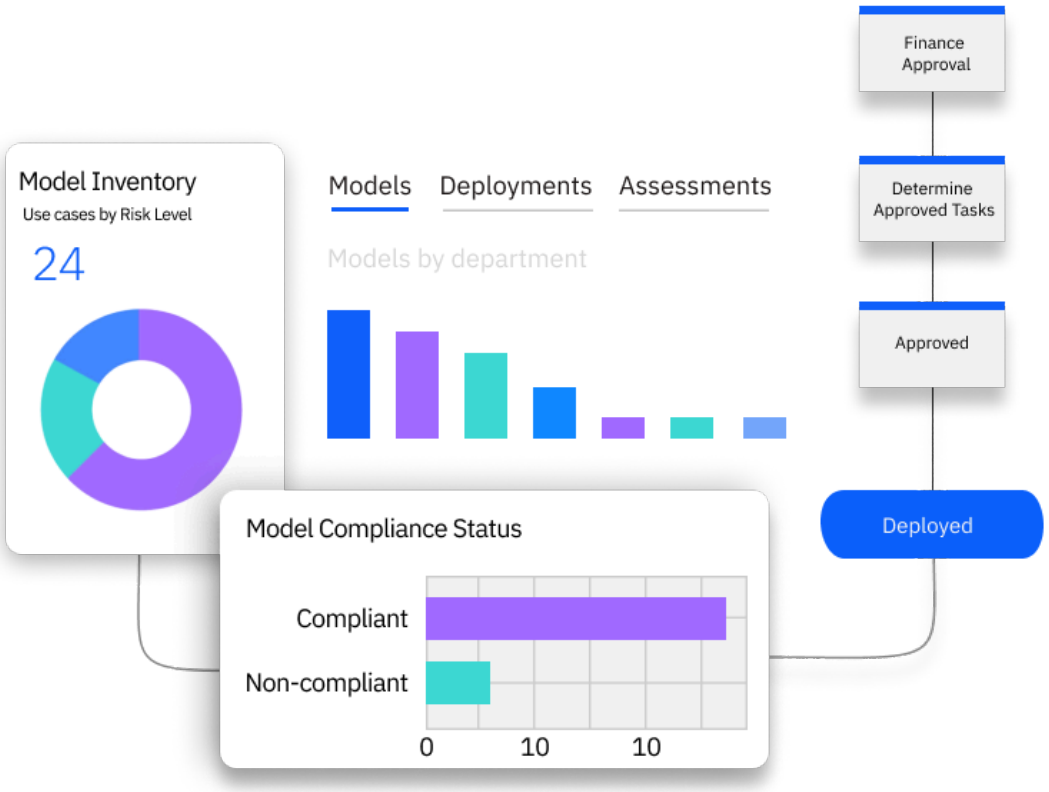
Govern generative models

watsonx.governance

Hands-on Lab Guide



Eric Martens
emartens@us.ibm.com
Information Developer, WW Technology

Contents

Govern generative models

Introduction

In this lab, you will go through the steps of the approval workflow you customized during the governance console configuration steps. The human resources department has received a large number of applications for open positions, and would like to use AI to summarize them to help save time for the hiring department, and process the applications more efficiently to improve the experience for the applicants.

Most use cases for generative models involve interacting with prompts and prompt templates, which help users provide clear input to a Large Language Model(LLM) by giving them a structured framework to follow, which in turn helps the model generate accurate responses.

Getting help - PLEASE READ

This is an extremely lengthy, highly technical lab that touches on multiple products and environments that are all under active development. Every effort has been made to address possible causes and issues in the instructions themselves; however, it is not uncommon for problems to arise, error messages to appear, or screens to sometimes differ from the lab instructions.

> ⛔ PLEASE refer to the Troubleshooting section of the lab first to see if your problem is addressed there. That section will be continually updated to respond to the most frequent issues encountered in running the lab.

> ⚠ If your issue is not addressed, PLEASE contact the author via Slack if at all possible. IBMers can reach Eric Martens via Slack. Business partners can reach out via email.

Leaving comments on the YourLearning page or attempting to address issues via a TechZone ticket will eventually get a response, but the above two methods are significantly preferred and will result in a much quicker resolution.

Getting started

This lab assumes that you have access to a fully-configured watsonx.governance environment, with installed and running instances of the governance console (OpenPages), monitoring console (OpenScale), and Db2. It also assumes that the GlobalCorp business entities from the user management lab exist on the system. It does NOT assume that your environment is equipped with graphics processing units (GPUs). For instructions on provisioning and configuring an environment in TechZone, see the watsonx.governance configuration lab. You will need the Cloud Pak for Data console URL and login credentials created in that lab.

Environment preparation

For this particular use case, your organization's AI engineers would like to test the summarization capabilities of OpenAI's ChatGPT model running on Microsoft Azure. You will begin by requesting access to an Azure environment in TechZone. Note that if you are performing a Proof of Experience (PoX) for a client, you can adapt the materials in this lab to use other third-party LLMs.

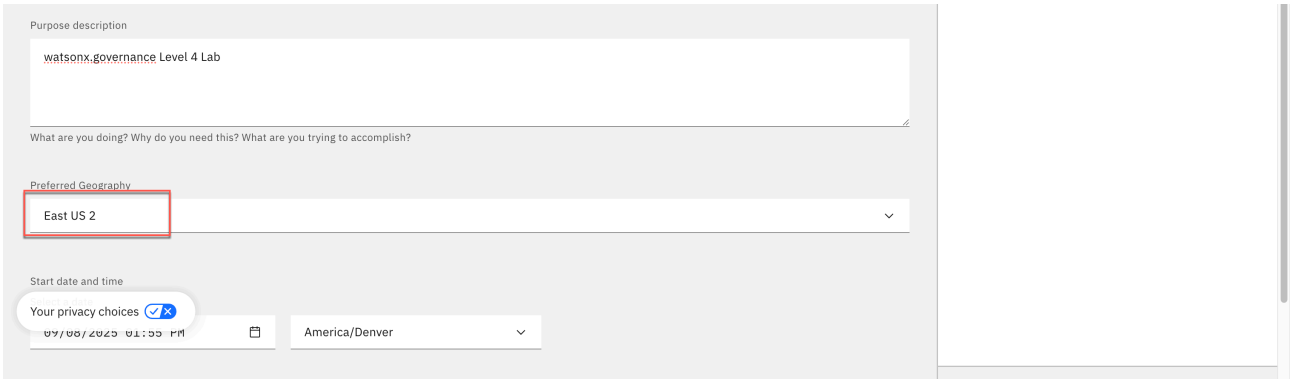
> ⚠ THE EVALUATIONS PERFORMED IN THIS LAB ARE NOT INTENDED TO SHOW THE RELATIVE STRENGTHS OF THE OPENAI OR AZURE PLATFORMS, AND SHOULD NOT BE PRESENTED AS SUCH.

1. Reserve an Azure instance

Temporary access to an OpenAI model can be reserved through IBM TechZone.

1. Click on the link for the Access to Pre-Deployed Azure OpenAI gpt-35-turbo LLM TechZone environment.

2. Sign in to TechZone and fill out the form. For the Preferred Geography field, select East US 2.



3. Check the box to agree to the terms and conditions, then click the Submit button. After a short time, your reservation will be available in your TechZone reservations list.

When the reservation has finished provisioning, you may proceed to the next step.

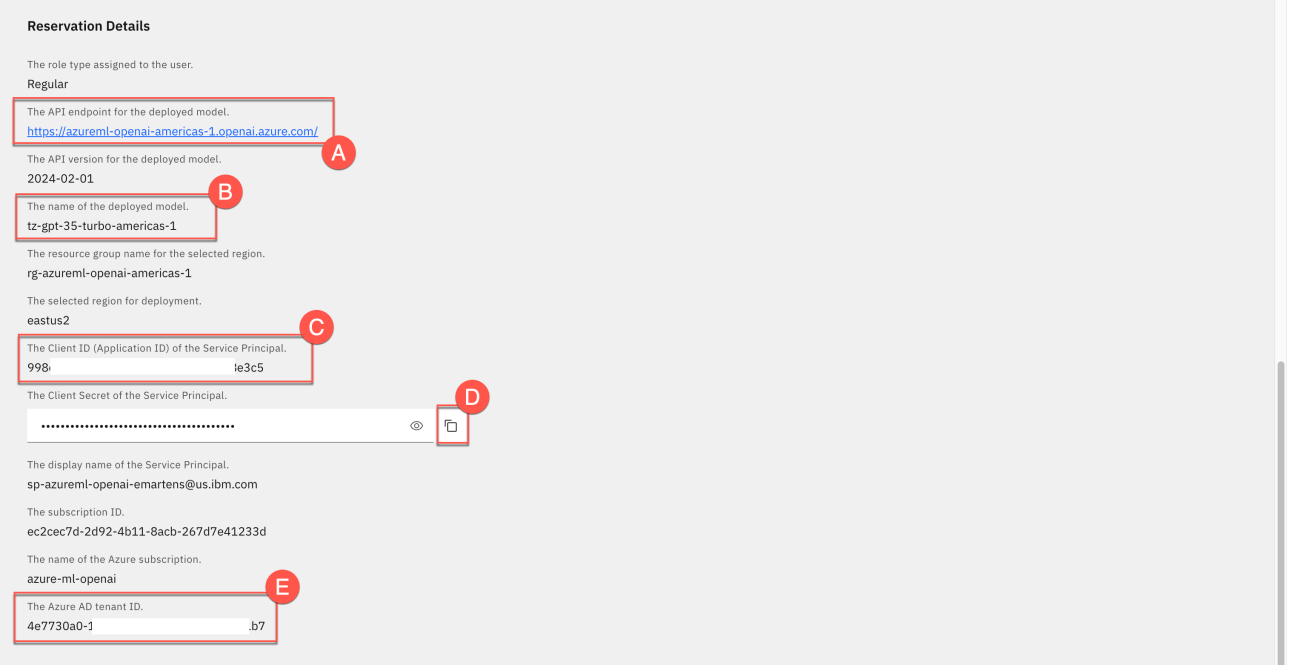
2. Collect Azure credentials

The Azure reservation comes with a set of credentials you can use to query the ChatGPT LLM. Later on in the lab, you will use these credentials in a Jupyter notebook. For now, you will need to copy them to a text file on your machine so they can be pasted into the notebook later.

1. From your TechZone reservations list, click on the tile for the Access to Pre-Deployed Azure OpenAI gpt-35-turbo LLM reservation.

2. Scroll down to the Reservation Details section of your reservation, and copy and paste the following values into a text file on your machine:

   - The API endpoint for the deployed model (A)
   - The name of the deployed model (B)
   - The Client ID (Application ID) of the Service Principal (C)
   - The Client Secret of the Service Principal (D)
   - The Azure AD tenant ID (E)



Now that you have the information necessary to access the LLM, you may proceed to the next step.

Create and progress a use case

To ensure that model use cases are tracked across the entire solution, they should be created using the watsonx governance console. During the configuration of the environment you are using, you turned on integration between the governance console (OpenPages) and watsonx, so any actions related to model use cases should now redirect you to the governance console interface.

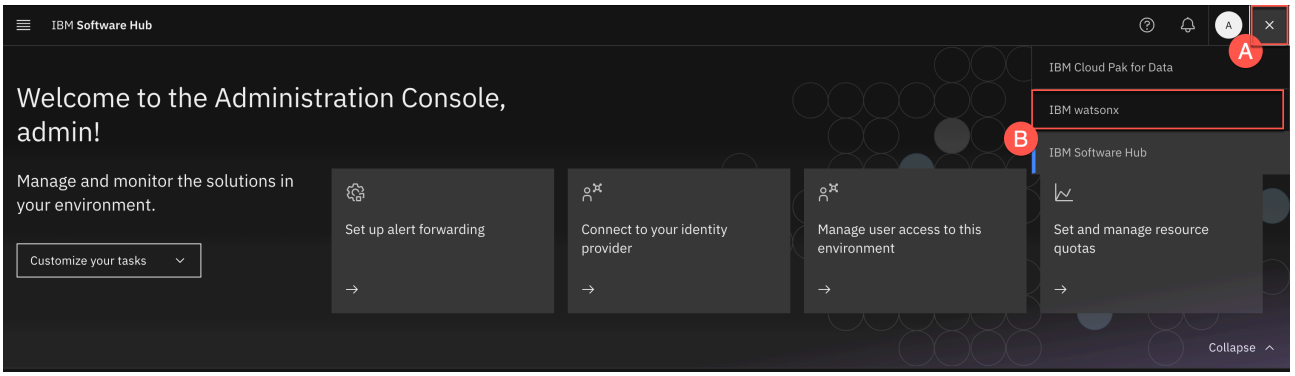
In a real-world scenario, this action would be performed by an organizational stakeholder who would like to request the development and implementation of a model; in this case, the manager of the human resources department, who is unable to keep up with the volume of resumes submitted for employment opportunities and would like help from an AI solution.

1. Sign in to watsonx.governance

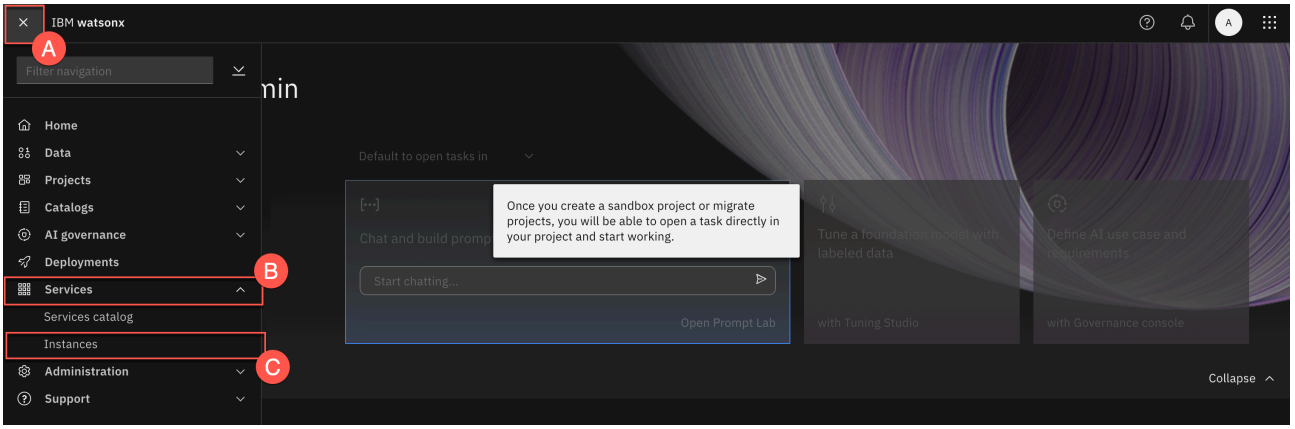
For this portion of the lab, you perform actions as the admin user. In a real-world scenario, you administrators would provide different levels of access to different users, taking advantage of pre-defined roles in Cloud Pak for Data and watsonx. While creating and managing multiple user personas and groups is beyond the scope of this lab, doing so can provide a more realistic PoX or demo for your client, particularly if they are unfamiliar with Cloud Pak for Data, watsonx, and the level of access control and collaboration provided.

