Event ends in 2 days 15 hours 31 minutes.

Event dashboard > Guardrails for Amazon Bedrock > **Test the guardrail**

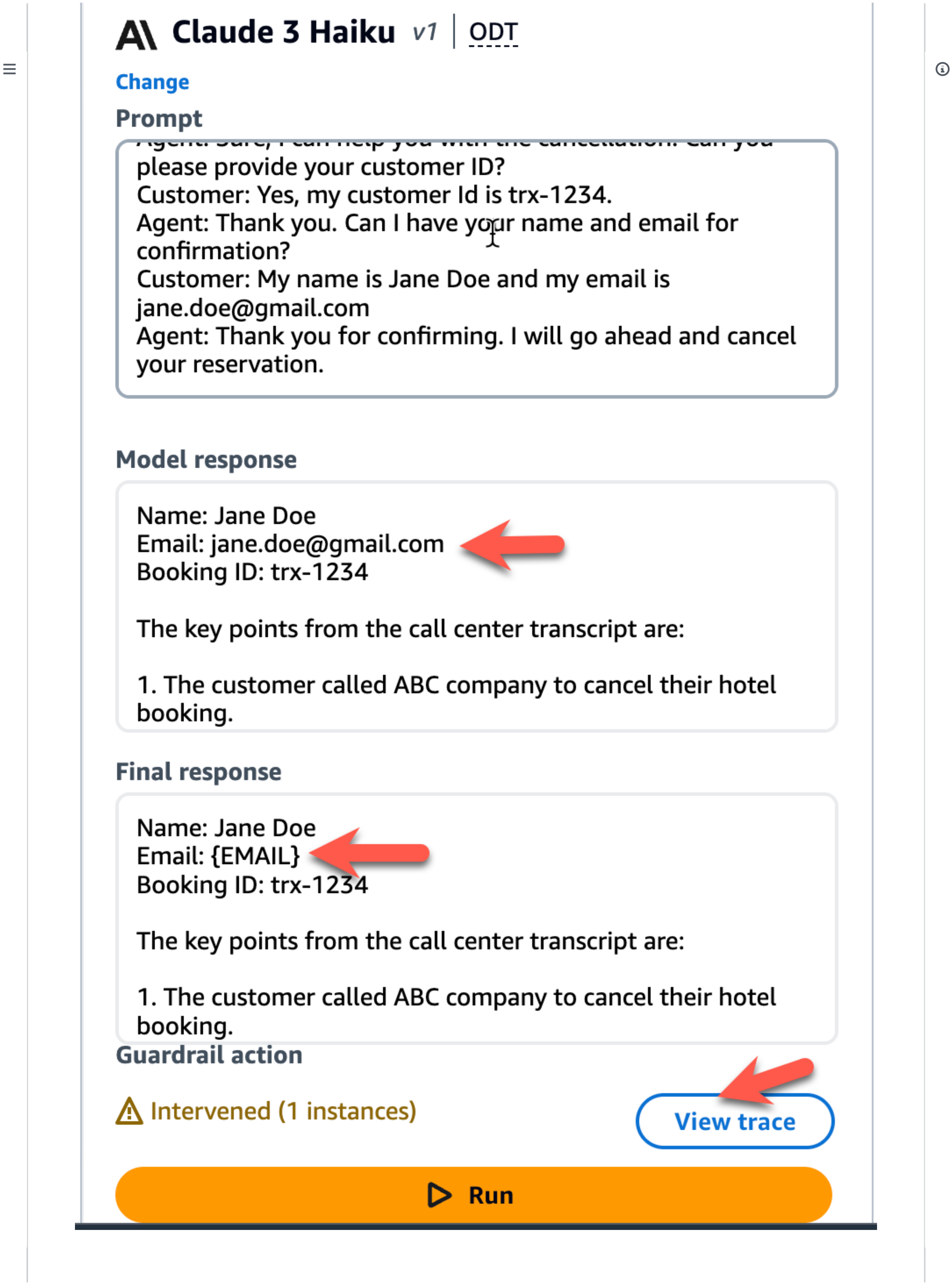
Test the guardrail

To test a guardrail

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at https://console.aws.amazon.com/bedrock/.
2. Choose Guardrails from the left navigation pane. Then, select the guardrail you created from the Guardrails section.
3. A test window appears on the right. Select the working draft of the guardrail to test. To select a model, choose Select model. After you make a choice, select Apply. To change the model, choose Change. We recommend selecting `Anthropic > Claude 3 Haiku` for this test.
4. Copy the following text and paste under the **Prompt** textbox of the test window. Then click **Run** button

```
Please summarize the below call center transcript. Put the name, email and the customer ID to the top:

Agent: Welcome to ABC company. How can I help you today?
Customer: I want to cancel my hotel booking.
Agent: Sure, I can help you with the cancellation. Can you please provide your customer ID?
Customer: Yes, my customer Id is trx-1234.
Agent: Thank you. Can I have your name and email for confirmation?
Customer: My name is Jane Doe and my email is jane.doe@gmail.com
Agent: Thank you for confirming. I will go ahead and cancel your reservation.
```



