- 16 -

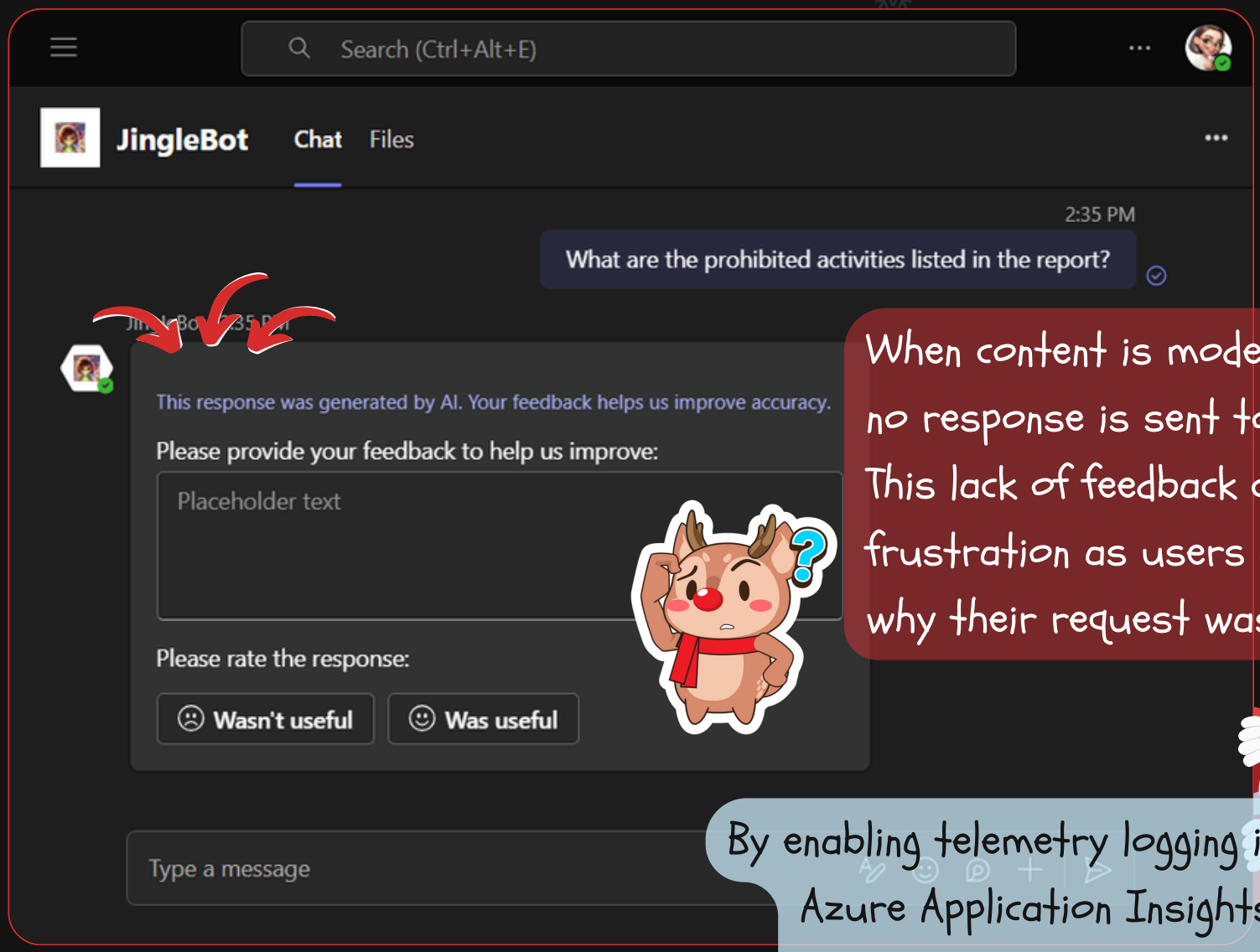
Catch responses
blocked by Content Filter

MONTHLY MASTERY

FEATURE-A-DAY

with Copilot Studio

Sprinkled with holiday magic by
Katerina Chernevskaya



When content is moderated, no response is sent to the user. This lack of feedback can cause frustration as users don't know why their request wasn't answered.