1. In a web browser window, navigate to the watsonx home page using the credentials for the environment you provisioned and configured in prior labs. If you are asked to log in, skip ahead to step 4. If you opened the home page and are signed in, you will need to log out.

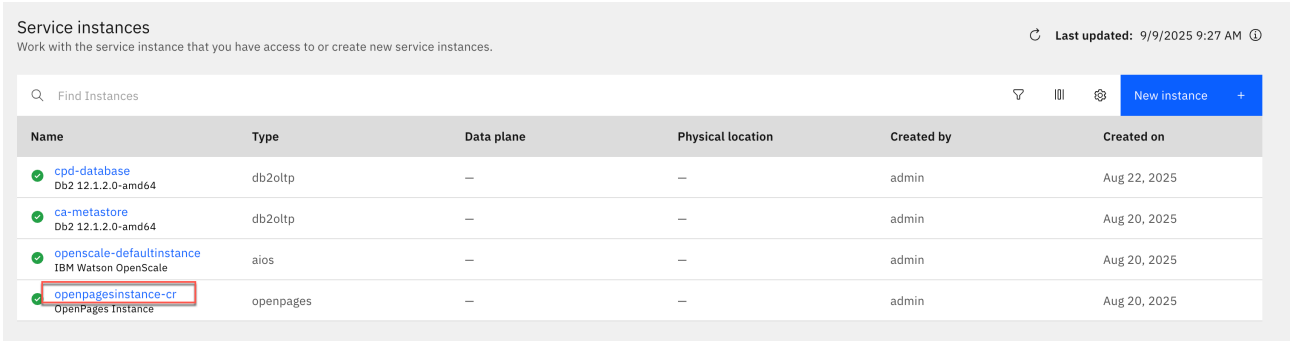
2. Click on the grid icon in the upper right to open the context menu (A). If necessary, click on the IBM watsonx menu item (B) to change the context. A Welcome to watsonx popup window may open. Close the popup window, or click the Take a tour button if you wish.



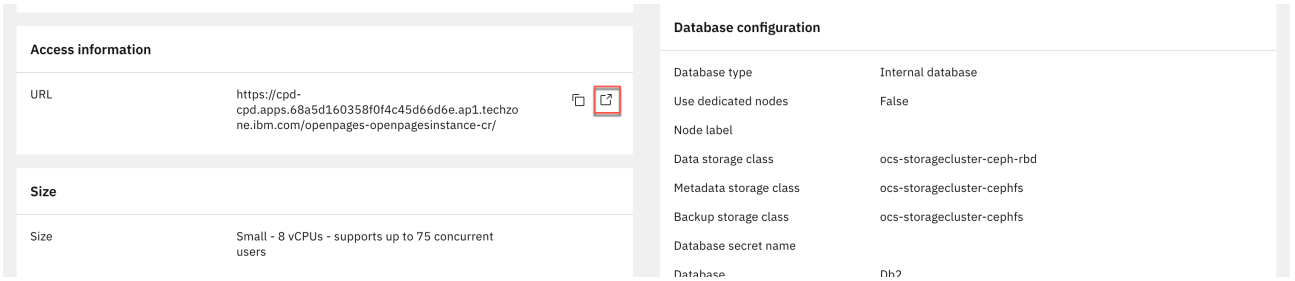
3. Click on the navigation menu in the upper left (A) to open it. Click on the Services menu item (B) to expand it. Click on the Instances menu item (C). The Instances screen opens.



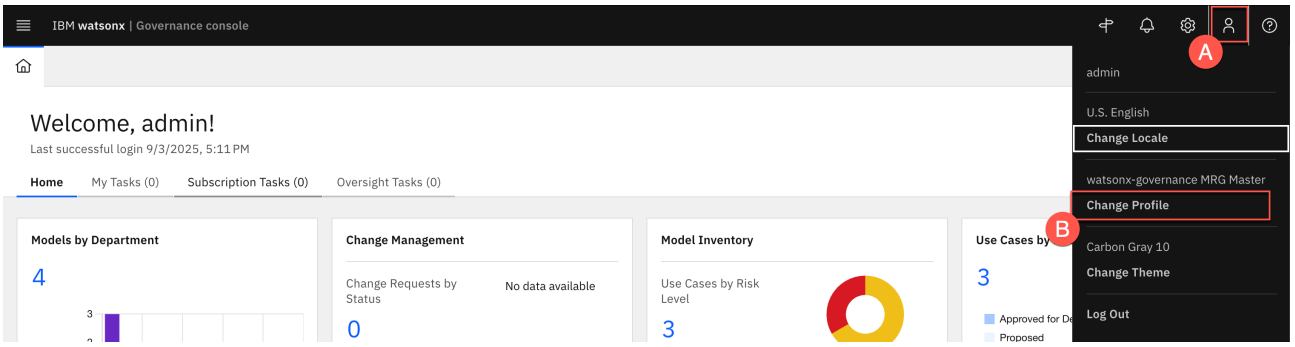
4. From the Instances list, locate and click on the OpenPages instance.



5. Scroll down to the Access information section of the screen and click on the Launch icon to launch the watsonx governance console (OpenPages).



6. Once the governance console opens, you may need to switch to the correct profile to see all the applicable fields. Click the avatar icon in the upper right (A). The User menu opens. Click the Change Profile menu item (B). The Select profile dialog opens.



Take a moment to review the different profile roles and descriptions available. Each of these can be customized, or new profiles created, to fit the structure and requirements of the organization. While this lab will deal primarily with the watsonx-governance MRG Master for governing models, pre-defined profiles also exist for regulatory compliance officers (watsonx-governance RCM Master) and for risk managers (watsonx-governance ORM Master).

7. Click on the watsonx-governance MRG Master profile from the list to select it.

8. Click Save to finalize your choice.

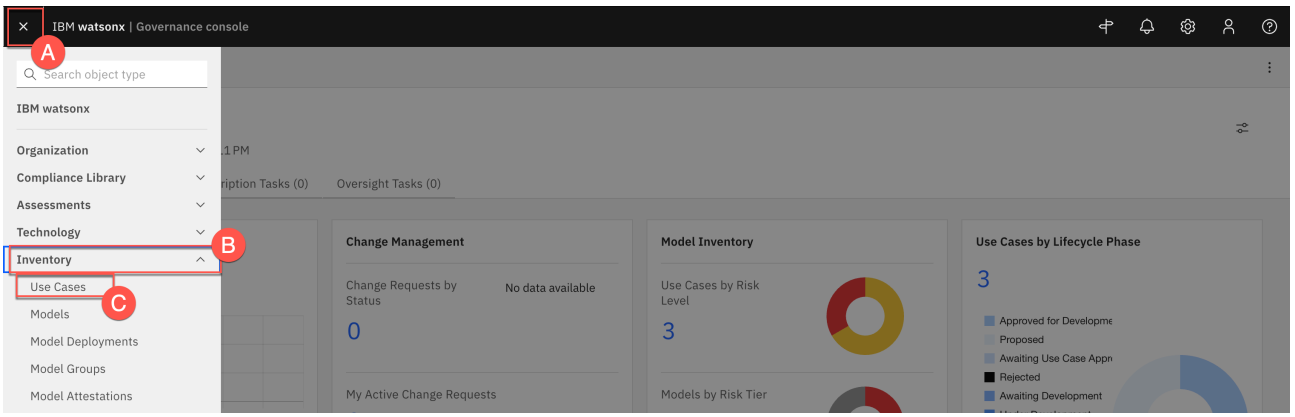
Note that when you return to the dashboard, it is populated with several charts displaying metrics for the sample models and use cases you loaded in the configuration step. The pre-defined roles have been created to display the information most useful for that role in their dashboards. Like all aspects of the governance console, the dashboard charts can be customized per role, or per individual user.

2. Create a model use case

The model governance process begins with the creation of a model use case. A use case is meant to track and capture information about a collection of models and prompts that will be built to serve a particular purpose. A use case should be created whenever there is a business need requiring the use of a model (AI or non-AI) to be built. Model records should then be added as a child of the use case.

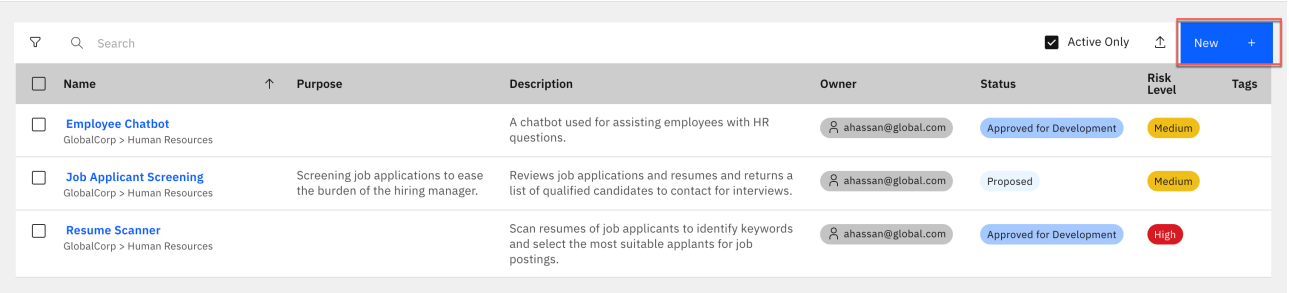
To ensure that model use cases are tracked across the entire solution, they should be created using the watsonx governance console. In the configuration lab, you turned on integration between the governance console (OpenPages) and watsonx, so any actions related to model use cases should now redirect you to the governance console interface.

1. Click on the Primary menu in the upper left (A) to open it. Click on the Inventory menu item (B) to expand it. Click on the Use Cases menu item (C). The Use Cases tab opens. Note that several sample uses cases were loaded during the FastMap import step you performed in the configuration lab.



2. Click the blue New button. The New Use Case tab opens.

Use Cases (3)



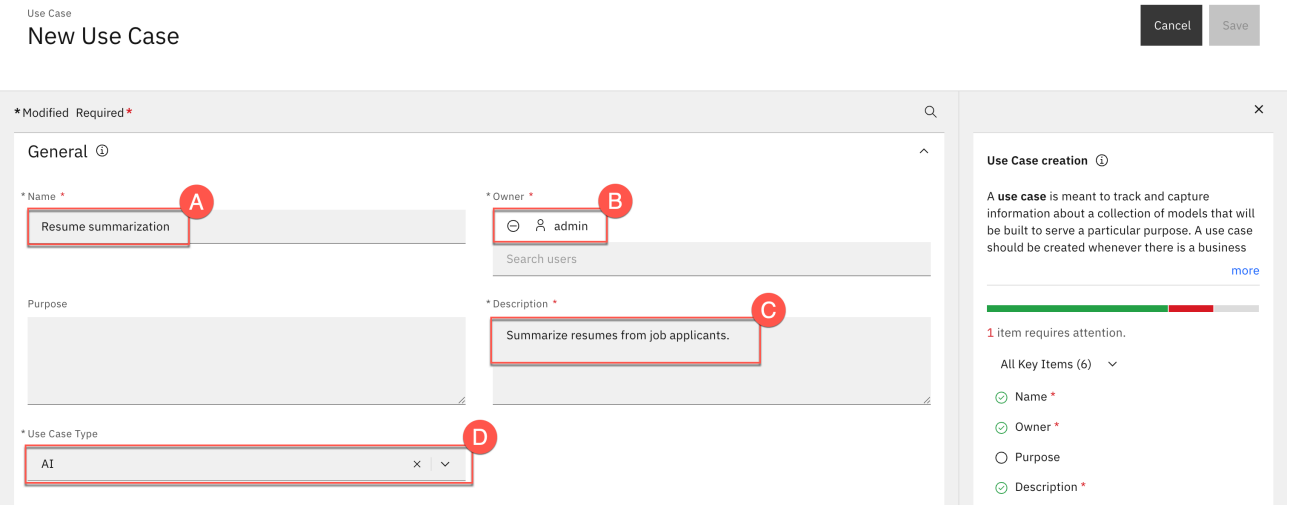
Note that the Model Use Case creation information panel on the right of the screen offers helpful information about model use cases, as well as a list of required fields. Clicking on any of the fields in that panel will scroll the screen directly to that portion of the form, helping you quickly rectify any items needing attention.

3. In the General section of the form, enter Resume summarization in the Name field (A). Note that when you enter a value in the field, the progress bar in the Model Use Case creation information panel updates.

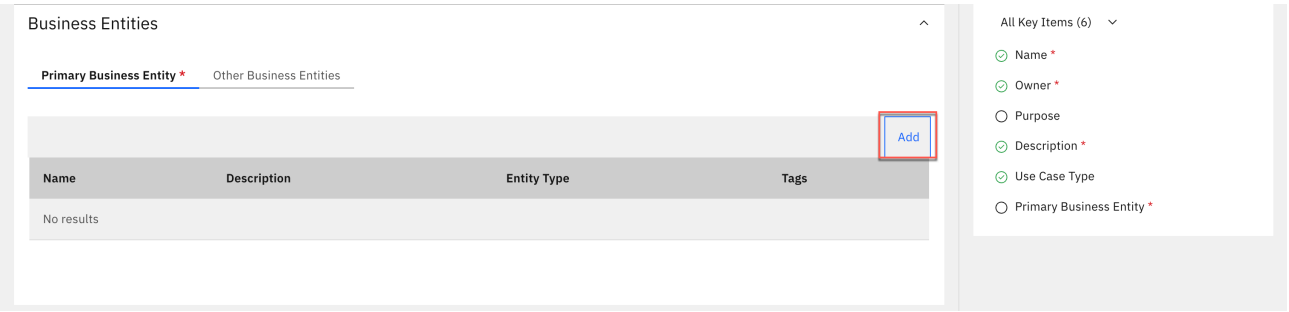
Click the Owner field and enter the admin user into this field (B). If you wish, you could enter the ahassan@global.com created user account for Aisha Hassan, the Human Resources manager you created during the configuration portion of the lab. However, you would need to log out of the environment as the admin user and then log back in as Aisha Hassan in order to proceed with the workflow, for simplicity's sake, you will use the admin user.

Enter a description in the Description field (C).

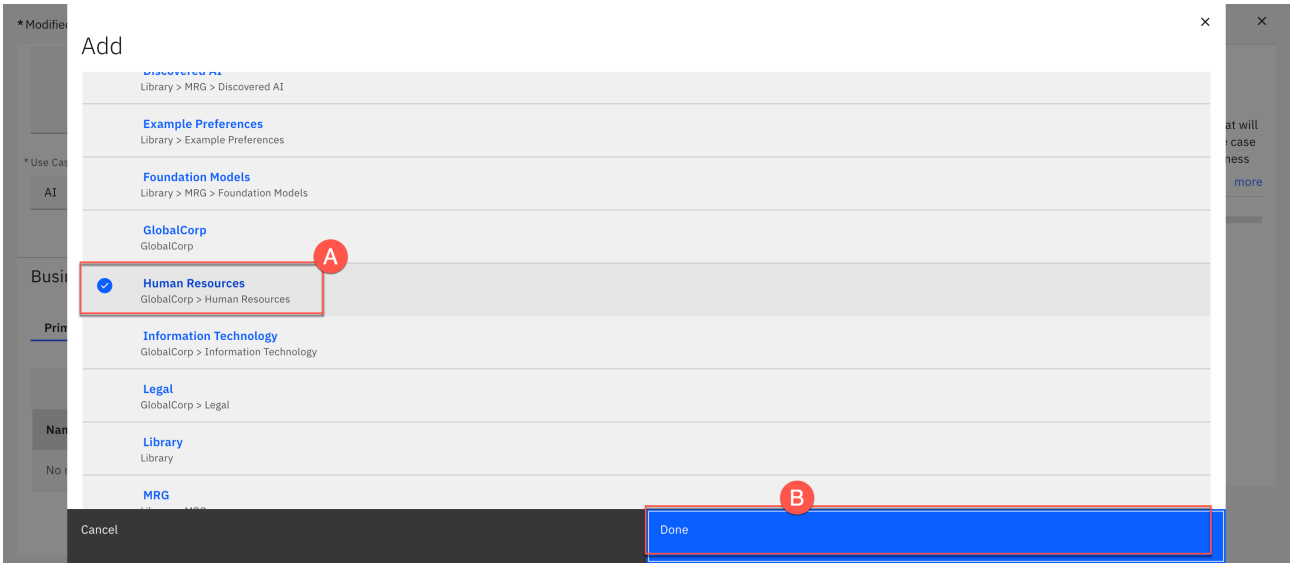
Click on the Use Case Type dropdown and select AI (D).