Prompt

Agent: Sure, I can help you with the cancellation. Can you please provide your customer ID?
Customer: Yes, my customer Id is trx-1234.
Agent: Thank you. Can I have your name and email for confirmation?
Customer: My name is Jane Doe and my email is jane.doe@gmail.com
Agent: Thank you for confirming. I will go ahead and cancel your reservation.

Model response

Name: Jane Doe
Email: jane.doe@gmail.com
Booking ID: trx-1234

The key points from the call center transcript are:

1. The customer called ABC company to cancel their hotel booking.

Final response

Name: Jane Doe
Email: {EMAIL}
Booking ID: trx-1234

The key points from the call center transcript are:

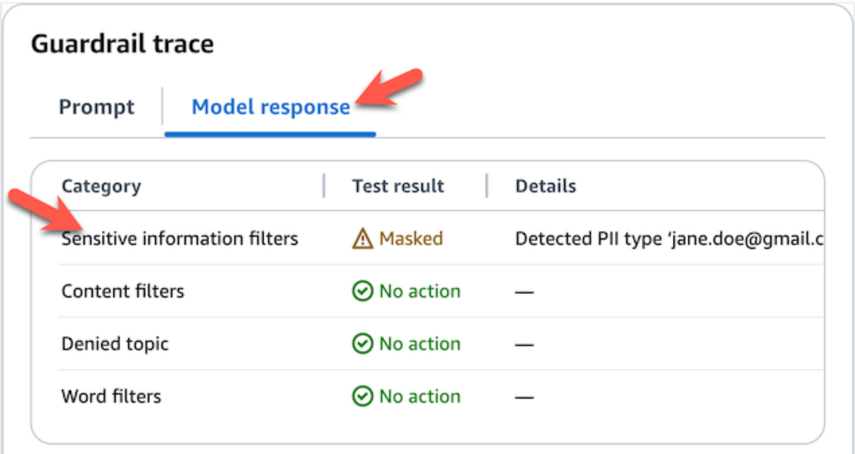
1. The customer called ABC company to cancel their hotel booking.

Guardrail action

⚠ Intervened (1 instances)

Notice the input prompt has an email in it. The FM responds with a summary that contains that email adress. However when the final response is generated, the guardrail kicks in and masks the email with **{EMAIL}**.

To view the topics or harmful categories in the prompt or response that were recognized and allowed past the filter or blocked by it, select **View trace**. In this case, it shows that the final response was `Masked` because of the presence of an email.