By enabling telemetry logging in Azure Application Insights, you can detect and analyze blocked responses. This helps identify problematic content and fine-tune your system.

Sprinkled with holiday magic by
Katerina Chernevskaya

Content moderation

What content gets blocked?

Copilot Studio moderates content to ensure responses are safe and compliant.

The types of content that get blocked include:

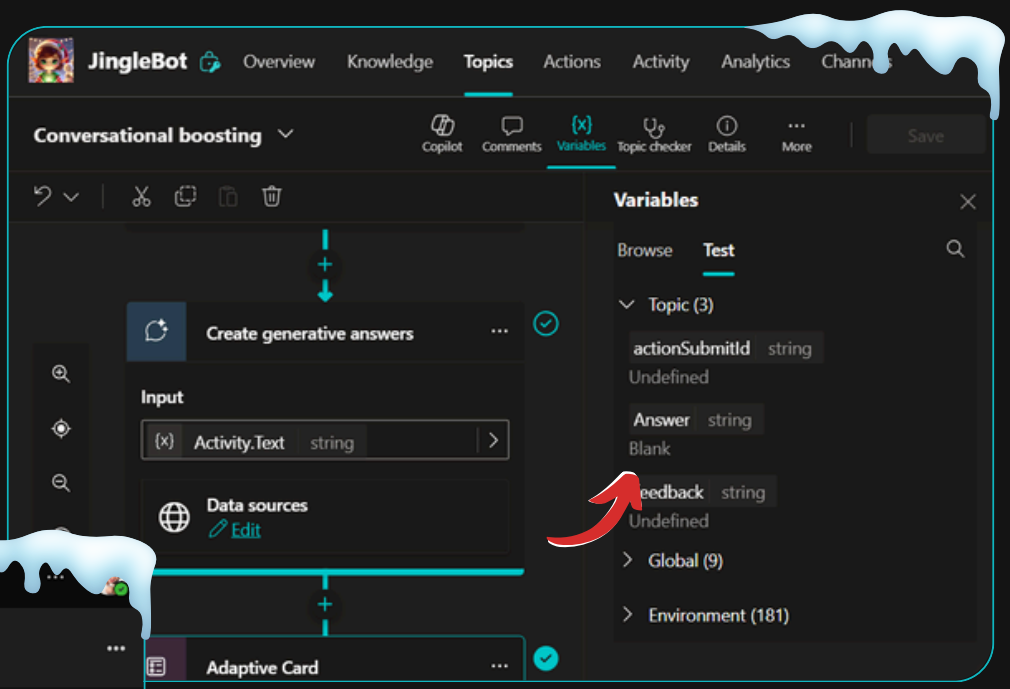
- Harmful or malicious text
- Noncompliant language or data
- Content that breaches copyrights or violates policies