4. All model use cases are owned by business entities, representing the part of the organization responsible for requesting the use case. In the Business Entities section of the form, click the Add button. The Add window opens with a list of business entities defined for the organization.



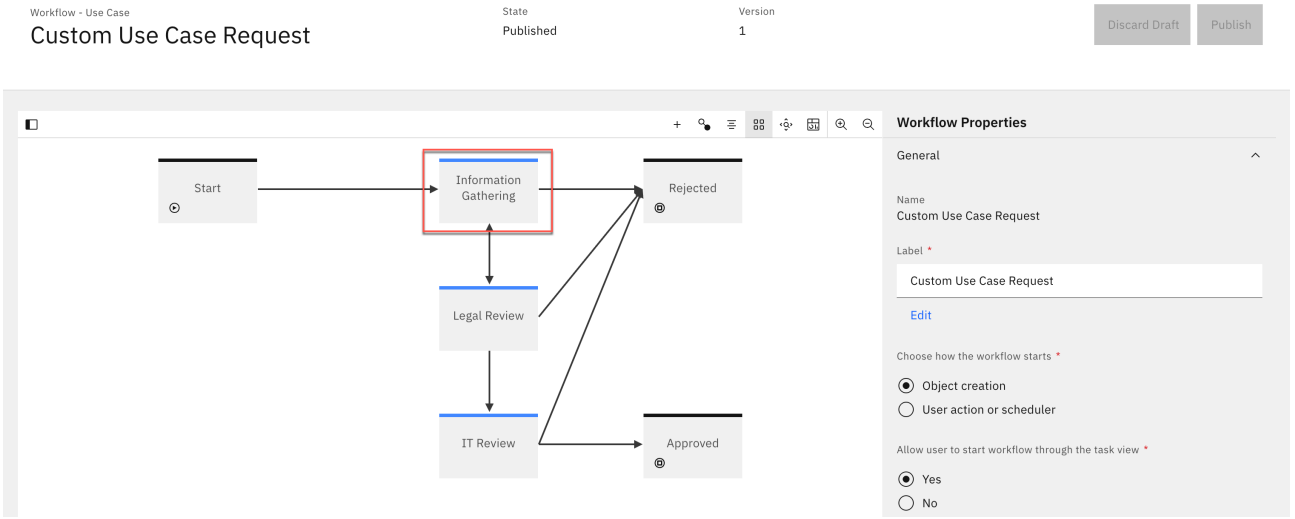
5. Locate the Human Resources entity from the list (A) and click on it to select it. Click on the Done button (B) to add the business entity to the use case. The Add window closes.



6. Click the Save button in the upper right to save the use case.

When the use case has finished saving, the screen will reload. At this point, the use case has been created and is now governed by the Custom Use Case Request workflow that you created in a previous lab. Specifically, it is in the Information Gathering stage of that workflow, as shown in the screenshot below.

> ⚠ Note that this screen is for informational purposes, and your screen will not look like this.

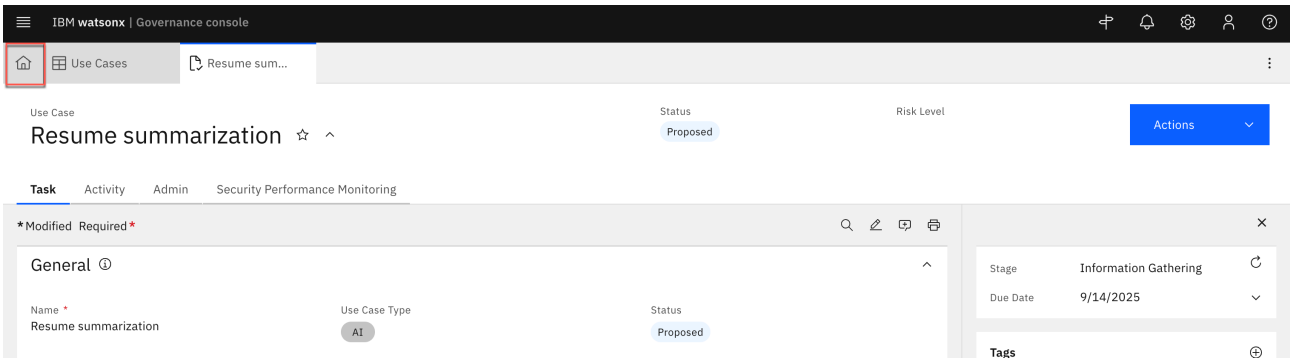


To progress the use case through the workflow, you will now need to perform the actions specified in the Action items in the workflow.

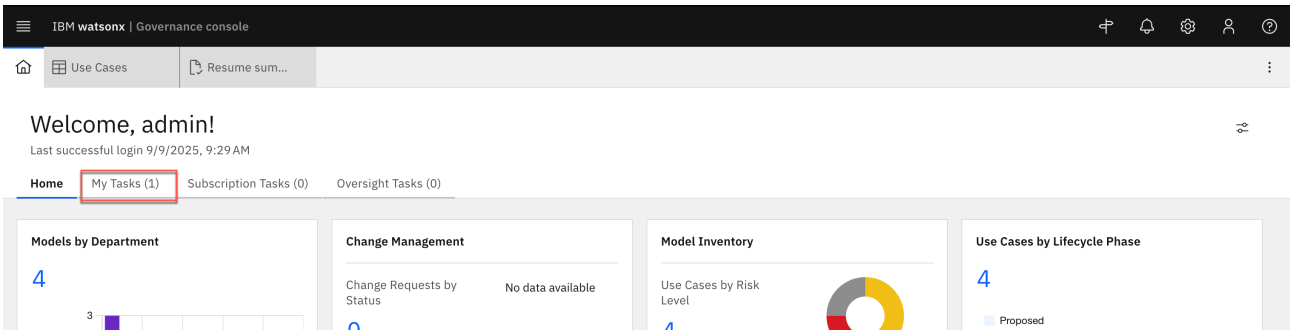
3. Progress the use case to the next phase

The use case request has progressed to the Information Gathering stage of the workflow, and has been assigned as an action for the appropriate owner. Recall that owners of each stage of the workflow can be configured, and alerts assigned.

1. Click on the Home icon in the upper left to return to the user's home tab.

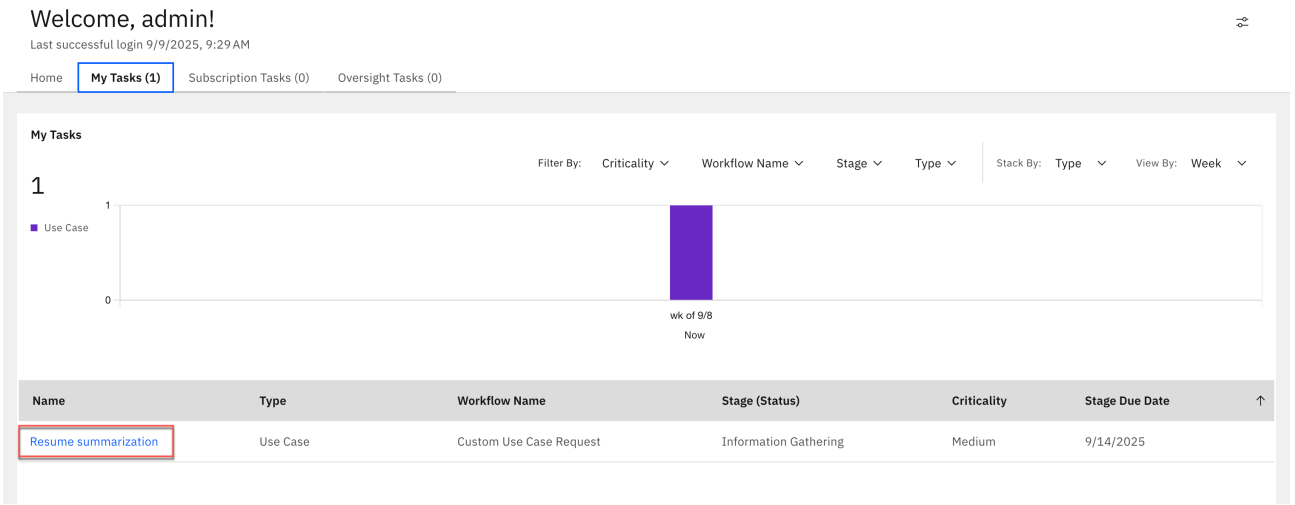


2. Note that the My Tasks tab now shows a new entry. Click on the tab to open it.



The My Tasks tab shows a list of all the current tasks assigned to the user. It can be filtered by a variety of fields. At the moment, it only contains a single task, showing that the use case request is in the data gathering stage and is in need of action, along with the stage due date.

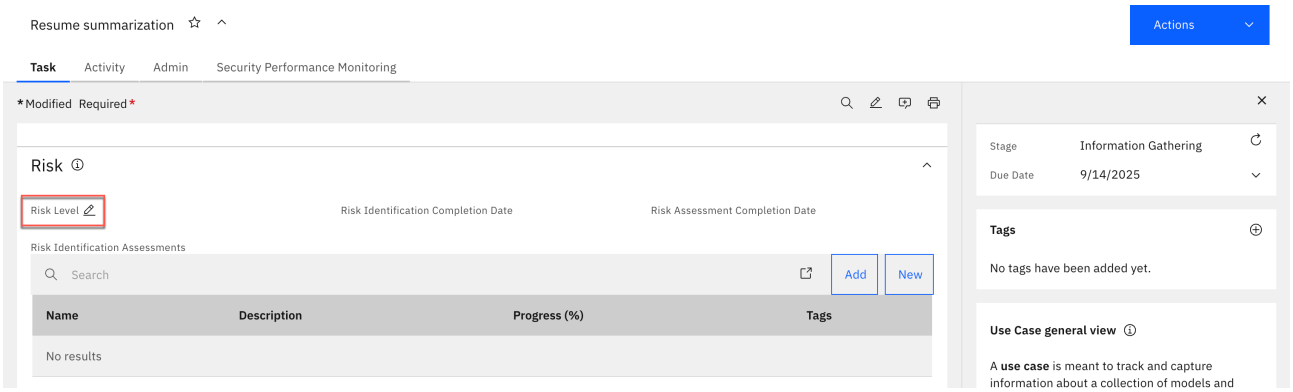
3. Click on the link for the Resume summarization link in the table to return to the use case request tab.



The Information Gathering phase of the workflow has been designed to allow stakeholders and subject-matter experts to contribute any additional materials that are relevant to the use case request. In this case, you will assign a risk level to the use case.

Risk Level represents the risk to the organization should issues arise with the models used to address the requirements laid out by the use case. A full risk assessment is beyond the scope of this lab; however, because hiring and employment violations can lead to expensive litigation damage to an organization's reputation, this use case will be marked as high risk.

4. In the Risk section of the form, hover your mouse pointer over the Risk Level field. Click on the pencil icon that appears to edit the field.

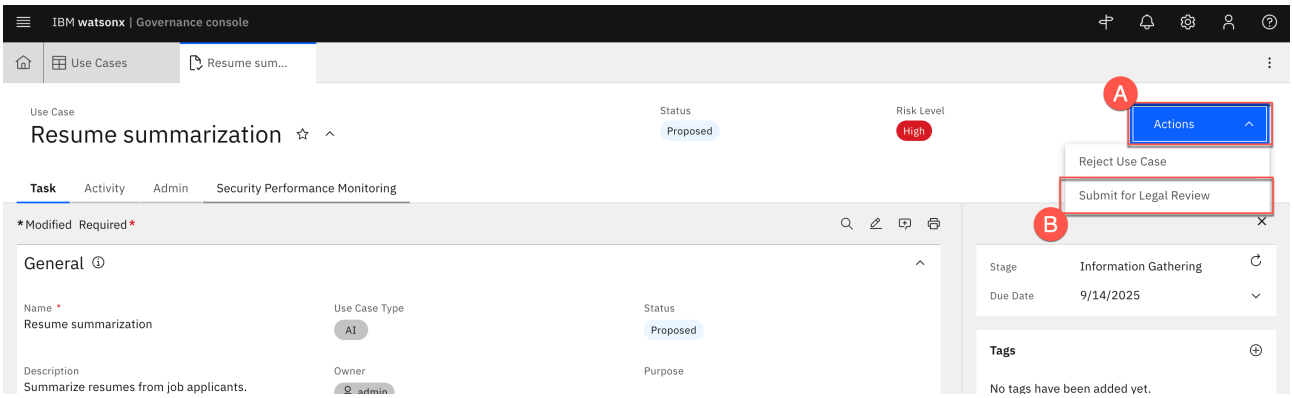


5. Select High from the dropdown.

6. Locate the Uses Generative AI field in the Use Case Details section. Hover your mouse over it, then click on the pencil icon that appears and select Yes to specify that this use case will use large language models.

7. Click the Save button in the upper right to save your changes.

8. Once the changes have been saved, click on the Actions button in the upper right to open the Actions menu (A). Click on the Submit for Legal Review action. Recall that the text of this action is defined by the Name field given to the action connecting the Use Case Data Gathering stage to the Initial Approval stage in the workflow you created in a previous lab. The Submit for initial approval confirmation dialog opens.

