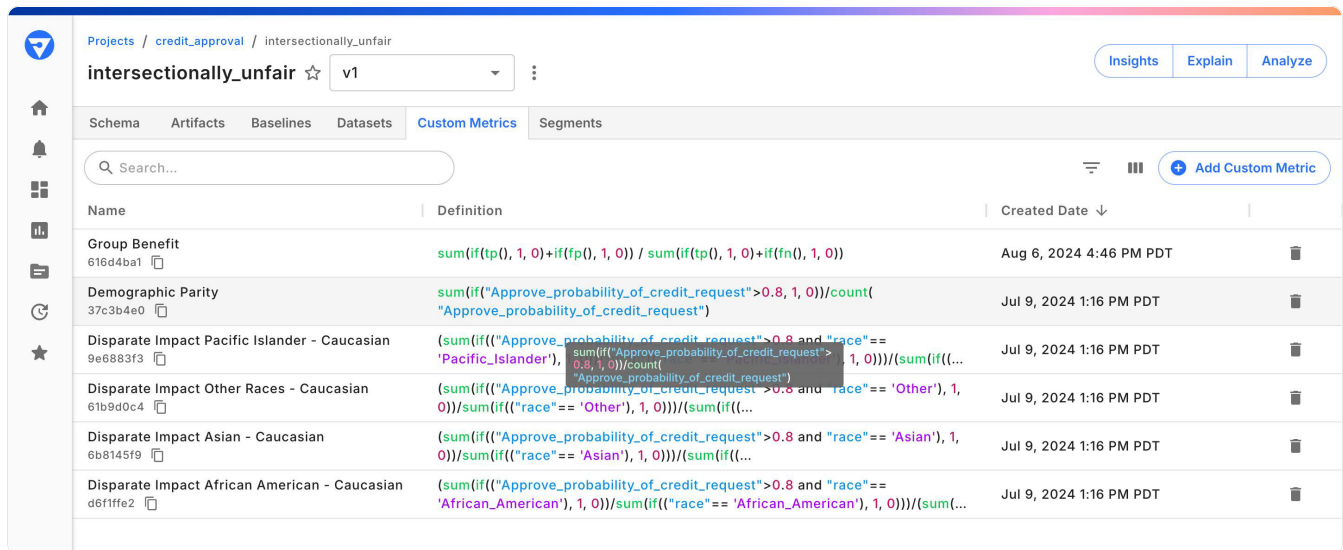
- Name \***: Pacific Islander Male Population
- Description**: Applicants identifying as Pacific Islander and Male
- Definition \***: "race" == 'Pacific\_Islander' and "gender" == 'Male'

Below the definition field, there is a link: "For a complete list of supported functions and operators, see our [documentation](#)." At the bottom right, there are 'Cancel' and 'Save' buttons.

Creating a new user segment that can be tracked

# Define Industry-Specific Fairness Metrics using Custom Metrics

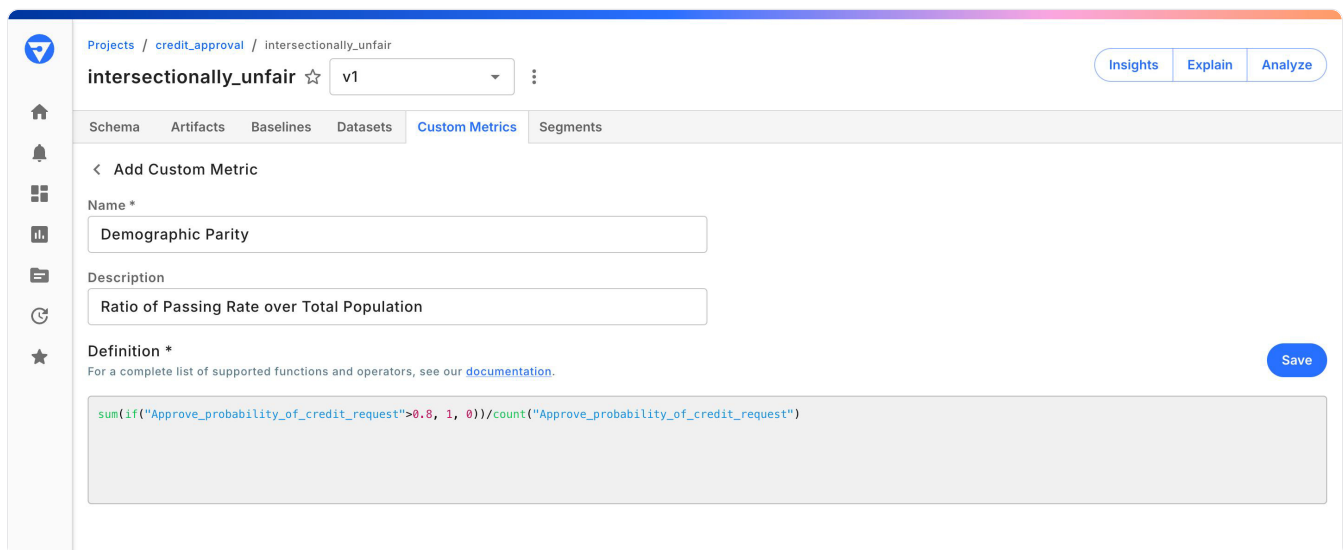
Custom metrics let you define metrics used in your industry to measure the fairness of outcomes for the decisions made by models used by your organization.



