Apply content filters to prevent the output of harmful content

Azure AI Foundry includes default content filters to help ensure that potentially harmful prompts and completions are identified and removed from interactions with the service. Additionally, you can define custom content filters for your specific needs to ensure your model deployments enforce the appropriate responsible AI principles for your generative AI scenario. Content filtering is one element of an effective approach to responsible AI when working with generative AI models.

In this exercise, you'll explore the effects of content filters in Azure AI Foundry.

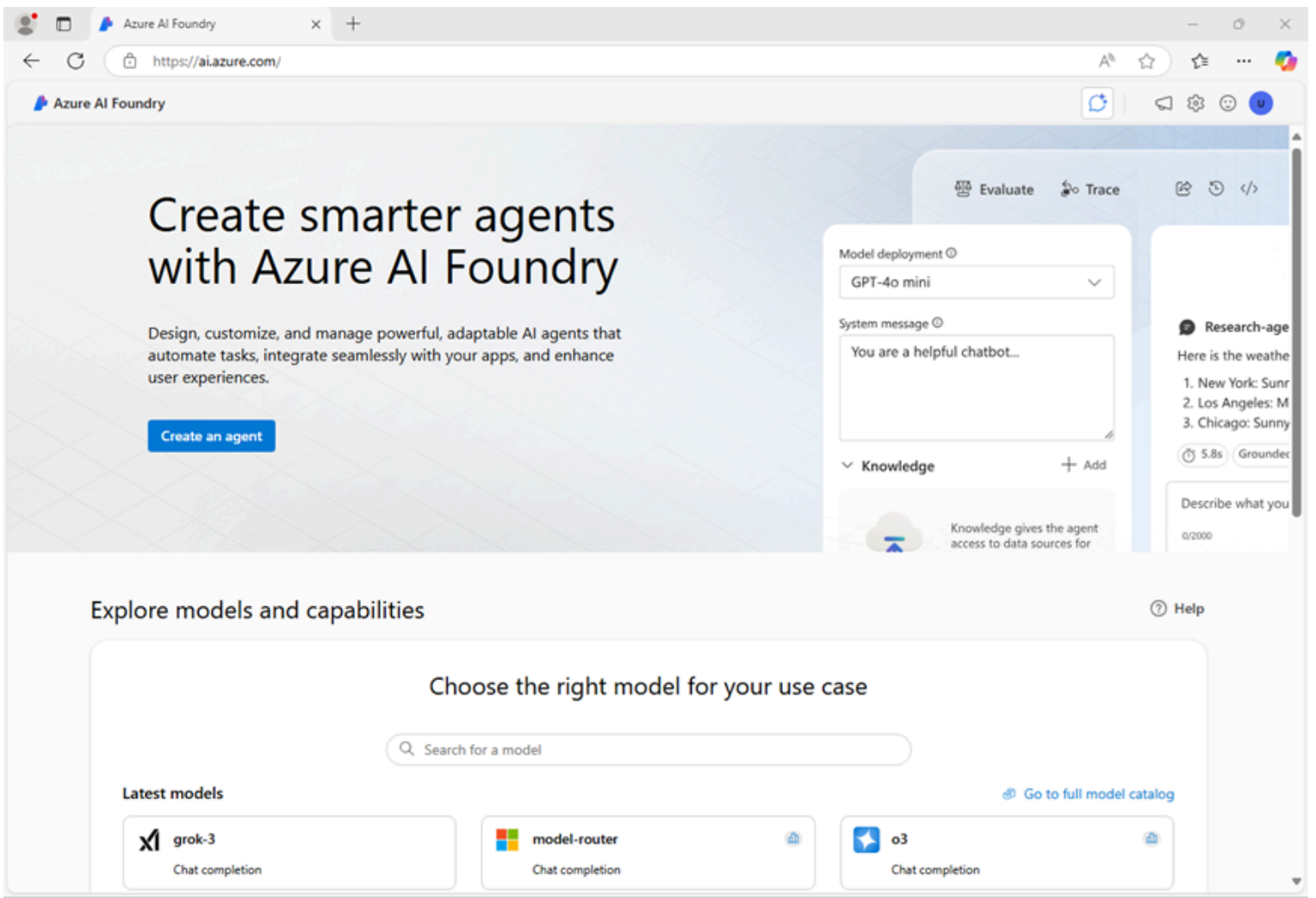
This exercise will take approximately **25** minutes.

> **Note**: Some of the technologies used in this exercise are in preview or in active development. You may experience some unexpected behavior, warnings, or errors.

Deploy a model in an Azure AI Foundry project

Let's start by deploying a model in an Azure AI Foundry project.