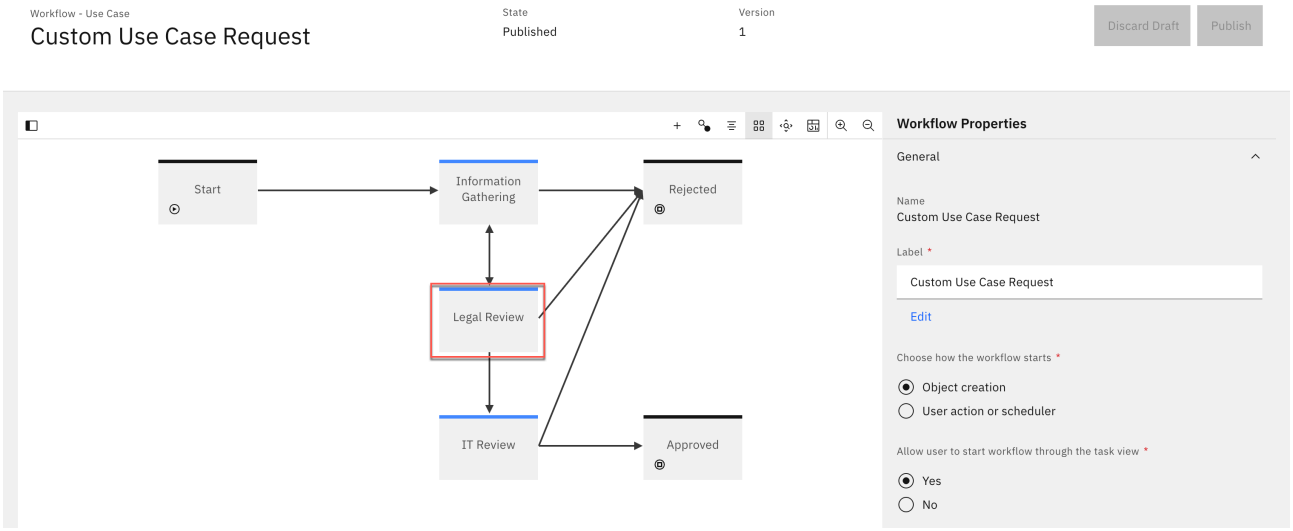


9. Click the black Continue button to confirm your action, but keep the use case tab open.

⚠ If you receive a Network error message, your change may have been recorded, but network issues may have prevented the screen from refreshing. Try submitting again; if the error persists, click the Refresh workflow info button to the right of the Stage field in the information panel on the right. The Stage should progress to Awaiting use case approval.

10. When the action completes, note that the Stage field in the information panel on the right has updated once again to Legal Review.

⚠ The screen below is provided for informational purposes. This workflow will not appear on your screen.



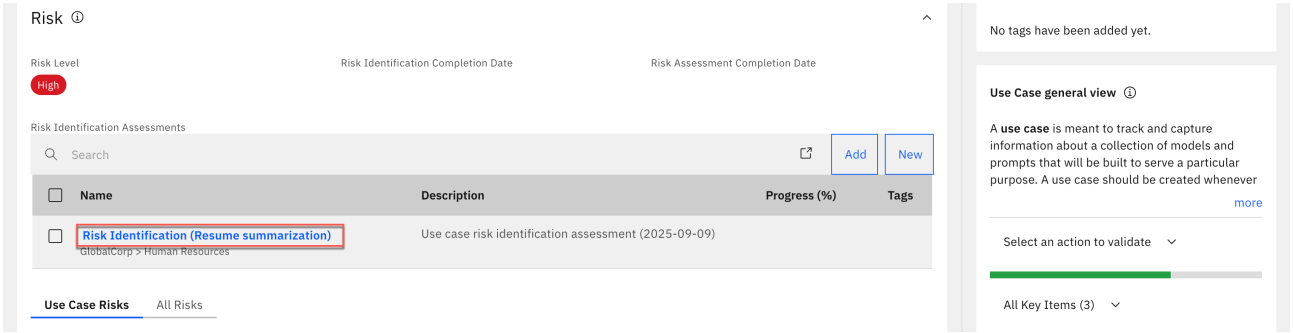
Recall that, in order to progress the use case to the next stage (IT Review) the action in the workflow requires the risk identification questionnaire to be filled out. As that questionnaire has yet to be completed, clicking on the Actions menu for the use case only shows two available actions: rejecting the use case (moving it to the Rejected stage, or returning it to the owner (moving it back to the Information Gathering stage). In order to continue forward, the questionnaire must be filled out.

4. Identify use case risks

In this section, you will fill out the risk assessment questionnaire you created in a previous lab. Note that this questionnaire, if built as instructed, is very basic and does not identify as wide a variety of risks from the AI risk atlas. If you need to demonstrate the capability to identify and manage a more diverse set of risks, you can restore the original questionnaire to the workflow by following the instructions in the custom questionnaire lab. If you do this, you will also need to create a new use case, as the workflow has already created and attached the custom questionnaire to the use case.

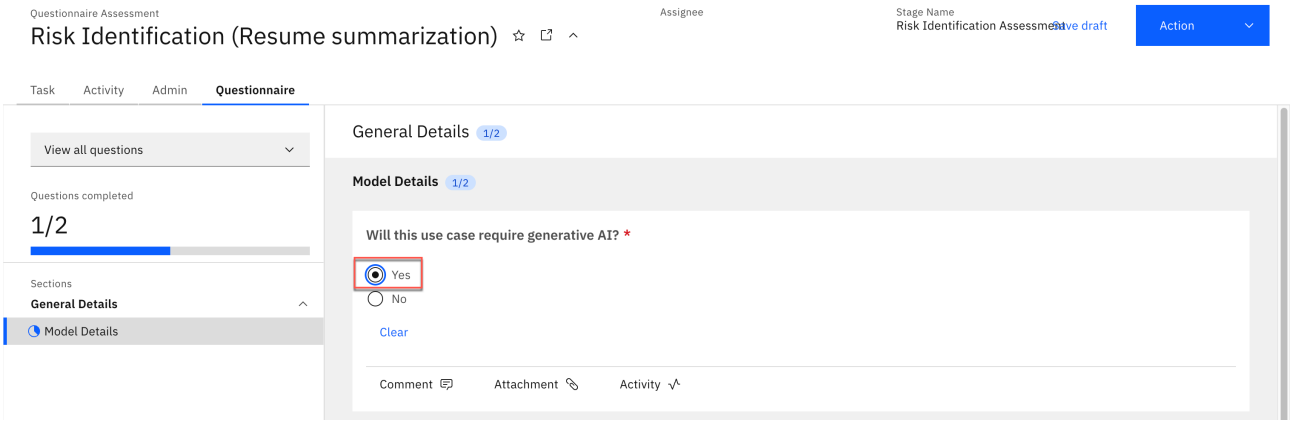
The waiting questionnaire can be found in the My Tasks menu of the Home tab. However, it is also linked on the use case record.

1. Scroll down to the Risk section of the use case record. Locate the Risk Identification (Resume summarization) assessment from the Risk Identification Assessments table, and click on it.

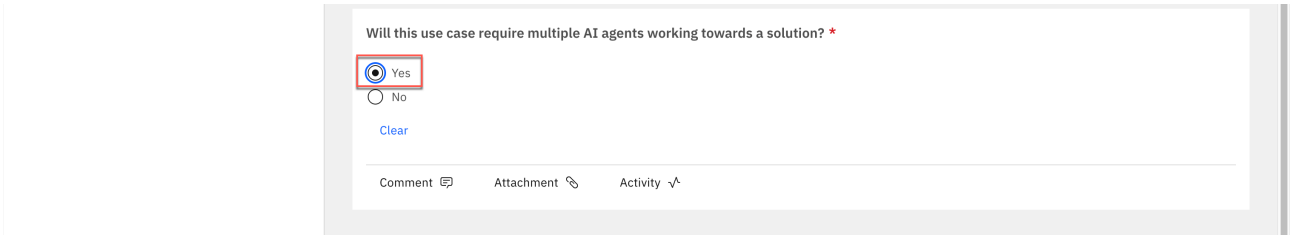


2. Fill out the questionnaire for a model that will perform summarization of resumes provided by human applicants. For the sake of this lab, you will answer in such a way that the risks you added to the risk inventory in the previous lab will be associated with this use case.

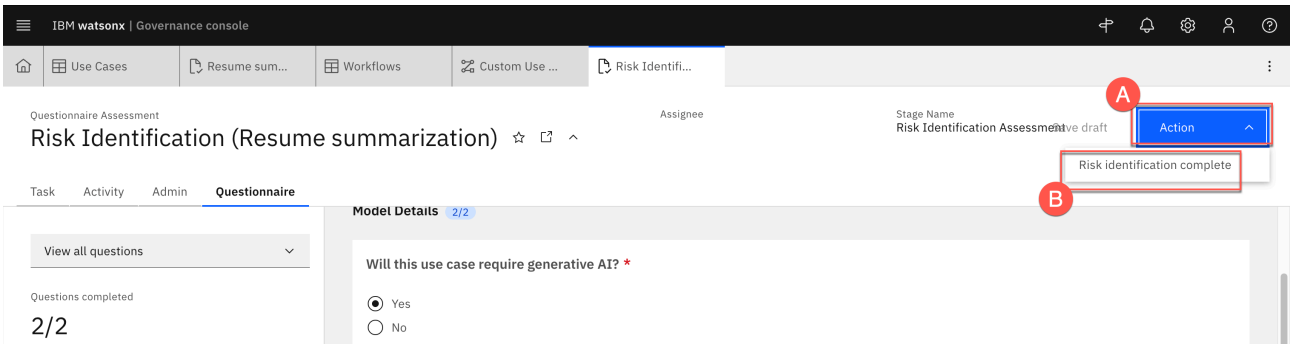
For the first question, select the Yes option to specify that the use case will use generative AI.



3. For the second question, select the Yes option to specify that the use case will use AI agents. Recall that answering Yes here means that the Identity Spoofing and Impersonation risk you created in the previous lab will be attached to this use case.



4. When you have finished filling out the survey, click the Action button in the upper right (A) to open it.. The Actions menu opens. Click the Risk identification complete button (B). A confirmation dialog opens.

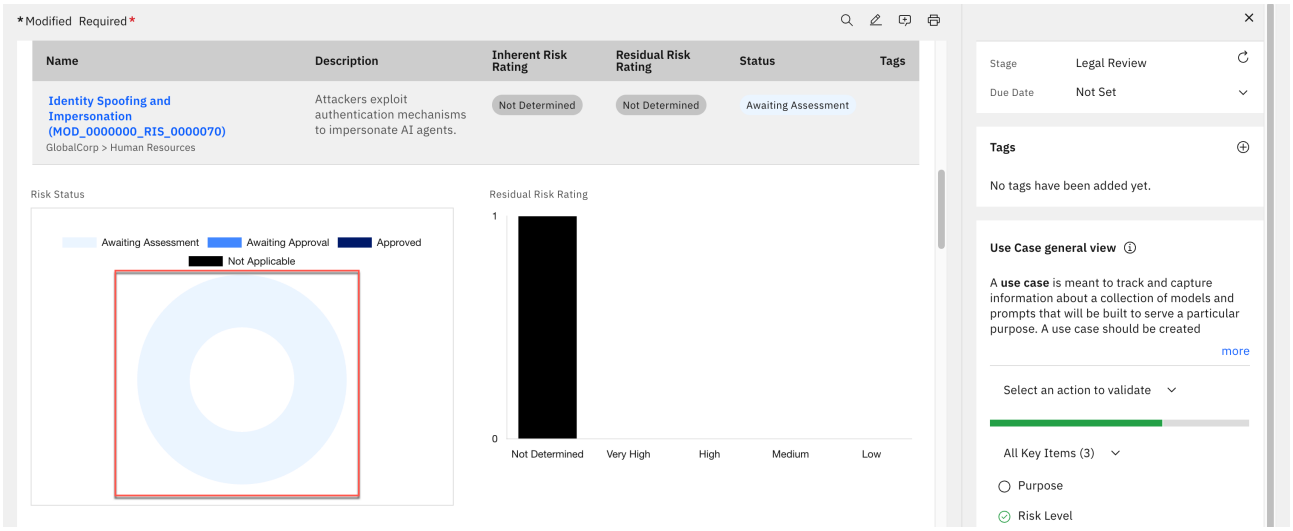


5. Click the Submit button to submit the risk identification questionnaire.

Based on the questionnaire answers, the governance console now assigns certain risks to the use case. If you have followed all lab directions to this point, the Identity Spoofing and Impersonation risk will be attached. You can view this by returning to the open tab in the governance console for the use case.

6. Scroll down to the Risk section of the page use case. Note that the Risk Identification Completion Date now has a value.

7. Scroll down to the Risk Status and Residual Risk Rating graphs. Click on the Risk Status graph. The Risks tab opens.



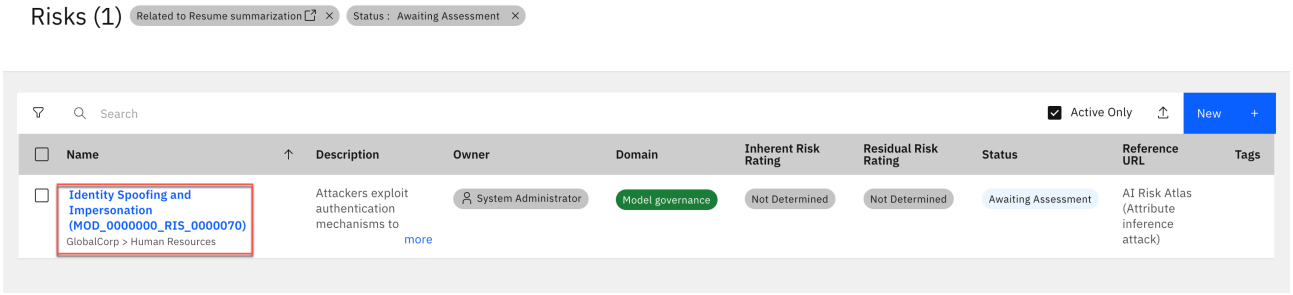
8. Examine the table of risks. Note that each has a description. Risks from the IBM AI Risk atlas will also have a reference URL for more information. They also have a Status of Awaiting assessment, indicating that a risk assessor must decide if they are relevant to the use case or not.

You have now completed an assessment regarding the model use case, which has been used to both automatically identify possible risks associated with using AI and helped insure regulatory compliance. Next, you will individually assess the risks identified by the questionnaire.

5. Assess individual risks

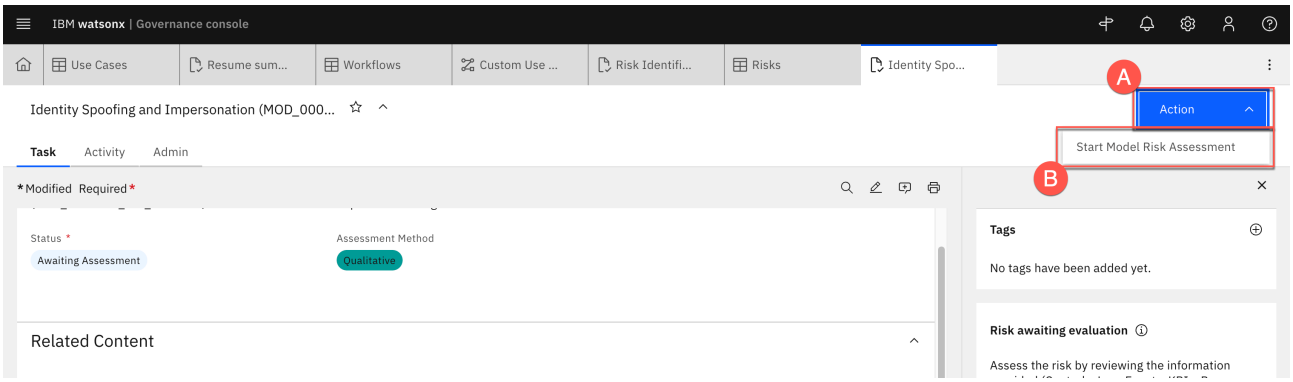
In this section, you will assess individual risks. For the sake of time, you will only perform a single in-depth assessment, to see how this is handled in the governance console.

