**Deploy and test a language understanding model**

**Deploy model**

After you have reviewed the model's performance and decide it's fit to be used in your environment, you need to assign it to a deployment to be able to query it. Assigning the model to a deployment makes it available for use through the [prediction API](https://learn.microsoft.com/en-us/rest/api/language/2023-04-01/conversation-analysis-runtime/analyze-conversation). It is recommended to create a deployment named production to which you assign the best model you have built so far and use it in your system. You can create another deployment called staging to which you can assign the model you're currently working on to be able to test it. You can have a maximum on 10 deployments in your project.

1. Select **Deploying a model** from the left side menu.

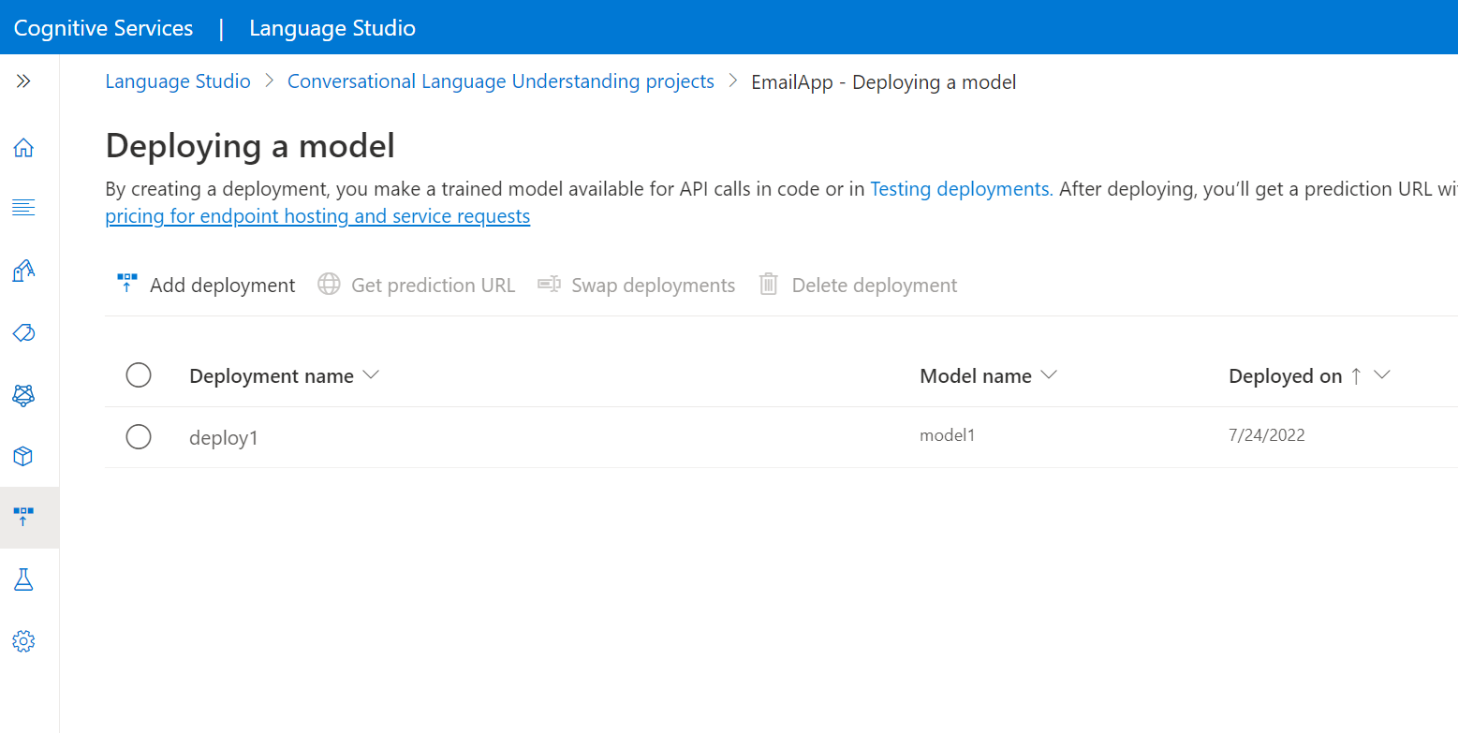
A screenshot of a computer

Description automatically generated

1. Select **Add deployment** to start the **Add deployment** wizard.

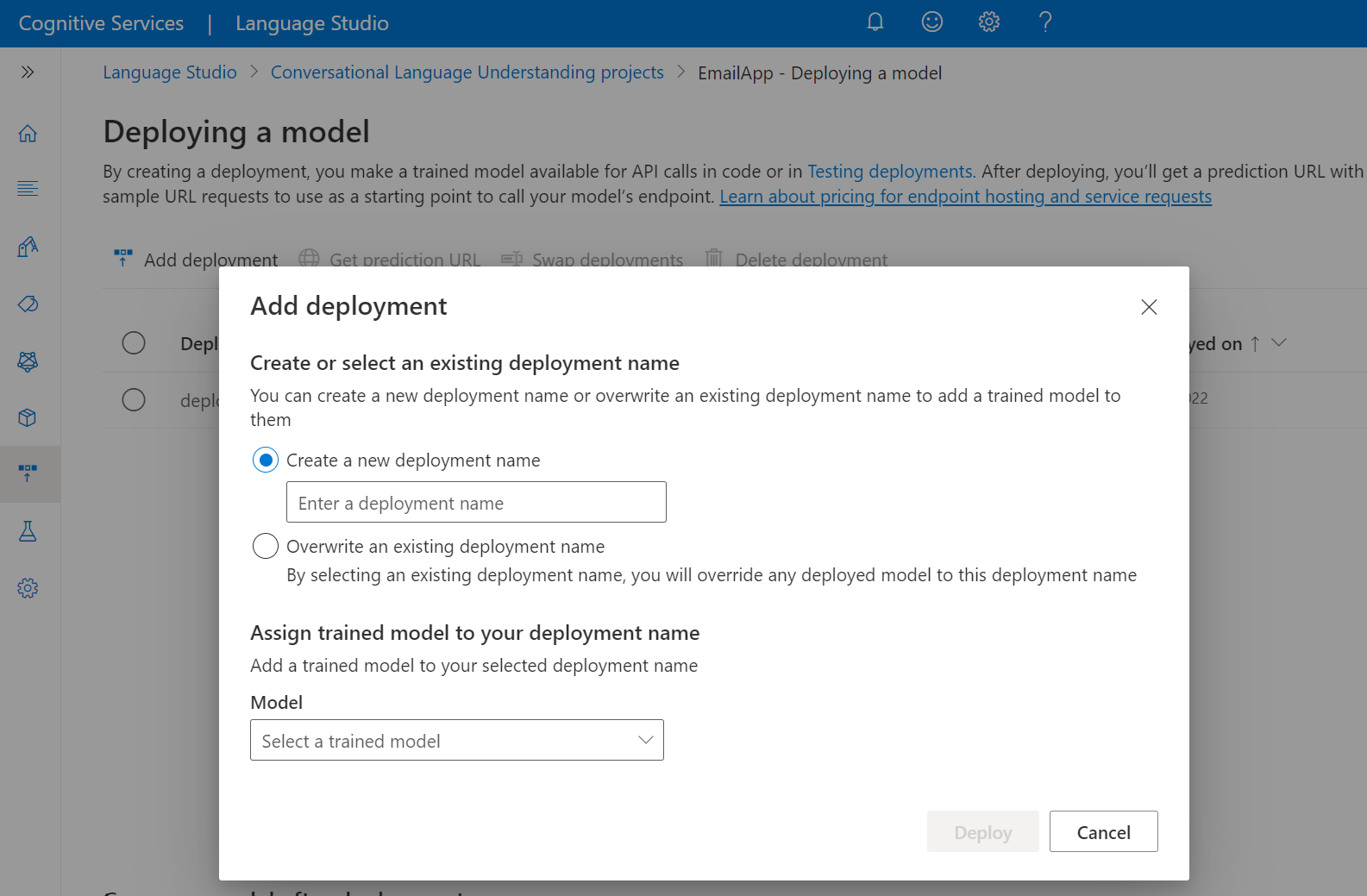
A screenshot of a computer

Description automatically generated

[](https://learn.microsoft.com/en-us/azure/ai-services/language-service/conversational-language-understanding/media/add-deployment-model.png#lightbox)

1. Select **Create a new deployment name** to create a new deployment and assign a trained model from the dropdown below. You can otherwise select **Overwrite an existing deployment name** to effectively replace the model that's used by an existing deployment.

**Note :** Overwriting an existing deployment doesn't require changes to your [**Prediction API**](https://aka.ms/clu-runtime-api) call but the results you get will be based on the newly assigned model.

[](https://learn.microsoft.com/en-us/azure/ai-services/language-service/conversational-language-understanding/media/create-deployment-job.png#lightbox)

A screenshot of a computer

Description automatically generated

1. Select a trained model from the **Model** dropdown.
2. Select **Deploy** to start the deployment job.
3. After deployment is successful, an expiration date will appear next to it. [Deployment expiration](https://learn.microsoft.com/en-us/azure/ai-services/language-service/concepts/model-lifecycle#expiration-timeline) is when your deployed model will be unavailable to be used for prediction, which typically happens **twelve** months after a training configuration expires.