The screenshot shows the 'Custom Metrics' tab for a project named 'intersectionally\_unfair'. The table lists several metrics, including 'Group Benefit', 'Demographic Parity', and several 'Disparate Impact' metrics for different racial groups. Each row shows the metric name, its definition (a complex SQL-like expression), and the creation date.

Name	Definition	Created Date
Group Benefit 616d4ba1	<code>sum(if(tp(), 1, 0)+if(fp(), 1, 0)) / sum(if(tp(), 1, 0)+if(fn(), 1, 0))</code>	Aug 6, 2024 4:46 PM PDT
Demographic Parity 37c3b4e0	<code>sum(if("Approve_probability_of_credit_request"&gt;0.8, 1, 0))/count("Approve_probability_of_credit_request")</code>	Jul 9, 2024 1:16 PM PDT
Disparate Impact Pacific Islander - Caucasian 9e6883f3	<code>(sum(if(("Approve_probability_of_credit_request"&gt;0.8 and "race"== 'Pacific_Islander'), sum(if("Approve_probability_of_credit_request"&gt;0.8, 1, 0))/count("Approve_probability_of_credit_request"))/(sum(if(...</code>	Jul 9, 2024 1:16 PM PDT
Disparate Impact Other Races - Caucasian 61b9d0c4	<code>(sum(if(("Approve_probability_of_credit_request"&gt;0.8 and "race"== 'Other'), 1, 0))/sum(if(("Approve_probability_of_credit_request"&gt;0.8 and "race"== 'Other'), 1, 0))</code>	Jul 9, 2024 1:16 PM PDT
Disparate Impact Asian - Caucasian 6b8145f9	<code>(sum(if(("Approve_probability_of_credit_request"&gt;0.8 and "race"== 'Asian'), 1, 0))/sum(if(("Approve_probability_of_credit_request"&gt;0.8 and "race"== 'Asian'), 1, 0))</code>	Jul 9, 2024 1:16 PM PDT
Disparate Impact African American - Caucasian d6f1ffe2	<code>(sum(if(("Approve_probability_of_credit_request"&gt;0.8 and "race"== 'African_American'), 1, 0))/sum(if(("Approve_probability_of_credit_request"&gt;0.8 and "race"== 'African_American'), 1, 0))</code>	Jul 9, 2024 1:16 PM PDT

Define industry or use case-specific fairness metrics using Custom Metrics



The screenshot shows the 'Add Custom Metric' form. It has fields for 'Name', 'Description', and 'Definition'. The 'Name' field is filled with 'Demographic Parity', the 'Description' with 'Ratio of Passing Rate over Total Population', and the 'Definition' with a SQL-like expression. A 'Save' button is visible at the bottom right.