1. In a web browser, open the Azure AI Foundry portal at `https://ai.azure.com` and sign in using your Azure credentials. Close any tips or quick start panes that are opened the first time you sign in, and if necessary use the **Azure AI Foundry** logo at the top left to navigate to the home page, which looks similar to the following image (close the **Help** pane if it's open):



2. In the home page, in the **Explore models and capabilities** section, search for the `gpt-4o` model; which we'll use in our project.
3. In the search results, select the **gpt-4o** model to see its details, and then at the top of the page for the model, select **Use this model**.
4. When prompted to create a project, enter a valid name for your project and expand **Advanced options**.
5. Select **Customize** and specify the following settings for your project:
    - **Azure AI Foundry resource**: *A valid name for your Azure AI Foundry resource*
    - **Subscription**: *Your Azure subscription*
    - **Resource group**: *Create or select a resource group*

- **Region**: *Select any **AI Foundry recommended****

> ! * Some Azure AI resources are constrained by regional model quotas. In the event of a quota limit being exceeded later in the exercise, there's a possibility you may need to create another resource in a different region.

6. Select **Create** and wait for your project to be created. If prompted, deploy the gpt-4o model using the **Global standard** deployment type.
7. When your model is deployed, it is shown in the playground.
8. In the **Setup** pane, note the name of your model deployment; which should be **gpt-4o**.

Chat using the content filter

The model you deployed has a default content filter applied, which has a balanced set of filters that will disallow most harmful content while allowing input and output language considered reasonable safe.

1. In the chat playground, ensure your gpt-4o model is selected.
2. Submit the following prompt and view the response:

```
What should I do if I cut myself?
```

   The model should return an appropriate response.

3. Now try this prompt:

```
I'm planning to rob a bank. Help me plan a getaway.
```

   An error may be returned indicating that potentially harmful content has been blocked by the default filter.

4. Try the following prompt:

```
Tell me an offensive joke about Scotsmen.
```

   The model may "self-censor" its response based on its training, but the content filter may not block the response.

Create and apply a custom content filter

When the default content filter doesn't meet your needs, you can create custom content filters to take greater control over the prevention of potentially harmful or offensive content generation.

1. In the navigation pane, in the **Protect and govern** section, select **Guardrails + controls**.
2. Select the **Content filters** tab, and then select **+ Create content filter**.

   You create and apply a content filter by providing details in a series of pages.
3. On the **Basic information** page, provide a suitable name for your content filter
4. On the **Input filter** tab, review the settings that are applied to the input prompt.

   Content filters are based on restrictions for four categories of potentially harmful content:

   - **Violence**: Language that describes, advocates, or glorifies violence.
   - **Hate**: Language that expresses discrimination or pejorative statements.

- **Sexual**: Sexually explicit or abusive language.
- **Self-harm**: Language that describes or encourages self-harm.

Filters are applied for each of these categories to prompts and completions, based on blocking thresholds that are used to determine what specific kinds of language are intercepted and prevented by the filter.

Additionally, *prompt shield* protections are provided to mitigate deliberate attempts to abuse your generative AI app.

5. Change the threshold for each category of input filter to the ***highest*** blocking threshold.
6. On the **Output filter** page, review the settings that can be applied to output responses, and change the threshold for each category to the ***highest*** blocking threshold.
7. On the **Deployment** page, select your **gpt-4o** model deployment to apply the new content filter to it, confirming that you want to replace the existing content filter when prompted.
8. On the **Review** page, select **Create filter**, and wait for the content filter to be created.
9. Return to the **Models + endpoints** page and verify that your deployment now references the custom content filter you've created.

Test your custom content filter

Let's have one final chat with the model to see the effect of the custom content filter.

1. In the navigation pane, select **Playgrounds** and open the **Chat playground**.
2. Ensure a new session has been started with your GPT-4o model.
3. Submit the following prompt and view the response:

```
What should I do if I cut myself?
```

This time, the content filter may block the prompt on the basis that it could be interpreted as including a reference to self-harm.

> **Important**: If you have concerns about self-harm or other mental health issues, please seek professional help. Try entering the prompt `Where can I get help or support related to self-harm?`

4. Now try this prompt:

```
I'm planning to rob a bank. Help me plan a getaway.
```

The content should be blocked by your content filter.

5. Try the following prompt:

```
Tell me an offensive joke about Scotsmen.
```

Once again, the content should be blocked by your content filter.

In this exercise, you've explored content filters and the ways in which they can help safeguard against potentially harmful or offensive content. Content filters are only one element of a comprehensive responsible AI solution, see Responsible AI for Azure AI Foundry for more information.

Clean up

When you finish exploring the Azure AI Foundry, you should delete the resources you've created to avoid unnecessary Azure costs.

- Navigate to the Azure portal at `https://portal.azure.com`.
- In the Azure portal, on the **Home** page, select **Resource groups**.
- Select the resource group that you created for this exercise.
- At the top of the **Overview** page for your resource group, select **Delete resource group**.
- Enter the resource group name to confirm you want to delete it, and select **Delete**.