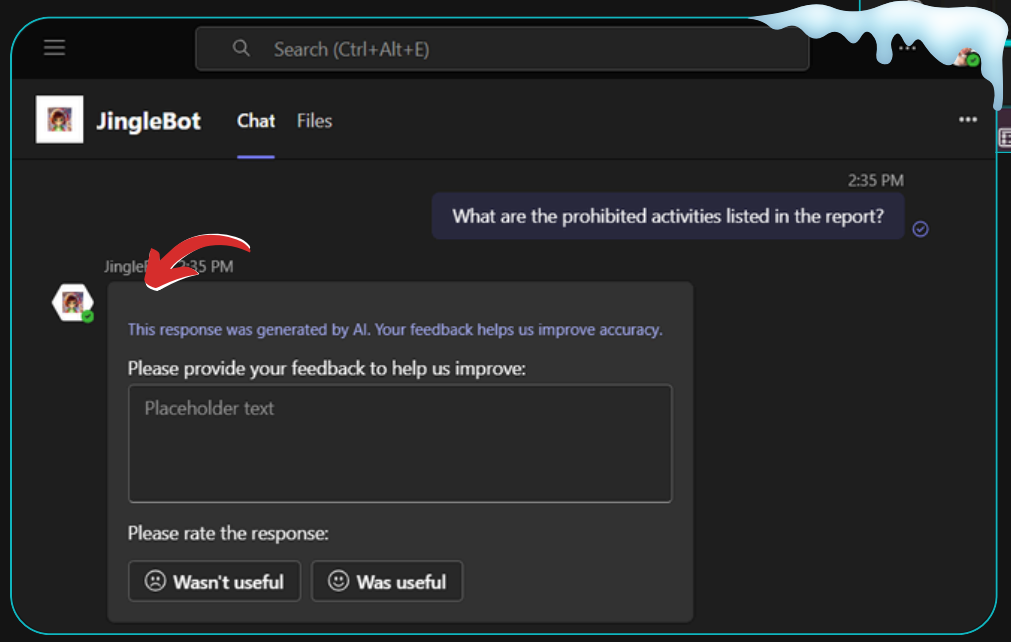
Sprinkled with holiday magic by
Katerina Chernevskaya

Content moderation

When content gets moderated:

The generative answer doesn't return a response.





Users don't get any indication that content was blocked.

Sprinkled with holiday magic by
Katerina Chernevskaya

Content moderation
Filter blocked responses

To catch responses blocked by the content filter, you can query Azure Application Insights data using Kusto Query Language (KQL).

Navigate to your Azure Application Insights instance and open the Logs tab to access Kusto Query editor.

Sprinkled with holiday magic by
Katerina Chernevskaya

Content moderation
Filter blocked responses

Use the following KQL query to filter moderation events:

```
customEvents
| extend cd = todynamic(customDimensions)
| extend message = tostring(cd.Message)
| extend result = tostring(cd.Result)
| where name == "GenerativeAnswers" and result contains "Filtered"
| project cloud_RoleInstance, timestamp, message, result
| order by timestamp desc
```

Adjust the filters to include specific keywords
or add further fields for analysis.

Key Components:

- customEvents: captures events for telemetry data.
- extend: creates new columns for Message and Result.
- where: filters events where responses are flagged as "Filtered".
- project: displays relevant fields like timestamp, message, and result.
- order by: sorts results by the latest events.