**Name \***  
Demographic Parity

**Description**  
Ratio of Passing Rate over Total Population

**Definition \***  
For a complete list of supported functions and operators, see our [documentation](#).

`sum(if("Approve_probability_of_credit_request">0.8, 1, 0))/count("Approve_probability_of_credit_request")`

**Save**

A Demographic Parity custom metric for a loan approval rates use case

# Monitoring Fairness in Generative AI Models

With the advent of LLMs, new challenges in maintaining fairness have surfaced. These models can unintentionally generate biased or harmful content, highlighting the need for proactive monitoring and prevention. Fiddler AI addresses this challenge with its Fiddler Trust Service, consisting of proprietary Fiddler Trust Models that are fast, scalable, secure, and cost-effective in monitoring.

Fiddler Trust Service helps enterprises with:



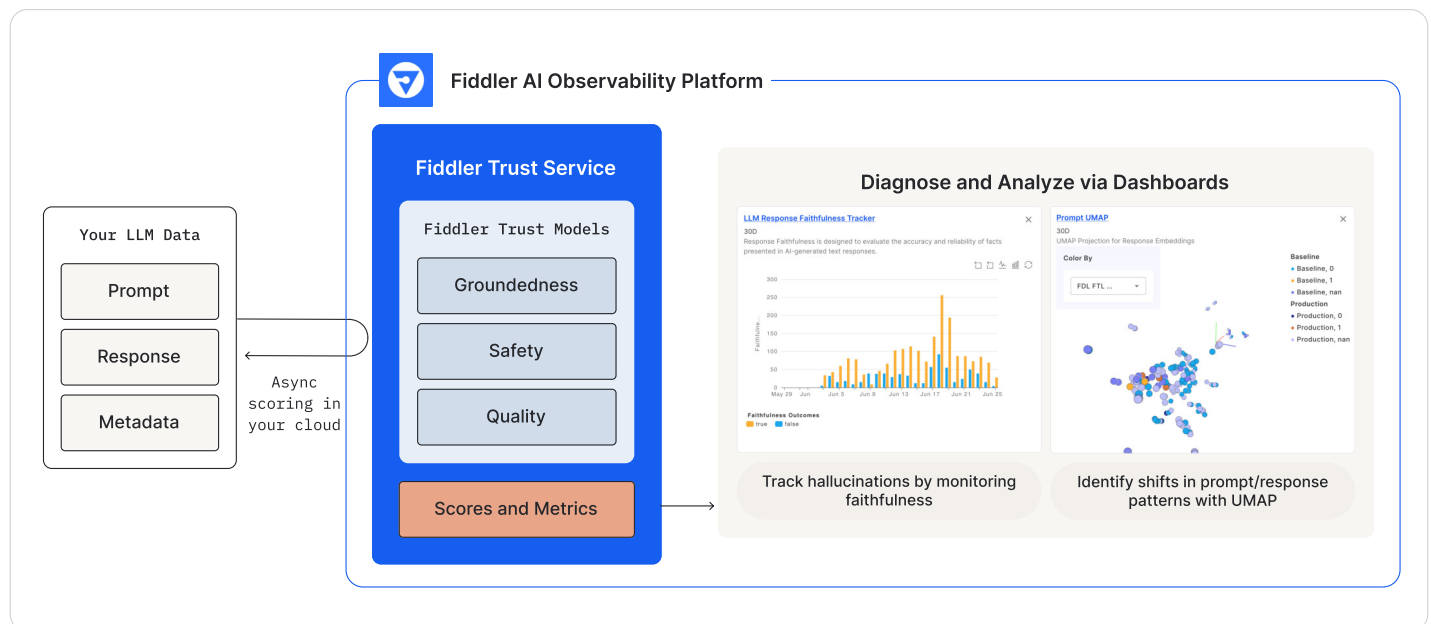
Fiddler Trust Models monitor LLM metrics by scoring trust-related dimensions in responses, user prompts, and context documents



Trust-related dimensions include: illegal, hateful, harassing, racist, sexist, violent, sexual, harmful, unethical, and jailbreaking content



Language is flagged for bias or unfair behavior and an alert is triggered for AI teams to investigate in Fiddler's 3D UMAP and Slice and Explain to address the incident



Similar to predictive models, custom metrics and segments can further stratify behavioral signals from safety models, allowing you to build targeted metrics and alerts that engage the appropriate teams effectively.

For example, the segment below tracks LLM responses containing biased language that could harm users or the business and generates alerts to notify a moderator when such instances are detected.

Add Segment

Name \*

Biased Responses

Description

LLM responses using biased language

Definition \*

For a complete list of supported functions and operators, see our [documentation](#).