1. Click on an entry from the Risks table to open it.

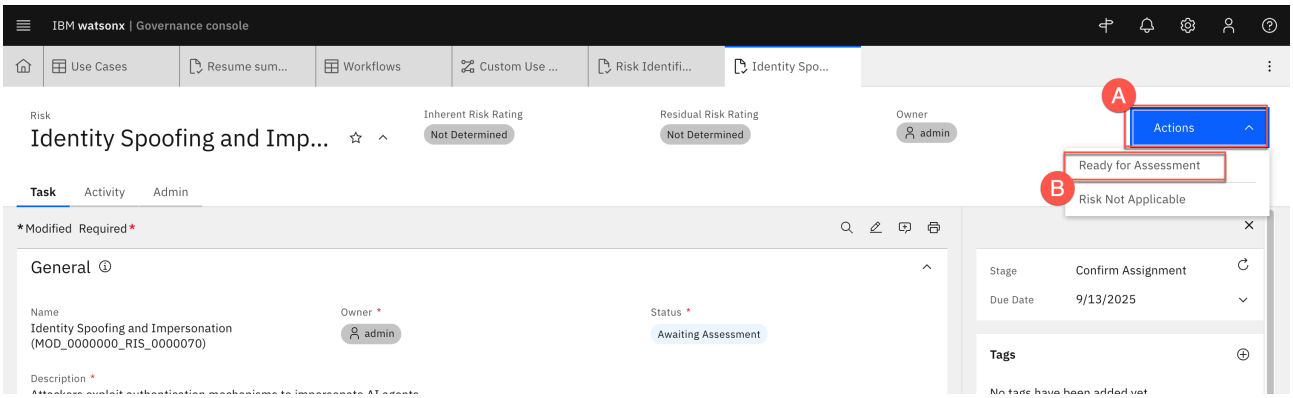


2. Scroll down to the Related Content section of the page, and note that the risk can be associated with mitigating controls, processes, or other issues. Take a moment to inspect some of the other sections on the page, including Internal Audit Risk Rating, and note how risks can be customized based on the threat they pose to a client's business.

3. Click on the Action button in the upper right (A)) to open the Actions menu. Click on the Start model risk assessment menu item (B). A confirmation dialog opens. Click the Continue button to begin assessing the risk. The risk assessment form opens.

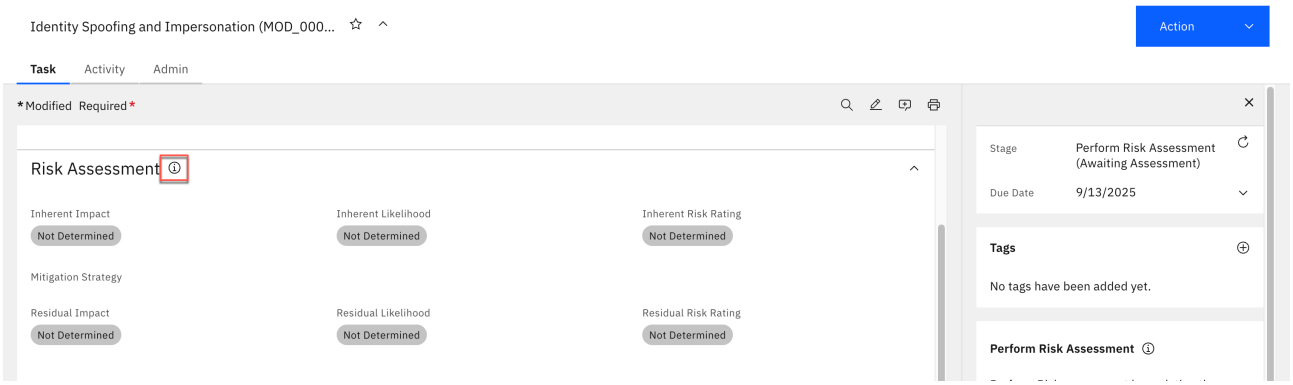


4. Click on the Actions button (A) to open the Actions menu. Click on the Ready for Assessment menu item. Click on the Continue button to confirm your choice and begin the assessment.



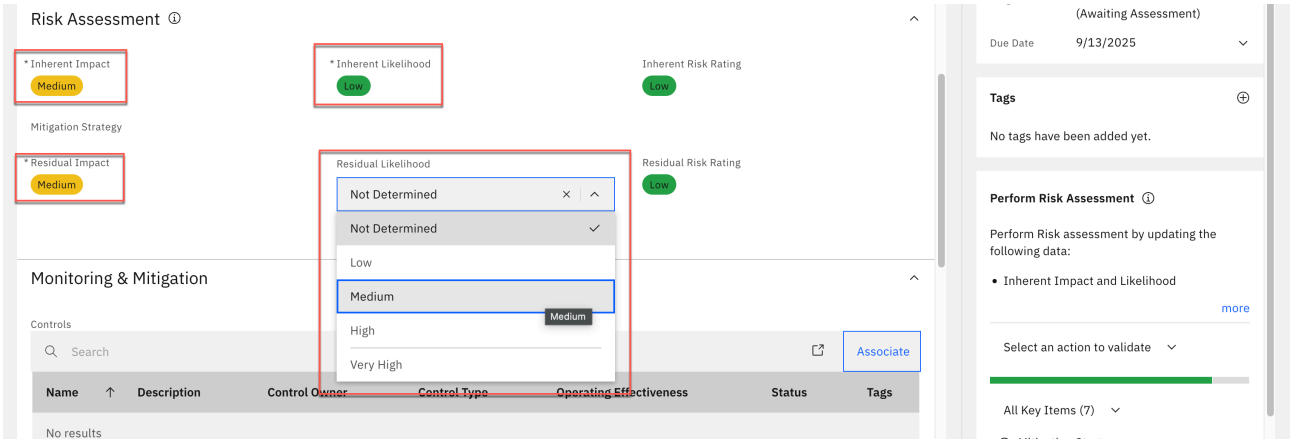
> ⚠ The assessment view is available to your user because you added it to the available views for MRG users in the User management lab.

5. Scroll down to the Risk Assessment portion of the page and click on the information icon next to the session header to open the Field Guidance window. Take a moment to read the descriptions of what each field represents.

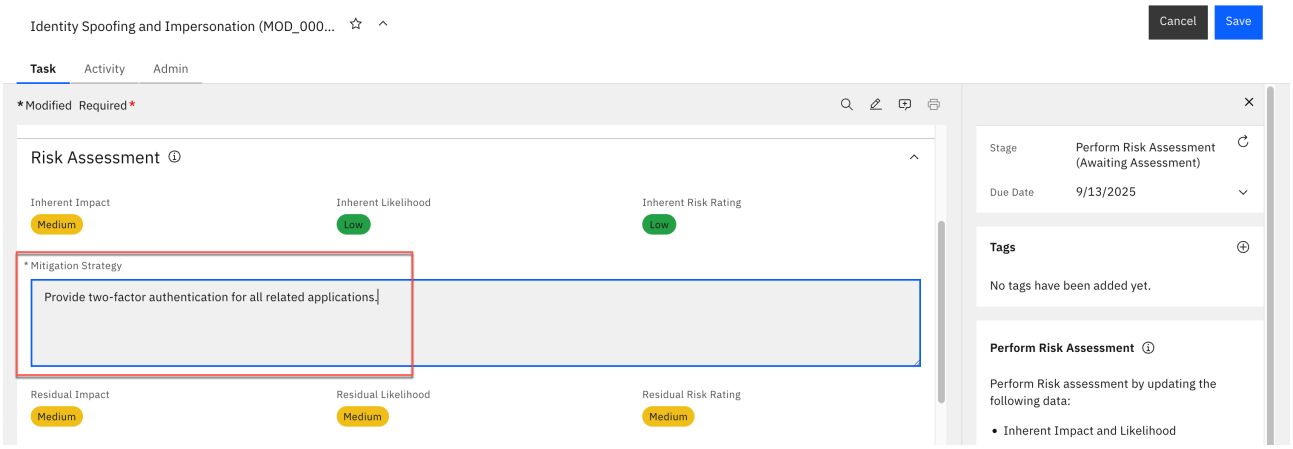


6. Close the Field Guidance window by clicking the X button in the upper right corner of the popup.

7. Hover your mouse over each field to make the edit icon appear. Click on the edit icon for each field and assign a rating.

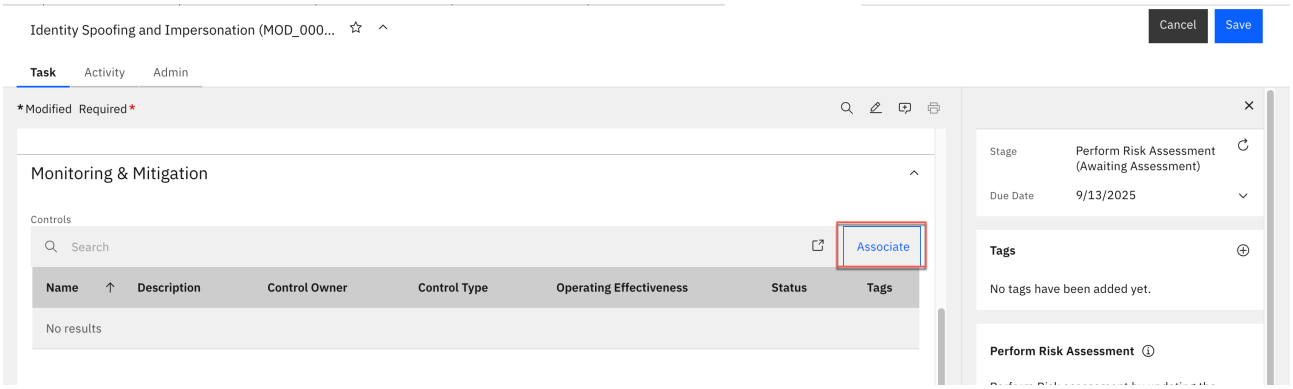


8. Finally, before completing the assessment, you will need to specify a mitigation strategy. Hover your mouse over the Mitigation Strategy field to make the edit icon appear, then click on the edit icon and enter text representing how the organization could mitigate this particular risk.

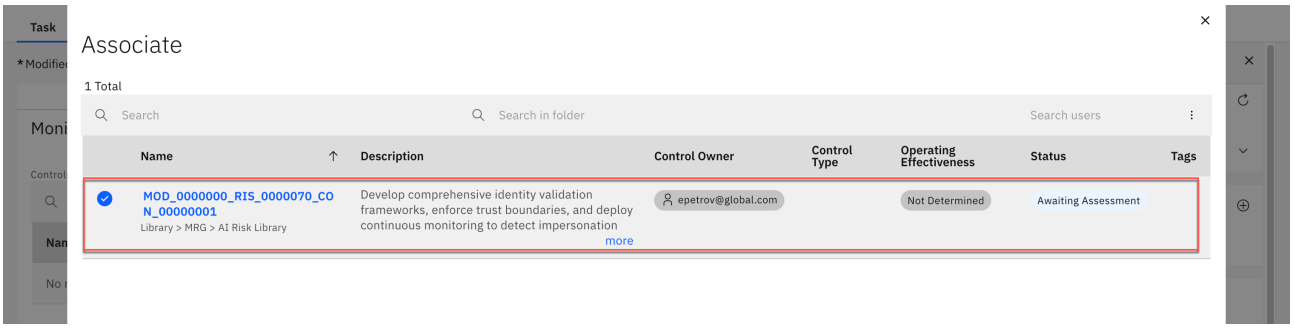


Next, you can add the mitigating control you created in the Questionnaires, risks and controls lab.

9. Scroll down to the Monitoring & Mitigation section of the page and click on the Associate button. The Associate window opens.

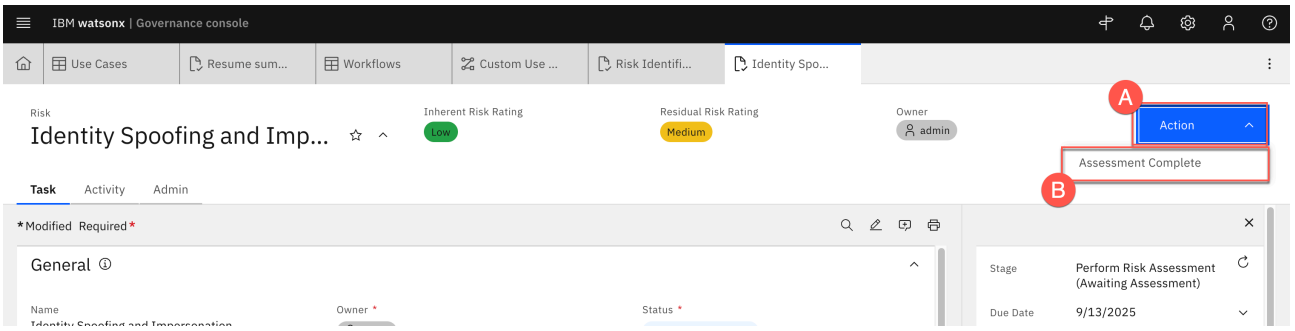


10. Click on the control to select it from the table, then click on the Done button to associate it with the risk and close the Associate window.



11. Click on the Save button in the upper right to save the changes to the risk assessment.

12. Click on the Action button in the upper right (A) to open the Actions menu. Click on the Assessment Complete menu item (B) to finish the risk assessment.



12. When asked to confirm your choice, click on the Continue and close tab button.

13. Return to the use case view, either by clicking on the tab or locating it from the inventory.

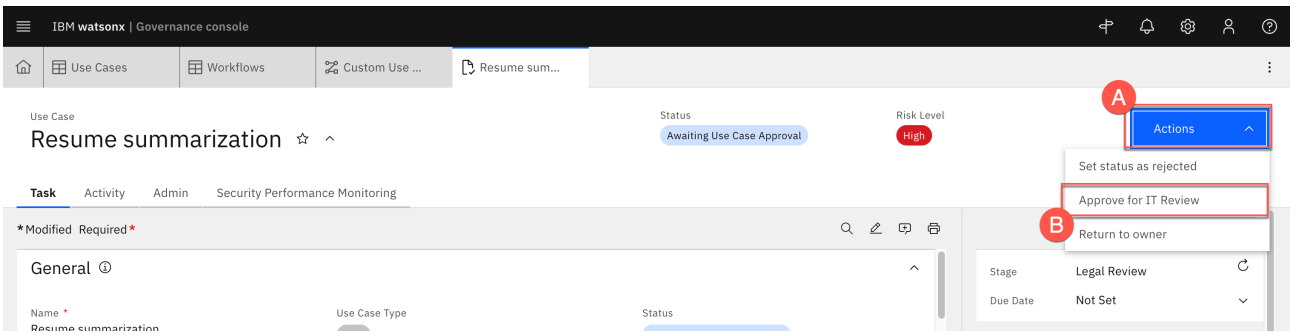
14. Scroll down to the Risks section, and note that the Inherent Risk Rating, Residual Risk Rating, and Status have been updated in the table.

The risk assessments are now complete. You can progress the use case to the next phase.

