Getting to Know AI Defense UI

The AI Defense console offers robust navigation, providing you with a quick snapshot of the current status of your applications, connections, and policies.

Note This dashboard is displayed after the initial onboarding.

Overview

Overview provides an at-a-glance view of key data, metrics, and features allowing you to monitor performance, track progress, and navigate to AI Defense features. The Overview page is divided into three sections:

- **Getting Started with AI Defense**: You can add applications, connections and create policies to protect your organization. Click on **Help** to leverage our user documentation.
- **Policy Analytics**: A visual chart provides a comprehensive overview of:
  - **Detected Prompts**: This shows the total number of prompts submitted to the AI model within the selected time period. It helps identify trends in user queries and potential risks posed by malicious or adversarial prompts.
  - **Detected Responses**: This tracks the responses generated by the AI model in response to the detected prompts. Monitoring responses helps ensure that the AI model in use is generating safe, compliant, and accurate output.
  - **Connectivity Status**: This indicates the current connection status between your system and AI Defense, allowing you to verify whether traffic is being monitored in real-time.
  - **Policy Action Summary:** The actions taken in response to detected rule violations or potential threats. This could include blocking harmful content, generating alerts, or allowing safe interactions to proceed.
  - **Guardrails Distribution**: This section provides a visual representation of how guardrails (security rules and policies) are applied across different applications and use cases. It helps you understand the extent of protective measures in place.
  - **Top Rule Matches**: Displays the triggered security rules within the defined time frame. This allows you to identify any recurring issues or common policy violations that need attention.

- **Event**: Provides details of all events, including the matched rule, application, timestamp, user ID, rule action, message type, and model. You can customize the displayed attributes in this tabular view by clicking on the settings icon in the top-right corner and selecting the desired attributes.

Note You can filter this data by time period and application.

Events

The Events page provides details of ongoing activities and interactions within your AI Defense environment. Key features include:

- **Event Logs**: Captures all events related to AI activity, including detected prompts, responses, and rule matches.
- **Filtering Options**: You can filter data by time period, application, or event type for more focused monitoring.

Validation

AI Validation systematically assess the performance, reliability, compliance, and safety of your AI systems. This ensures that AI models and applications meet your specific organizational, ethical, and regulatory requirements before and during deployment. You can run validation test. You can view the past and in progress validation tests.

AI App Discovery

AI Defense integrates with **Secure Access** to enhance the detection and management of third-party generative AI applications within an organization. Secure Access works as a monitoring and enforcement layer, providing visibility into the use of external AI tools and applications that interact with internal systems or data. It helps identify and assess potential risks associated with these applications, as well as ensure they comply with the organization's security and policy frameworks. You can filter the results by **Risk**.

AI Assets

AI Defense integrates with **Multicloud Defense** to extend security capabilities across multiple cloud environments, ensuring comprehensive protection for AI models and cloud-based infrastructure. This integration helps organizations maintain visibility into their cloud environments and manage the risks associated with generative AI models, regardless of whether they are deployed on-premises or across various cloud providers (e.g., AWS, Azure, GCP).

Cloud visibility tab: Shows a count of the models, agents, and knowledge bases discovered.

External assets: Detects traffic that goes to third-party models hosted outside your AWS cloud. For these, no validation is run, but AI Defense can show details about the network traffic to the model.

Policies

The Policies page allows you to manage and configure AI Defense policies. Key elements include:

- **Policy Creation and Management**: You can create, update, or delete policies governing how generative AI systems should behave, with options to define rule sets for data usage, model access, and compliance protocols.

- **Guardrails Configuration**: Establishes boundaries for AI actions, ensuring ethical use and compliance with regulations. These guardrails are defined as a part of a policy. AI Defense has security, privacy and safety guardrails:
  - Security guardrails: Cybersecurity and Model Vulnerabilities: Involve exploiting system weaknesses or assisting in cybersecurity attacks to compromise the security, integrity, or functionality of models and systems.
  - Privacy guardrails: PII (Personally Identifiable Information) and Intellectual Property Theft: Theft of sensitive personal information and intellectual property, including private data and protected organizational assets, with the intent to cause economic harm or competitive disadvantage to the victim organization.
  - Safety guardrails: Financial harm entails the loss of monetary assets through theft, fraud, or other malicious actions, while reputational harm involves a decline in public trust due to scandals or false impersonations.

Applications

The Applications page provides a detailed view of all AI applications being monitored and secured within your organization. From this page, you can easily manage applications, track their security status, and review connection details. The page is divided into two main tabs:

- **Applications Tab**: Shows a list of all applications along with their associated connection details and the current status of those connections.
  - **Application List**: Displays all active and inactive AI applications within the organization. This is also where you can add new applications to the system.
- **Connections Tab**: Displays connection-specific information, including connection status, details of the connection, and a connection guide to help you set up and troubleshoot connections.
  - **Endpoint**: Displays the endpoint for the connection.
  - **Models**: Displays the AI models used in the connection.
  - **Policy**: Lists the policies that have been triggered within a specific application, helping you monitor potential threats and violations.
  - **Last active**: Timestamp of the last activity detected by the connected proxy.
  - **Filtering and Sorting**: Allows you to filter applications by security status, application type, or other customizable parameters, ensuring you can quickly find relevant information.
  - **Connection guide**: The guide provides a cURL command to update the proxy URL, ensuring that all AI interactions are effectively monitored.

Administration

In AI Defense, you can input critical details regarding their **AI-powered applications** and the **AI model endpoints** they are utilizing. This information is essential for ensuring that AI Defense can effectively monitor, manage, and secure AI interactions across the infrastructure.