"FDL FTL Safety (response) racist" == true or "FDL FTL Safety (response) sexist" == true

Cancel

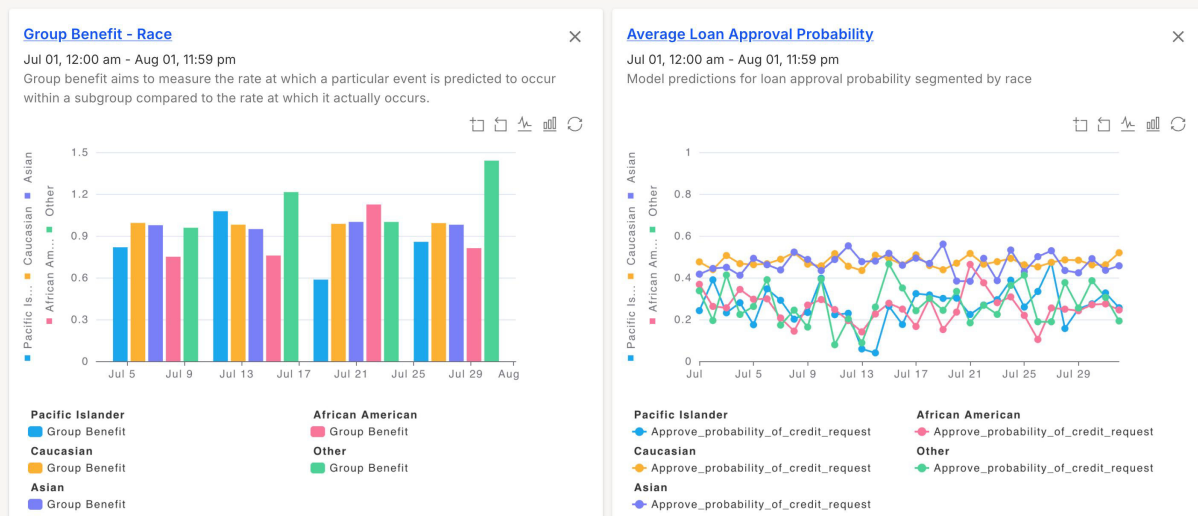
Save

Create segments to detect AI bias in LLM responses

# Customized Dashboards with Fairness Reports

Start tracking model behavior by incorporating fairness reports into dashboards customized for Trust and Safety, and Risk and Compliance teams. These dashboards and reports can also be shared for third party reviews and for GRC purposes.

When any of these metrics fall outside the enterprise's accepted thresholds, alerts notify the model development, engineering, or risk and compliance teams to intervene and identify the root cause of unfair model outcomes. This proactive approach helps address issues and tune the model to reduce the likelihood of bias in the future and remain compliant with GRC standards.



Comprehensive LLM and ML model fairness tracking dashboards



Tracking sexism and racism in LLM prompts and responses

# Monitoring Fairness Metrics is Essential for Responsible AI and GRC Compliance

Gaining oversight on AI bias and fairness metrics has become easier for enterprises deploying ML and GenAI models, thanks to the Fiddler AI Observability platform. By leveraging custom segments, custom metrics, and the Fiddler Trust Service for LLM monitoring, AI teams can proactively detect unfair outcomes, reduce AI risks, and remain compliant with GRC standards. Enterprises can avoid AI disasters caused by bias and unfair outcomes, ensuring their AI implementations are equitable and fair for all.

Connect with our Fiddler AI experts to learn how to integrate fairness metrics into your responsible AI framework and ensure GRC compliance.





Fiddler is a pioneer in AI Observability for responsible AI. The unified environment provides a common language, centralized controls, and actionable insights to operationalize ML models and GenAI and LLM applications with trust. Monitoring, analytics, and explainable AI capabilities address the unique challenges of building in-house stable and secure LLM and MLOps at scale. Fiddler helps you grow into advanced capabilities over time and build a framework for responsible AI practices.

Fortune 500 organizations use Fiddler across pre-production and production to deliver high performance AI, reduce costs, and be responsible in governance.

 [fiddler.ai](https://fiddler.ai)     [sales@fiddler.ai](mailto:sales@fiddler.ai)