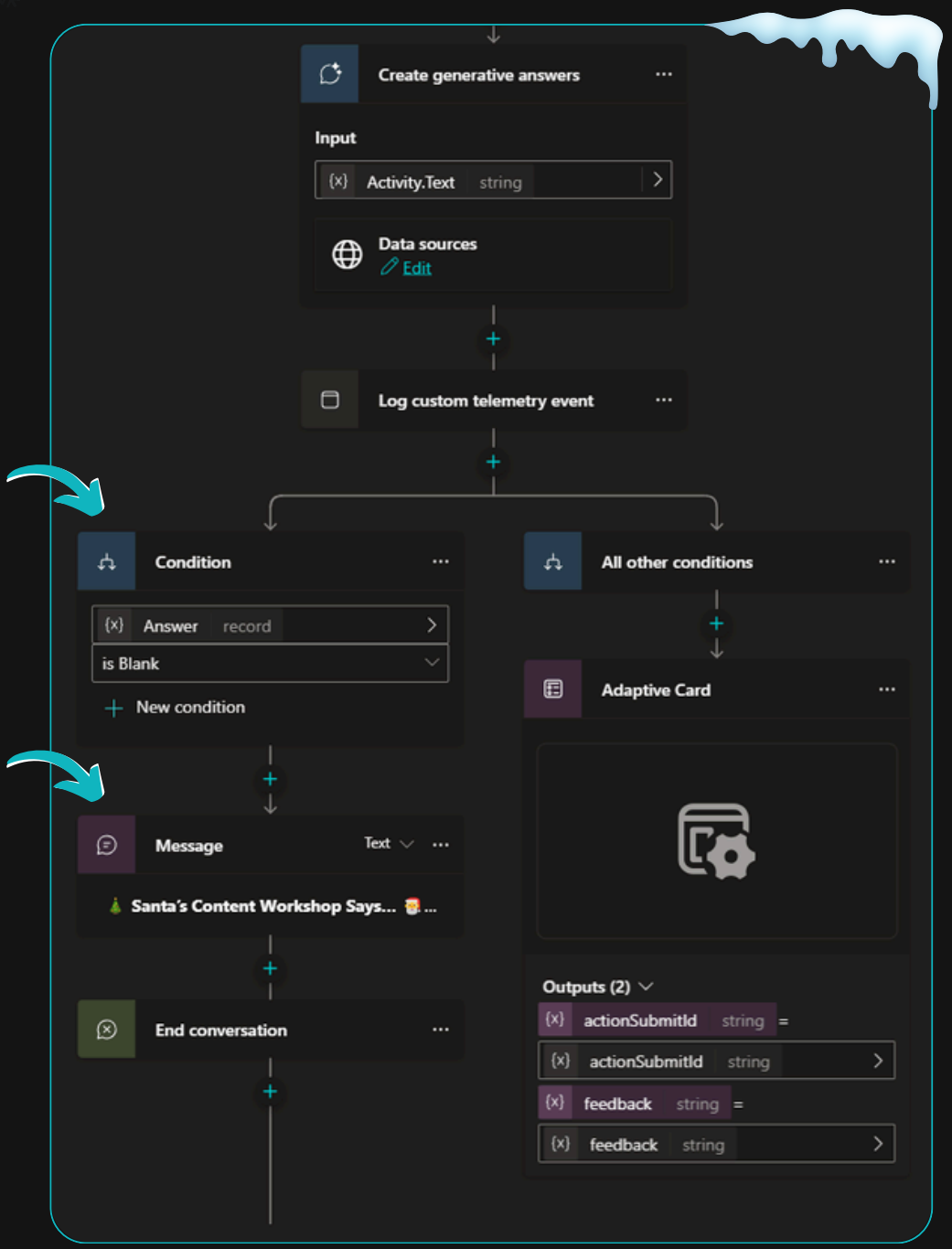
Sprinkled with holiday magic by
Katerina Chernevskaya

Content moderation
Inform users

To ensure a smooth and clear user experience, add a handler to the topic to manage situations where the content filter blocks a response.

Since the response will be blank when blocked, you can include a condition to check if the response is empty.