A screenshot of a computer

Description automatically generated

After the deployment is added successfully, you can query the deployment for intent and entities predictions from your utterance based on the model you assigned to the deployment. You can query the deployment programmatically through the [prediction API](https://aka.ms/ct-runtime-swagger) or through the client libraries (Azure SDK).

**Test deployed model**

You can use Language Studio to submit an utterance, get predictions and visualize the results.

To test your deployed models from within the [Language Studio](https://aka.ms/LanguageStudio):

1. Select **Testing deployments** from the left side menu.

A screenshot of a computer

Description automatically generated

1. For multilingual projects, from the **Select text language** dropdown, select the language of the utterance you're testing.
2. From the **Deployment name** dropdown, select the deployment name corresponding to the model that you want to test. You can only test models that are assigned to deployments.
3. In the text box, enter an utterance to test. For example, if you created an application for email-related utterances you could enter *Delete this email*.
4. Towards the top of the page, select **Run the test**.

A screenshot of a computer

Description automatically generated

1. After you run the test, you should see the response of the model in the result. You can view the results in entities cards view or view it in JSON format.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Send a conversational language understanding request**

1. After the deployment job is completed successfully, select the deployment you want to use and from the top menu select **Get prediction URL**.

A screenshot of a computer

Description automatically generated

1. In the window that appears, copy the sample request URL and body into your command line.
2. Replace <YOUR\_QUERY\_HERE> with the actual text you want to send to extract intents and entities from.
3. Submit the POST cURL request in your terminal or command prompt. You'll receive a 202 response with the API results if the request was successful.