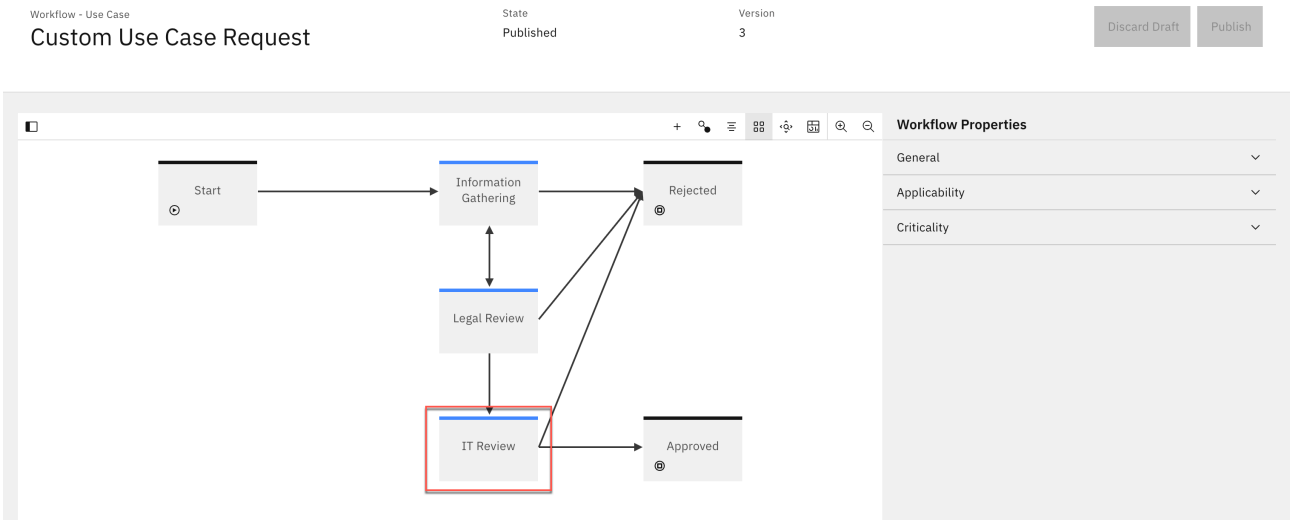
7. Approve the use case for development

Now that the risks have been identified and assessed, the use case can be approved for the next stage of the lifecycle.

1. The use case is now ready to be progressed to the next stage of the workflow. Click on the Actions button in the upper right (A). The Actions menu opens. Click on the Approve for IT Review menu option. A confirmation dialog opens.

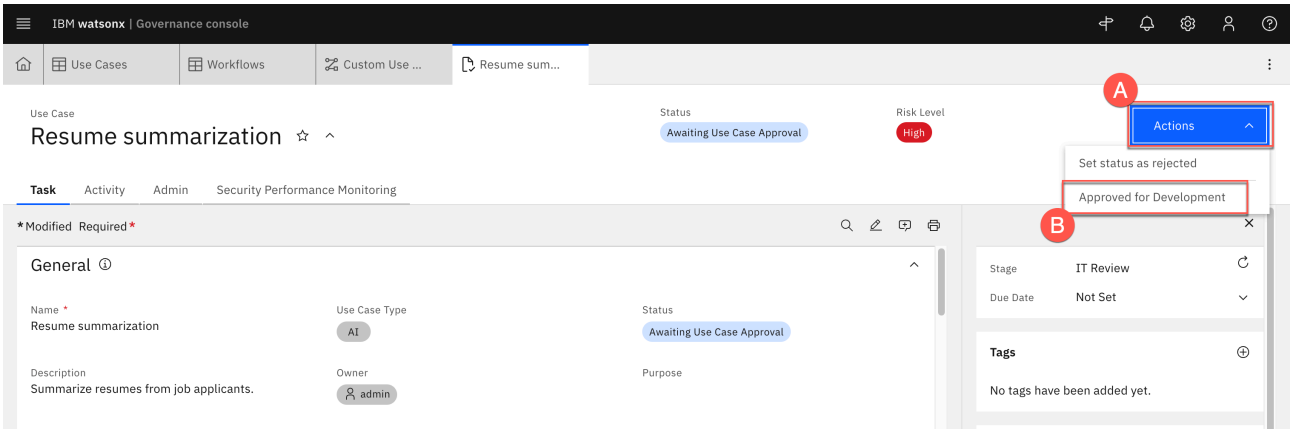


2. Click on the Continue button to confirm your choice. The use case progresses to the IT Review stage. Once again, the graphic below is informational, and will not be shown on your screen.



At this point in the process, the IT department would review the use case, including the answers provided in the risk identification questionnaire.

3. Click on the Actions button (A) once more. The Actions menu opens. Click on the Approved for development menu item (B) to approve the use case. A confirmation dialog opens.



4. Click the Continue button to confirm your choice. The Status field changes to Approved for Development.

If you click on the Action button again, you will see options to progress the use case to the next stage of its lifecycle. However, for the purposes of this lab, the focus will shift away from progressing the use case view to the development and monitoring of models. Feel free to update the workflow and continue progressing the use case through the different phases if your client would like to see the entire process.

At this point in the lifecycle, the model use case has been created, reviewed for risks, and approved by the various stakeholders. Personas involved are mostly non-technical, from the business user who requested the model to the risk and compliance officer who evaluated it. Next, the model would be developed by teams of data scientists and AI engineers. The following steps of the lab will take actions from the point of view of those personas.

Develop the prompt

1. Create the prompt template

In this case, the AI engineers have elected to work with the Azure OpenAI service on a prompt template to summarize the resumes.

⚠ THESE EVALUATIONS ARE NOT INTENDED TO SHOW THE RELATIVE STRENGTHS OF THE OPENAI OR AZURE PLATFORMS, AND SHOULD NOT BE PRESENTED AS SUCH.

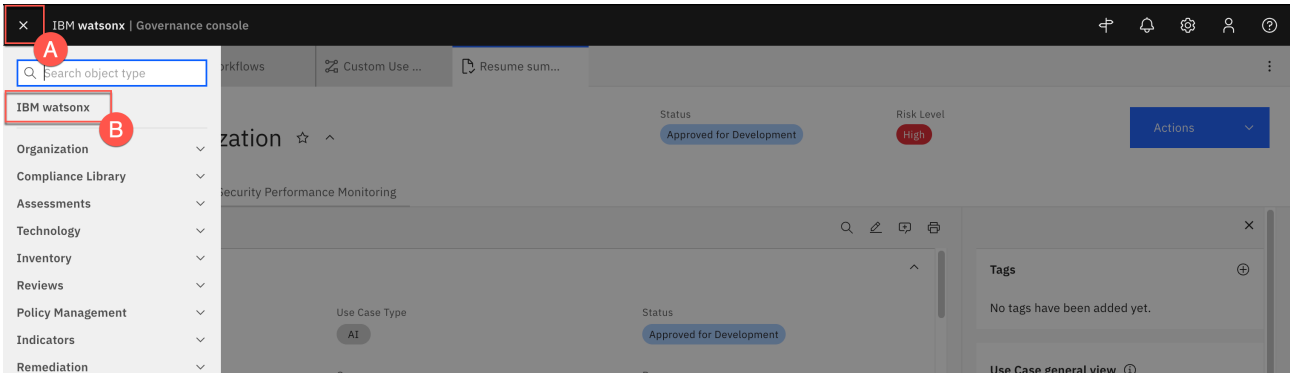
The prompt used in this lab is fairly simple, and in a real-world scenario would be tuned and optimized for the individual use case. The evaluations here are presented to show how the watsonx.governance platform can collect facts and metrics for hybrid environments with models deployed on any platform.

ⓘ Watsonx.governance supports the evaluation of third-party generative models via a method known as detached prompt templates, which are generative AI models not hosted on the same platform as the watsonx.governance service. At the time of writing, working with detached prompt templates is done through the use of Jupyter notebooks.

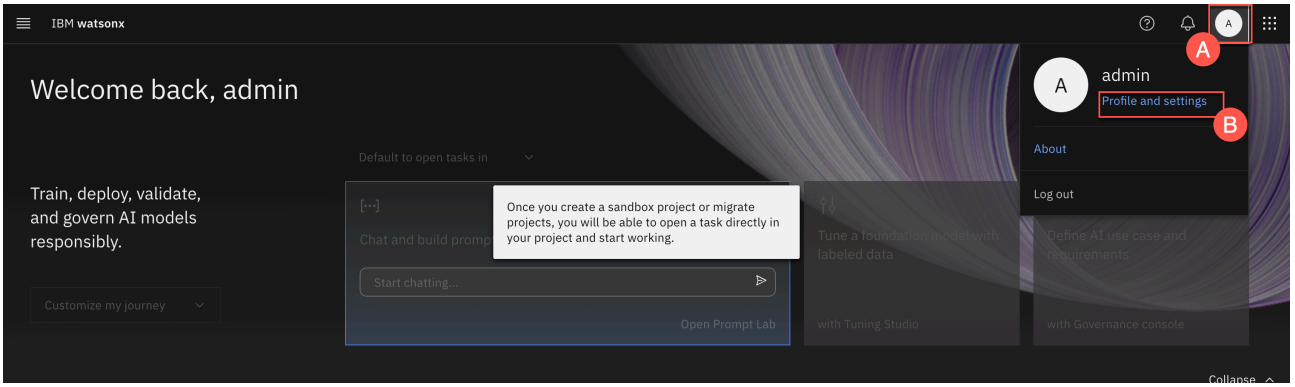
To begin, you will need to gather credentials used by the notebook. From watsonx, you will need the base Cloud Pak for Data URL, as well as the username and password of the created user. You will also create an API key for the user.

For Azure, you will need the API Endpoint, API key, name of the deployed model, Client ID, and Client Secret that you gathered at the beginning of this lab.

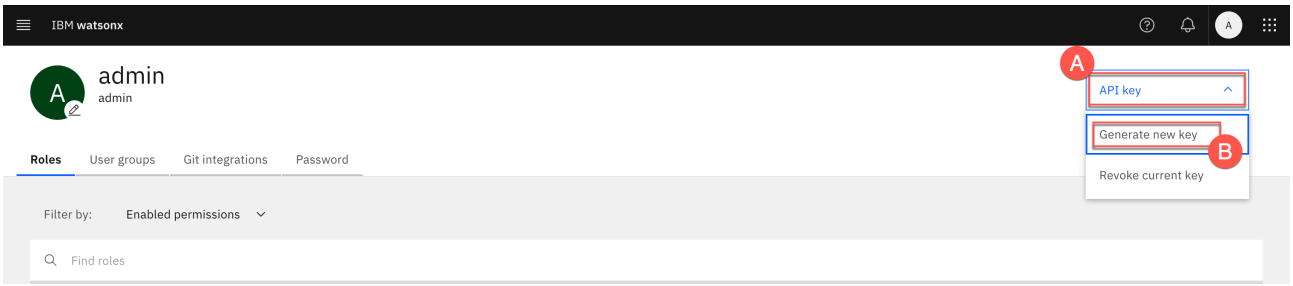
1. From the watsonx governance console, click on the Primary menu in the upper left (A). Click on the IBM watsonx menu item (B). The watsonx home page opens in the watsonx context (as opposed to the Cloud Pak for Data context).



2. Click on the avatar icon in the upper right (A) to open the user menu. Click on the Profile and settings item from the menu (B). The user profile screen opens.

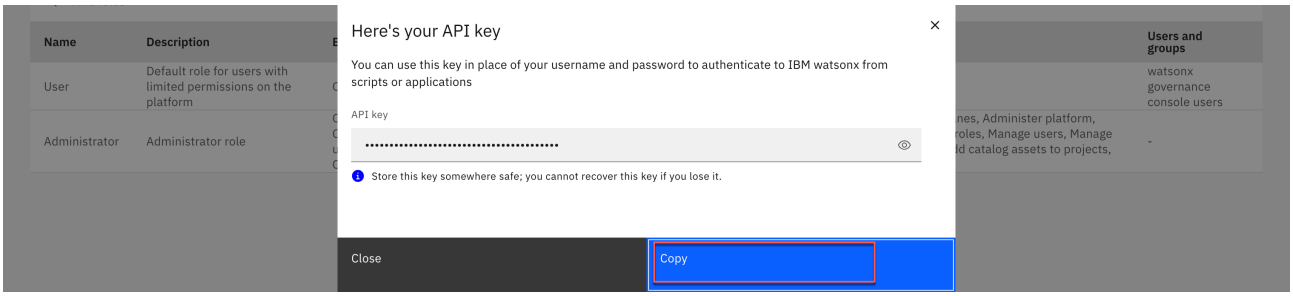


3. Click on the API key button in the upper right (A). The API key menu opens. Click on the Generate new key menu item (B). The Generate new API key? dialog window opens.



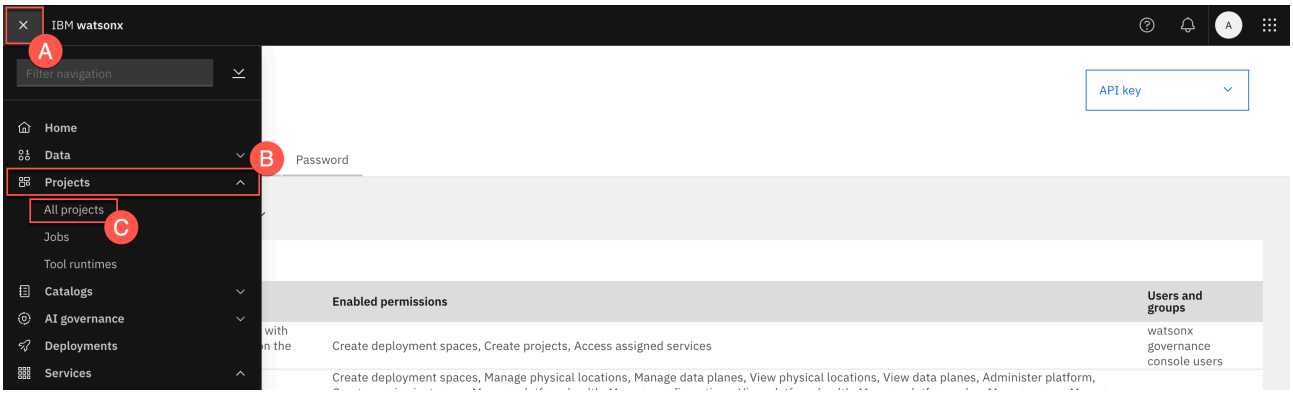
4. Click the red Generate button to confirm API key creation. Note that, as the warning states, generating a new key will invalidate any existing keys you have.

5. Click the Copy button to copy your new key to the clipboard. Paste it into a text file for later use in the notebook, where it will represent the CPD_API_KEY value.



