AI Defense Policies

Create policies using built-in guardrails and templates to ensure data protection, legal compliance, and responsible AI use. These policies are applied to your connected endpoints, helping to secure and govern their interactions.

To create a new policy, click **Create Policy** on the right top corner of the Policies page. Select from Gateway and API.

To create a new policy:

Note A left slide-in pane is displayed. This pane has a step-by-step guide for creating a new policy.

Procedure

Step 1 On the **New Policy** page, under **Policy Details**, enter a name for the policy.

The **Description** is optional, you can add details if required.

a) Click **Next**.

Step 2 A list of connections is displayed, you can select the checkbox for one or multiple connections to apply the policy. Click **Next**.

Step 3 Under **Security guardrails**(prompt injection and code detection), configure security rules.

a) Move the slider to **Enabled** for the rule.
b) You can click the dropdown for **Rule directionality** to apply this rule to prompts, responses or both.
c) Click the dropdown for **Action** to select the action this policy rule will take. You can select: Block or Allow.
d) Click **Next**.

Step 4 To protect data and maintain confidentiality, under **Privacy guardrails**, there are three rules:

a) Protected health information (PHI):

1. Move the slider to **Enabled**.
2. For **Rule directionality**, click the dropdown for **Directions** to select Prompts, Responses or Both.

3. Click the dropdown for **Action** to select the action this policy rule will take. You can select: Block or Allow.

b) PII (Personally Identifiable Information):

1. Move the slider to **Enabled**.
2. Slide to **Enable** the one or multiple PII entity.
3. Click the dropdown for **Directions** to select Prompts, Responses or Both.
4. Click the dropdown for **Action** to select the action this policy rule will take. You can select : Block or Allow.

c) Payment Card Industry (PCI):

1. Move the slider to **Enabled**.
2. Choose the entities for which you want to apply the subcategories and their direction and rule action.
3. For **Rule directionality**, click the dropdown for **Directions** to select Prompts, Responses or Both.
4. Click the dropdown for **Action** to select the action this policy rule will take. You can select: Block or Allow.

d) Click **Next**.

Step 5 Under **Safety guardrails**, configure security rules.

a) Move the slider to **Enabled** for the rule.
b) For **Rule directionality**, click the dropdown for **Directions** to select Prompts, Responses or Both.
c) Click the dropdown for **Action** to select the action this policy rule will take. You can select: Block or Allow.
d) Click **Next**.

Step 6 On the **Summary** page, review your policy details. Click **Save** to create the policy.

The policies are **disabled** by default, you would need to enable the policies.

The policy is created and displayed under Policy.