Guardrail trace

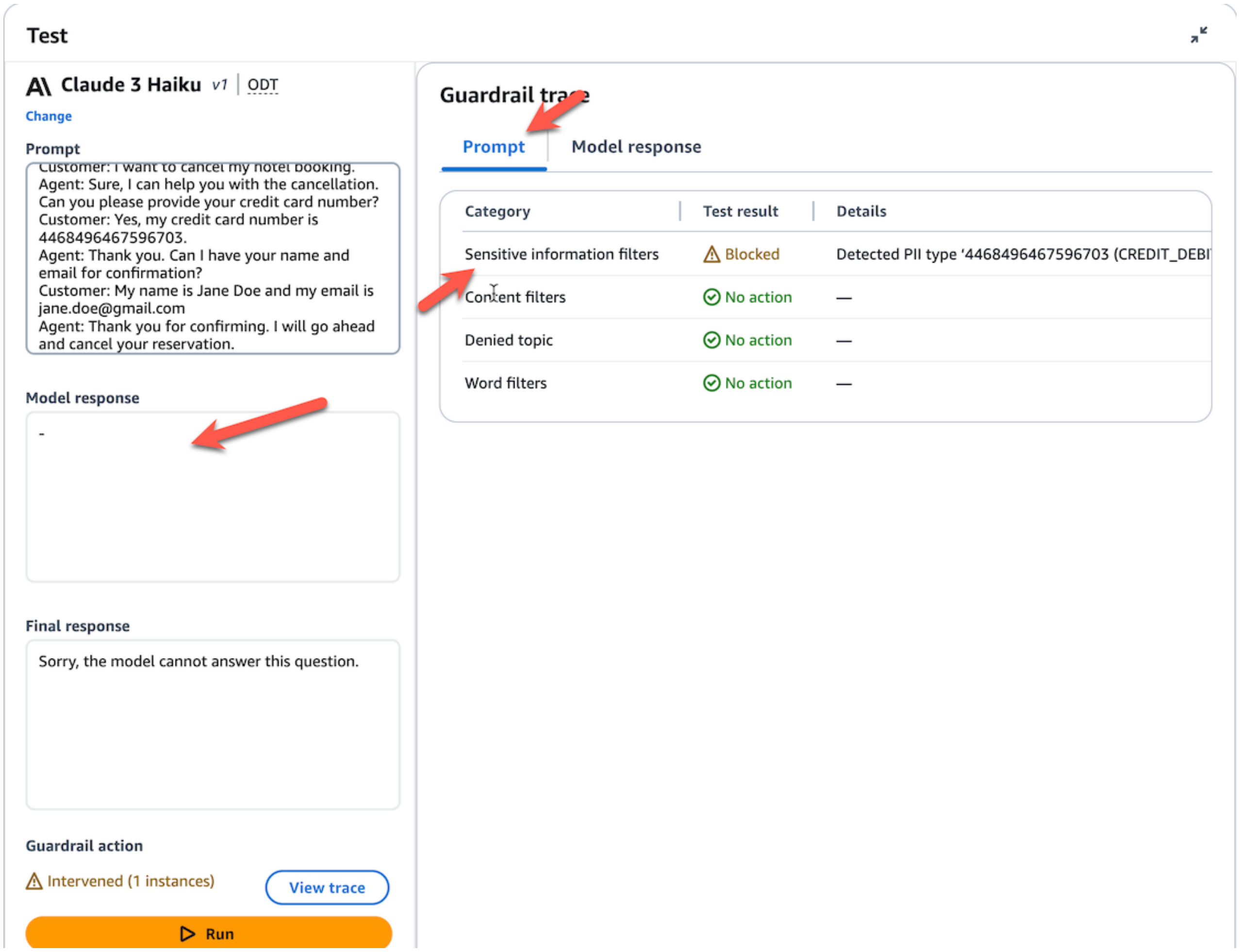
Category	Test result	Details
Sensitive information filters	⚠ Masked	Detected PII type 'jane.doe@gmail.c
Content filters	✓ No action	—
Denied topic	✓ No action	—
Word filters	✓ No action	—

5. Try the same process with the following prompt. Notice that customer Id has been replaced by a credit card number.

```
Please summarize the below call center transcript. Put the name, email and the credit card number to the top:

Agent: Welcome to ABC company. How can I help you today?
Customer: I want to cancel my hotel booking.
Agent: Sure, I can help you with the cancellation. Can you please provide your credit card number?
Customer: Yes, my credit card number is 4468496467596703.
Agent: Thank you. Can I have your name and email for confirmation?
Customer: My name is Jane Doe and my email is jane.doe@gmail.com
Agent: Thank you for confirming. I will go ahead and cancel your reservation.
```

The user prompt itself is blocked and is not sent to the foundational model for response generation.



Test

Claude 3 Haiku v1 | ODT

Change

Prompt

Customer: I want to cancel my hotel booking.
Agent: Sure, I can help you with the cancellation. Can you please provide your credit card number?
Customer: Yes, my credit card number is 4468496467596703.
Agent: Thank you. Can I have your name and email for confirmation?
Customer: My name is Jane Doe and my email is jane.doe@gmail.com
Agent: Thank you for confirming. I will go ahead and cancel your reservation.

Model response

-

Final response

Sorry, the model cannot answer this question.

Guardrail action

⚠ Intervened (1 instances)

Guardrail trace

Category	Test result	Details
Sensitive information filters	⚠ Blocked	Detected PII type '4468496467596703 (CREDIT_DEBI
Content filters	✓ No action	—
Denied topic	✓ No action	—
Word filters	✓ No action	—

6. Following the steps above, repeat this test by changin the user prompt to hit the different conditions you had configured the guardrail for and see how the final response generation is impacted. For example try asking for investment advice using Bitcoins.

Previous Next