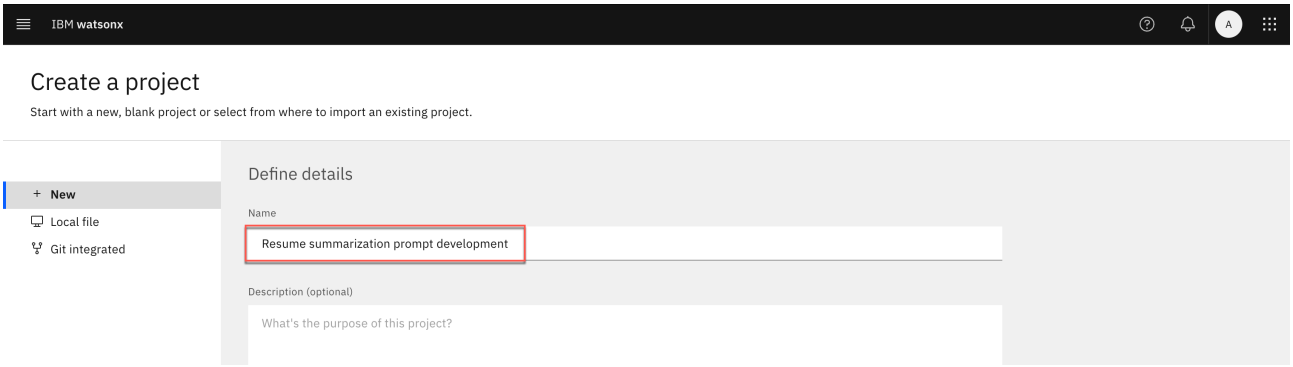
6. Once you have pasted the key into a text file, click the Close button to close the window.

7. Click on the navigation menu in the upper left (A) to open the menu. Click on the Projects menu item (B) to expand it. Click on the All projects menu item (C). The Projects screen opens.



8. Click on the New project button. The Create a project screen opens.

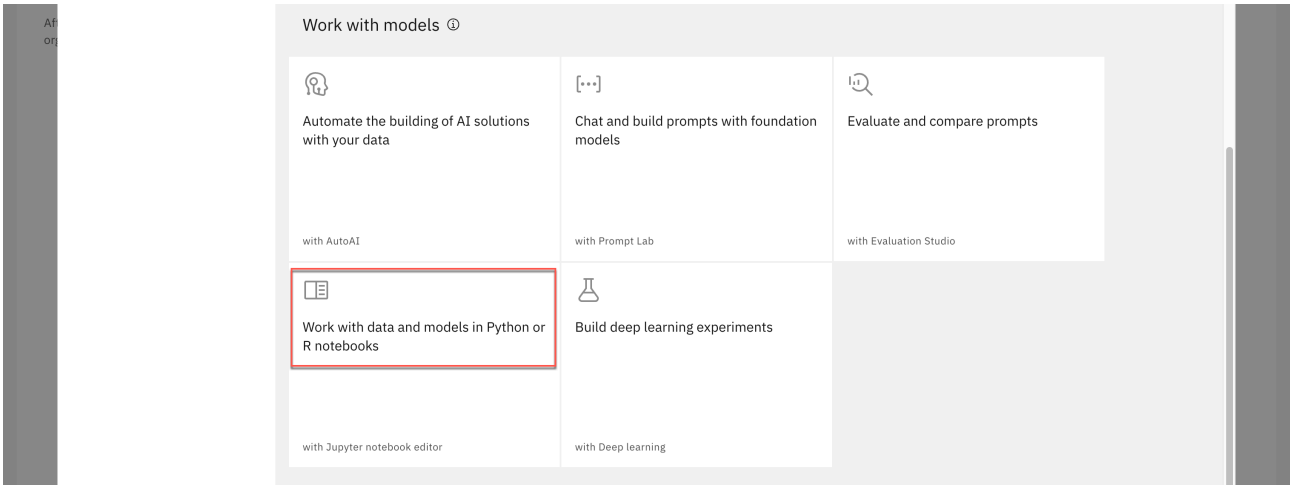
9. Give your project a Name.



10. Click the Create button. Your project will be created.

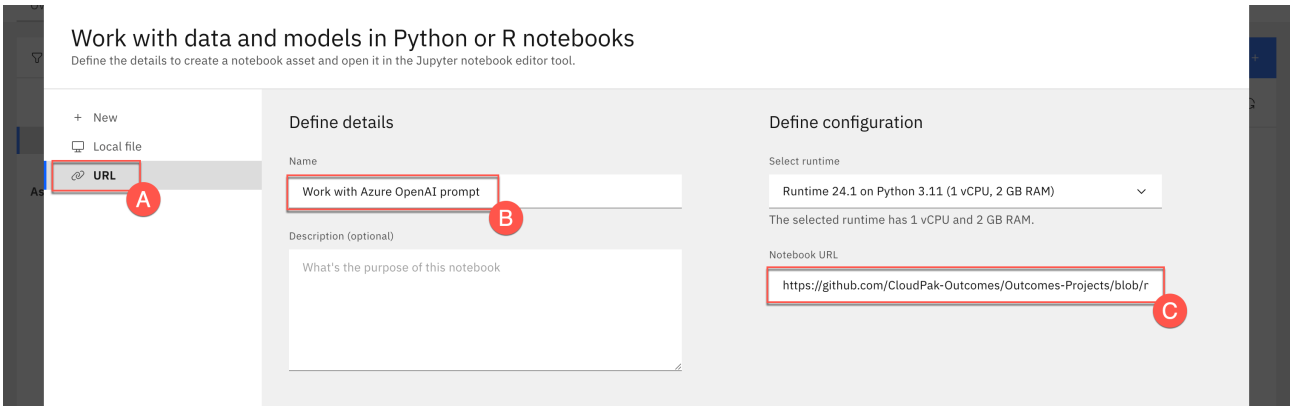
11. Click on the Assets tab (A). Click on the New asset button (B). The What do you want to do? window opens.

12. Locate and click on the Work with data and models in Python or R notebooks tile.



13. Click on the URL option (A). Give your notebook a Name (B). Copy and paste the following URL into the Notebook URL field (C):

```
https://github.com/CloudPak-Outcomes/Outcomes-Projects/blob/main/watsonx-governance-l4-deploy/governance/create_prompt_template_51.ipynb
```



14. Click on the Create button. Your notebook will be created.

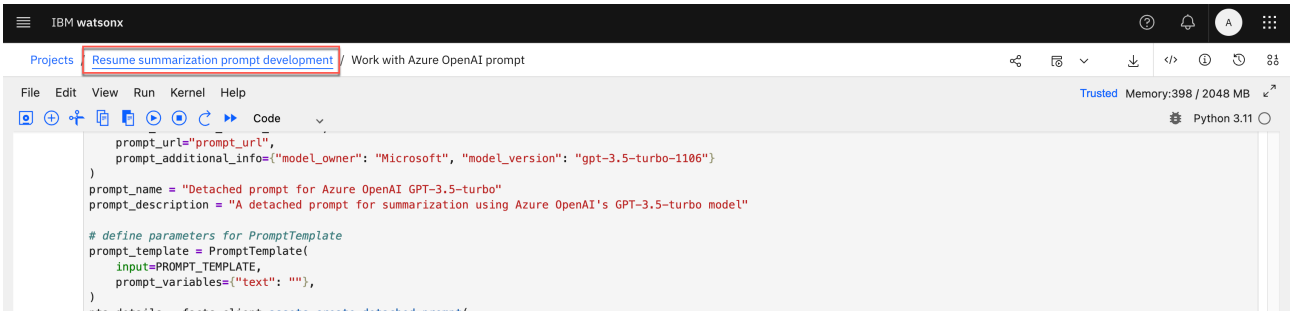
15. Scroll down to the first code cell and edit the string values. Your CPD_USERNAME will be the user that you are logged in as (admin if you have followed the lab instructions exactly). The CPD_PASSWORD will be the password for that user. The CPD_API_KEY is the API key you generated in a previous step.

The Azure credentials can all be found in your TechZone reservation for the Azure service that you created in the environment setup portion of this lab.

```
[18]: import os
from rich import print
from IPython.display import display, Markdown

CPD_USERNAME = "admin"
CPD_PASSWORD = "VIfjI                    nWt"
CPD_API_KEY = "OBQt                                qnutb"

AZURE_OPENAI_ENDPOINT = "https://azureml-openai-americas-1.openai.azure.com/"
AZURE_OPENAI_DEPLOYMENT_NAME = "tz-gpt-35-turbo-americas-1"
AZURE_CLIENT_ID = "998e0665                              :3c5"
AZURE_CLIENT_SECRET = "84E8                              :eqUImcOObdM"
AZURE_TENANT_ID = "4e7730a0                              :1b7"

CPD_URL = os.environ.get('RUNTIME_ENV_APSX_URL')
PROJECT_ID = os.environ.get('PROJECT_ID', "<YOUR_PROJECT_ID>")
print(f"Your project id is '{PROJECT_ID}' and your environment URL is {CPD_URL}")
```

16. Run through the code cells in the notebook one at a time. The notebook will connect to the OpenAI model, use it to perform resume summarization on sample resumes, and finally save the prompt template to your project. It will also save the summaries to a CSV file in your project that you will use to evaluate the template's performance.

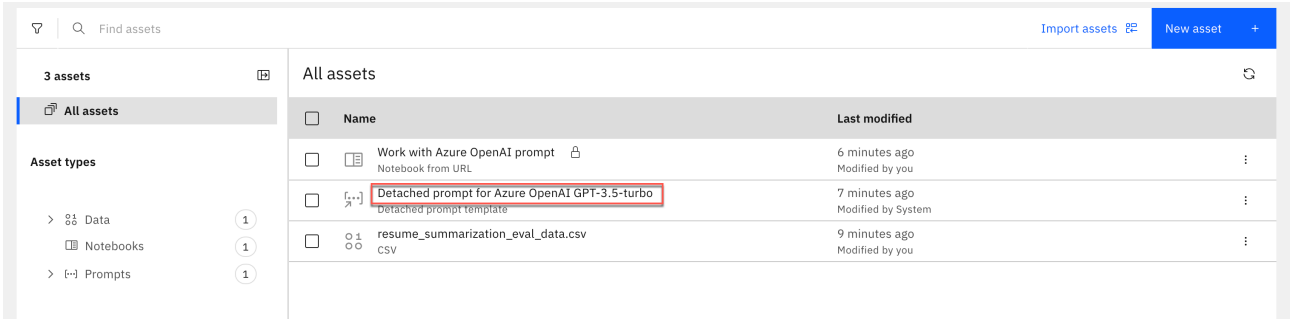
17. Click on project breadcrumb link at the top of the screen to navigate back to your project.



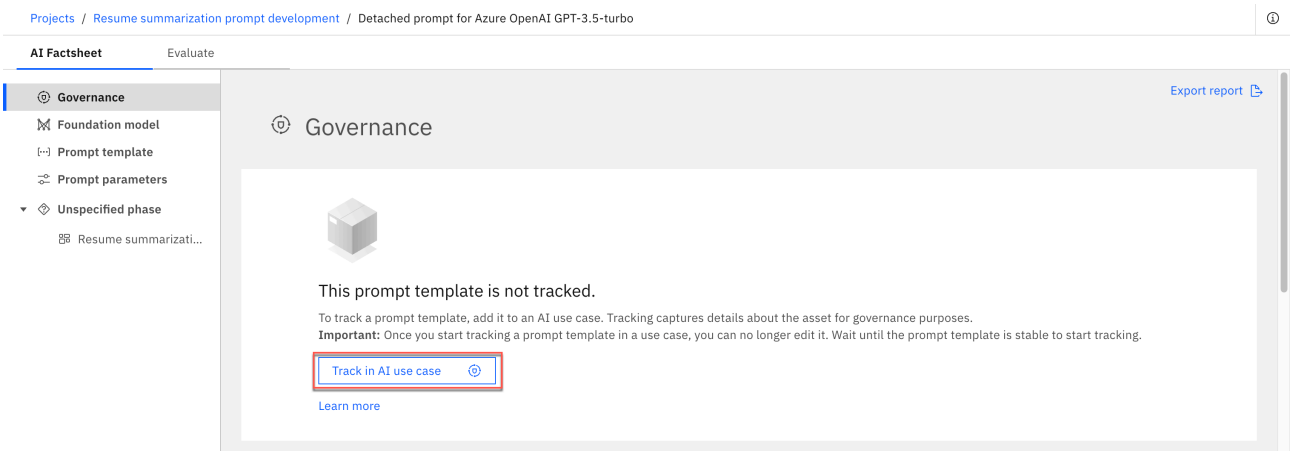
2. Associate workspaces

Next, you will associate the prompt template with the use case that you took through the approval process.

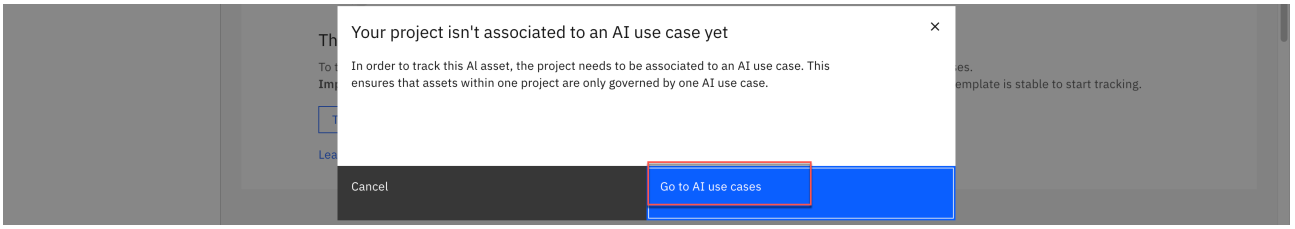
1. From the Assets tab in your project, click on the Detached prompt for Azure OpenAI... asset. Click on the X button to close the Learn about your AI asset window.



2. Click on the Track in AI use case button. A popup window will open, informing you that your project has not been associated with an AI use case yet.

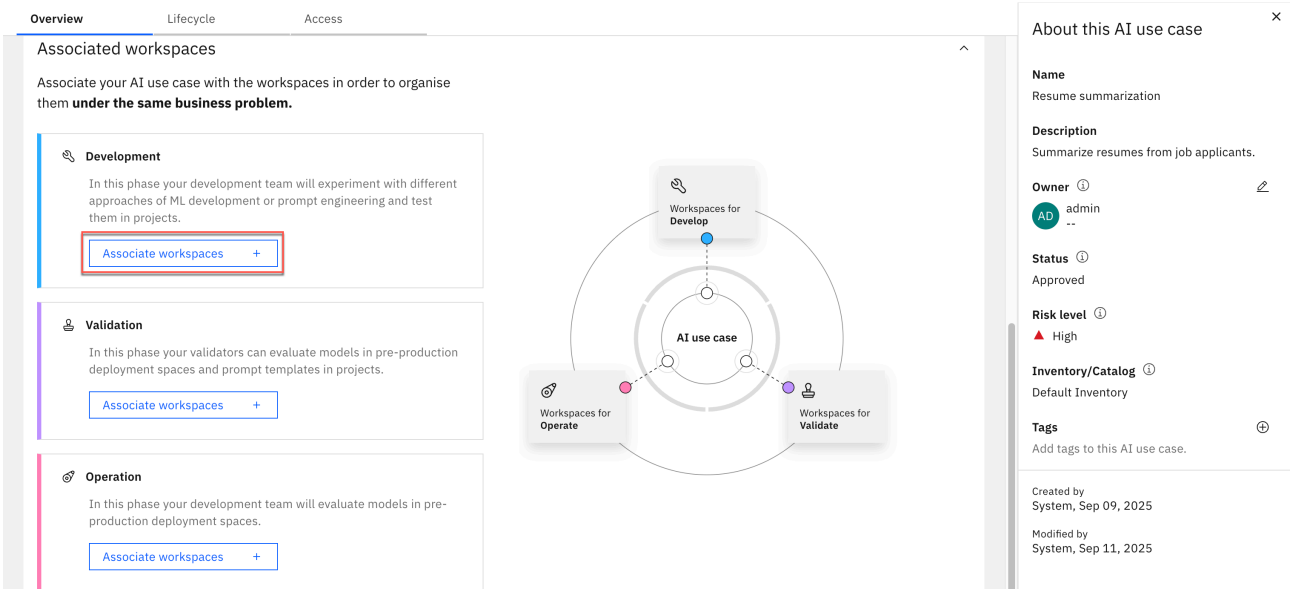


3. Click on the Go to AI use cases button. A new browser tab will open, listing all the use cases in the inventory.

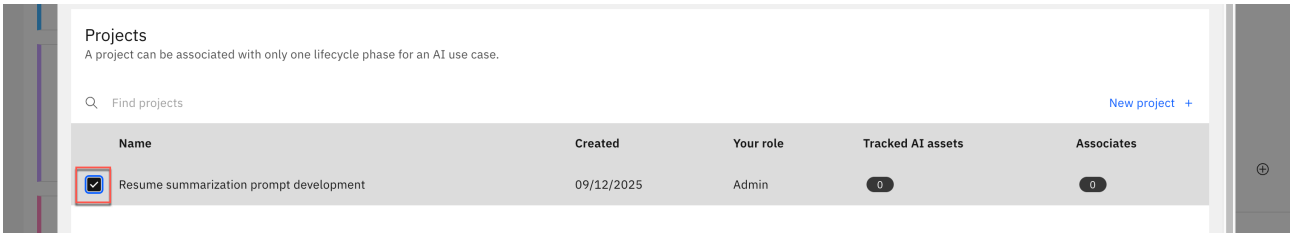


4. Click on the Resume summarization use case from the list. The use case opens in your browser. Note that this view differs from the view of the same use case in the governance console. The governance console view contains much of the same information, but presented for a different audience (risk managers and subject-matter experts). The current view is targeted more at AI engineers and data scientists.