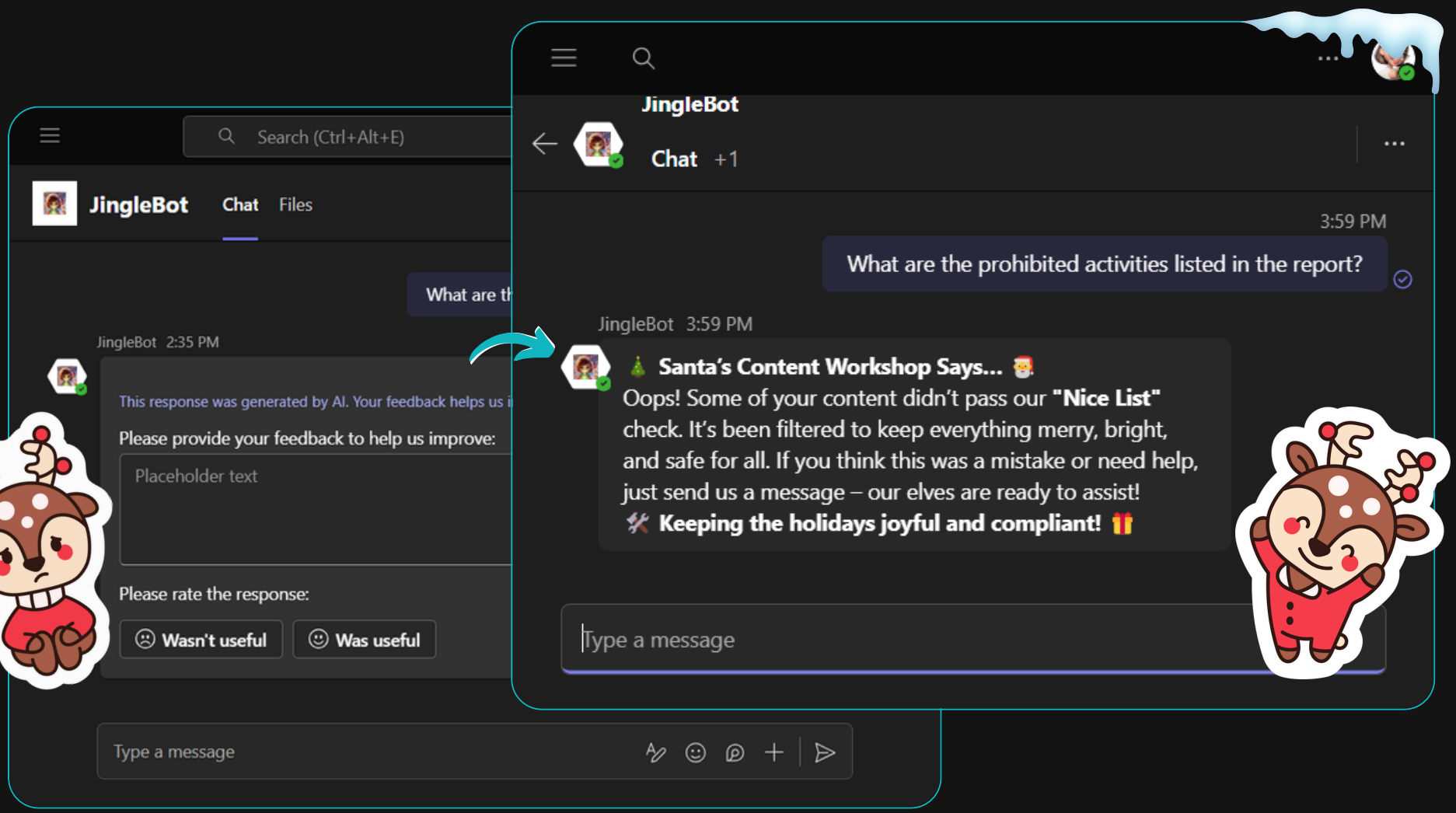
If it is, provide the user with an appropriate explanation.



Sprinkled with holiday magic by
Katerina Chernevskaya

Content moderation
Inform users

This handler ensures users receive appropriate responses
and remain informed about the situation,
providing a seamless and transparent experience.



Sprinkled with holiday magic by
Katerina Chernevskaya

Content moderation Benefits

Identify issues

Detect content that triggers filters and blocks AI responses.

Refine data sources

Improve your content sources to reduce moderation events.

Ensure transparency

Keep track of blocked responses to analyze patterns.

Optimize AI behavior

Enhance your agent's response quality and compliance.

Mitigate risks

Avoid repeated violations of compliance and policies.

Sprinkled with holiday magic by
Katerina Chernevskaya

Today's Task:
Catch blocked responses

1. Prepare test data

Create dummy content that includes sensitive or policy-violating text to trigger content moderation.

2. Run a test in Copilot Studio

Ask your agent a question that generates the blocked response.
Ensure telemetry logging is enabled in Azure Application Insights.

3. Analyze moderation events

Go to Azure Application Insights and use the Kusto Query provided. Find the blocked content in the logs and review its details.

Sprinkled with holiday magic by
Katerina Chernevskaya

MONTHLY MASTERY

FEATURE-A-DAY

with Copilot Studio

Sprinkled with holiday magic by
Katerina Chernevskaya

Follow for more!