



Event ends in 2 days 15 hours 35 minutes.

[Event dashboard](#) > [Guardrails for Amazon Bedrock](#) > [Create a guardrail](#) >

Configure Content Filters

Configure Content Filters

Guardrails support content filters to detect and filter both harmful **user inputs** and **FM-generated outputs**. Content filtering depends on the confidence classification of user inputs and FM responses across each of the six harmful categories. All input and output statements are classified into one of four confidence levels (NONE, LOW, MEDIUM, HIGH) for each harmful category.

On the Configure content filters page, set up how strongly you want to filter out content related to the categories defined in Content filters by doing the following:

1. To configure filters for prompts to a model, select **Enable filters for prompts** in the Filter strengths for model prompts section. Configure how strict you want each filter to be for prompts that the user provides to the model. For this workshop, we will just accept the defaults.

Filter strengths for prompts

[Reset](#)

Use a higher filter strength to increase the likelihood of filtering harmful content in a given category.

☒ Enable filters for prompts



2. To configure filters for model responses, select **Enable filters for responses** in Filter strengths for responses. Configure how strict you want each filter to be for responses that the model returns.
3. Choose Next.



Filter strengths for responses

Reset

Use a higher filter strength to increase the likelihood of filtering harmful content in a given category. These filters evaluate and override model responses, but don't modify the model behavior.

☒ Enable filters for responses

Hate	<div><div></div><div></div><div></div><div></div></div> <div>NoneLowMediumHigh</div>
Insults	<div><div></div><div></div><div></div><div></div></div> <div>NoneLowMediumHigh</div>
Sexual	<div><div></div><div></div><div></div><div></div></div> <div>NoneLowMediumHigh</div>
Violence	<div><div></div><div></div><div></div><div></div></div> <div>NoneLowMediumHigh</div>
Misconduct	<div><div></div><div></div><div></div><div></div></div> <div>NoneLowMediumHigh</div>

Cancel

Skip to Review and create

Previous

Next

Previous

Next