5. Scroll down to the Associated workspaces section of the use case entry. Note that there are three stages of a use case lifecycle: Development, Validation, and Operation. Workspaces such as projects and deployment spaces can be associated with these phases. Because the project you created to work with the prompt represents an early attempt to use AI to solve a business use case, you will associate it with the Development phase. Click on the Associate workspaces button in the Development tile.



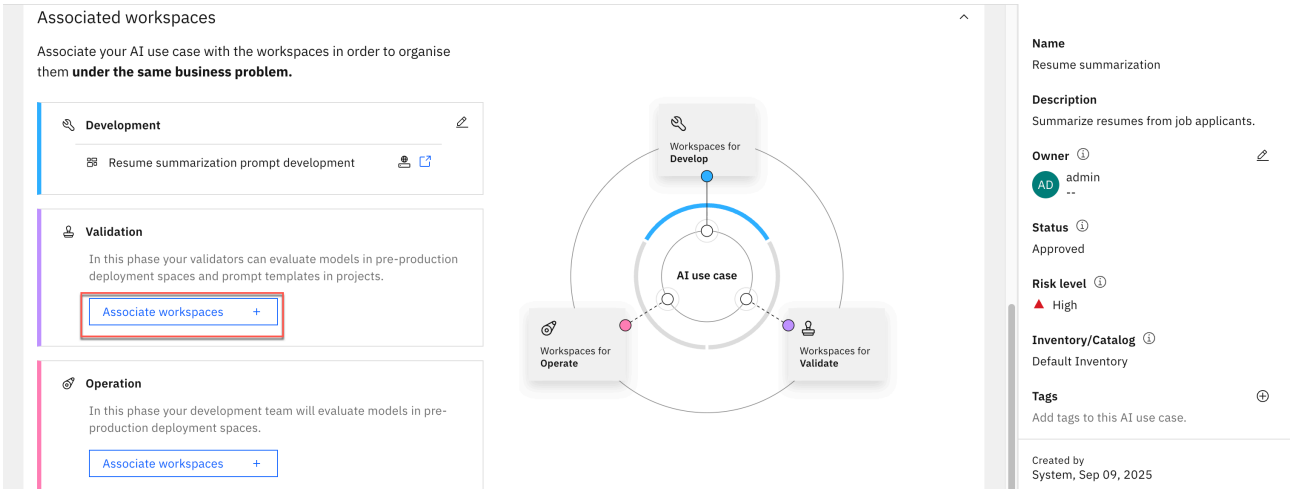
6. Locate the Projects section of the Associated workspaces window, and check the box to the left of your project.



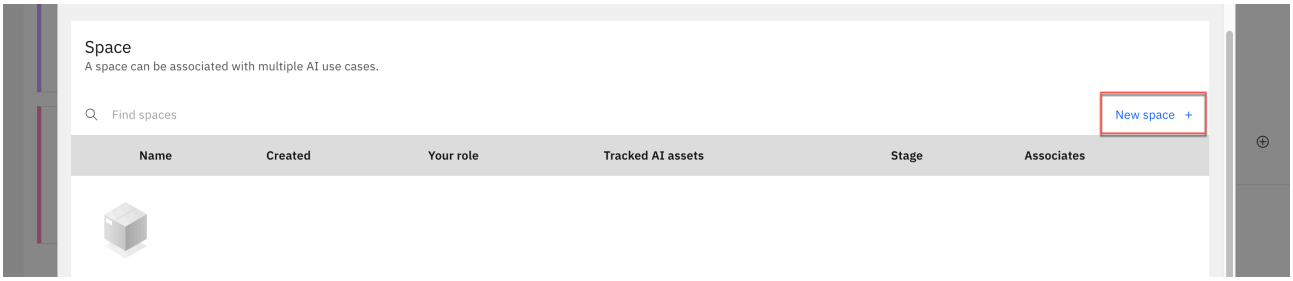
7. Click on the Save button to save the association and close the window.

The project has been associated with the use case. In the next steps, you will enable tracking of the detached prompt template. However, for the sake of convenience, you will first create and associate a deployment space for evaluating the prompt.

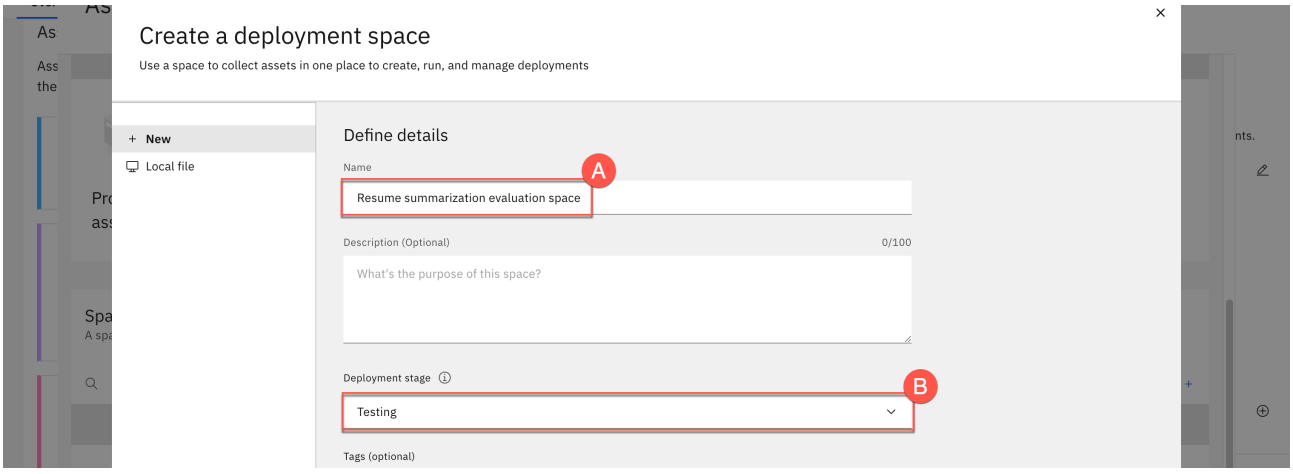
8. Click on the Associate workspaces button in the Validation tile. The Associate workspaces window opens again.



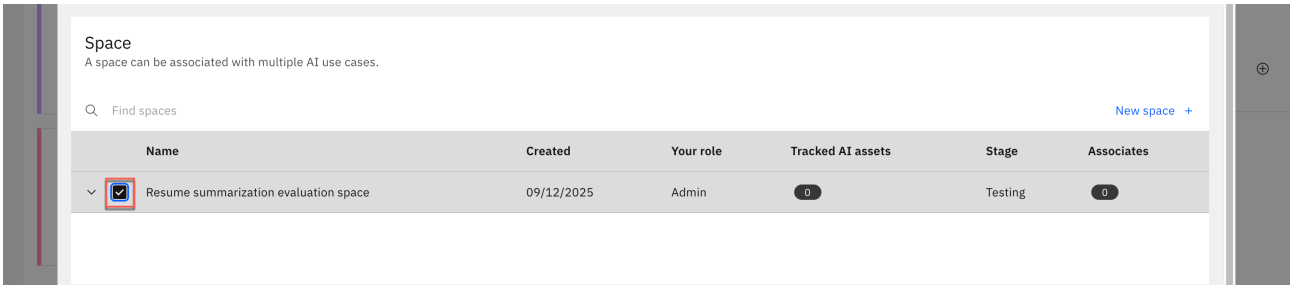
9. Scroll down to the Space section of the window and click on the New space button. The Create a deployment window opens.



10. Give your space a name (A). Click on the Deployment stage dropdown and select the Testing option (B).

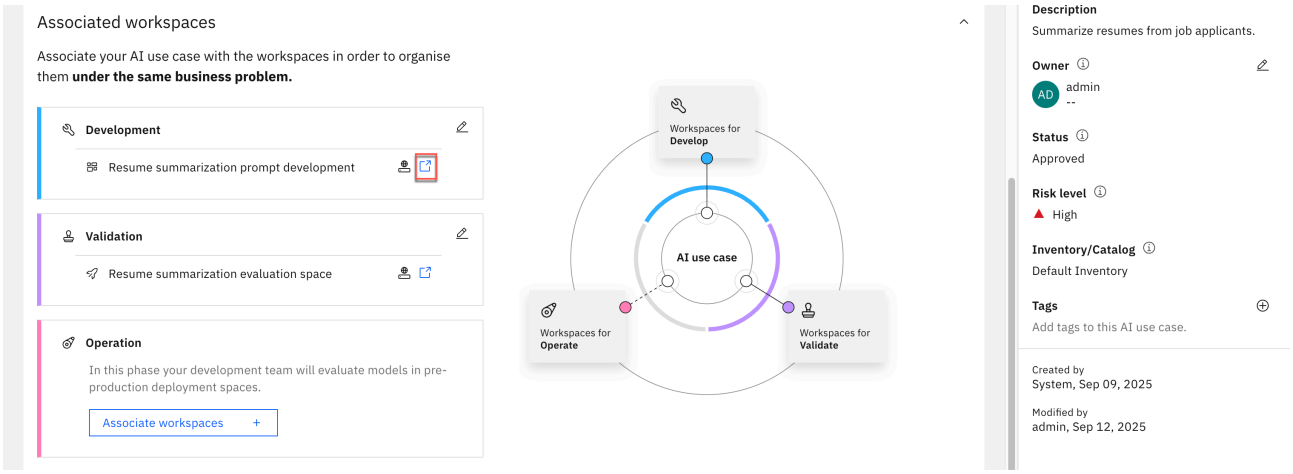


11. Click on the Create button to create your space and close the window. Check the box to the left of your newly-created space and click on the Save button to associate it with the use case and close the window.



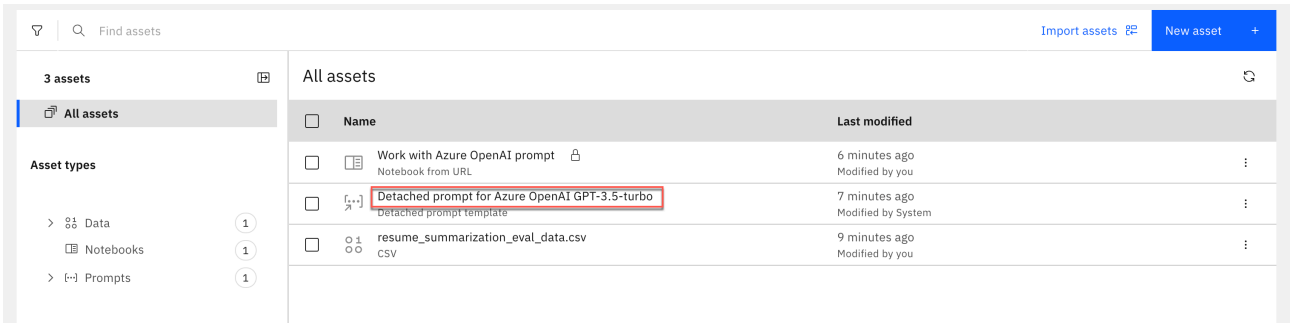
Your use case screen should now show a project associated with the Development phase of the lifecycle, and a deployment space associated with the Validation phase of the lifecycle. You can now return to the project and begin tracking the prompt template as part of the use case.

12. Click on the open link for your project in the Development tile. A new browser window will open for the project.

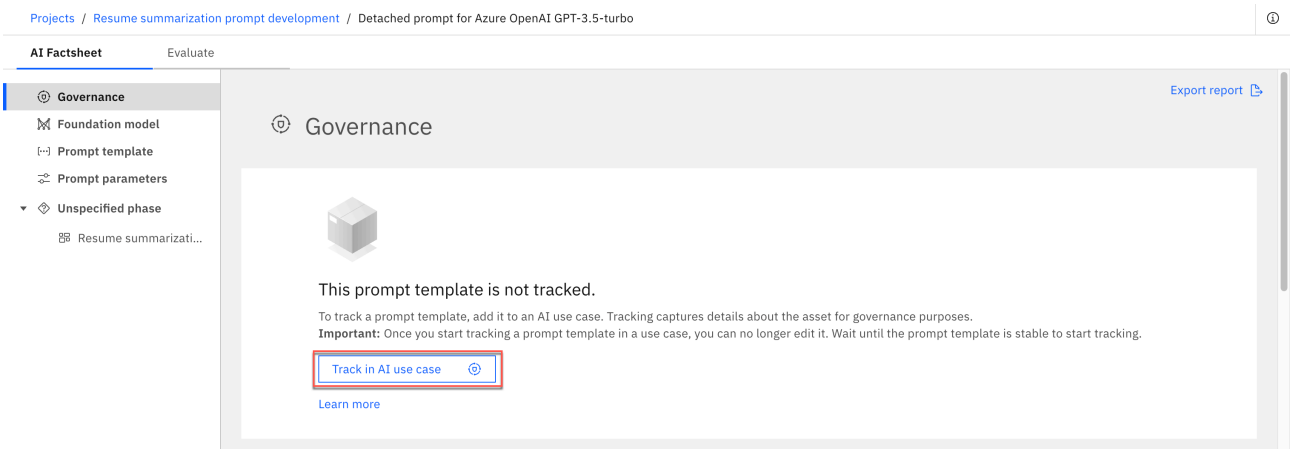


3. Track the prompt in the use case

1. From the Assets tab in your project, click on the Detached prompt for Azure OpenAI... asset. Click on the X button to close the Learn about your AI asset window.

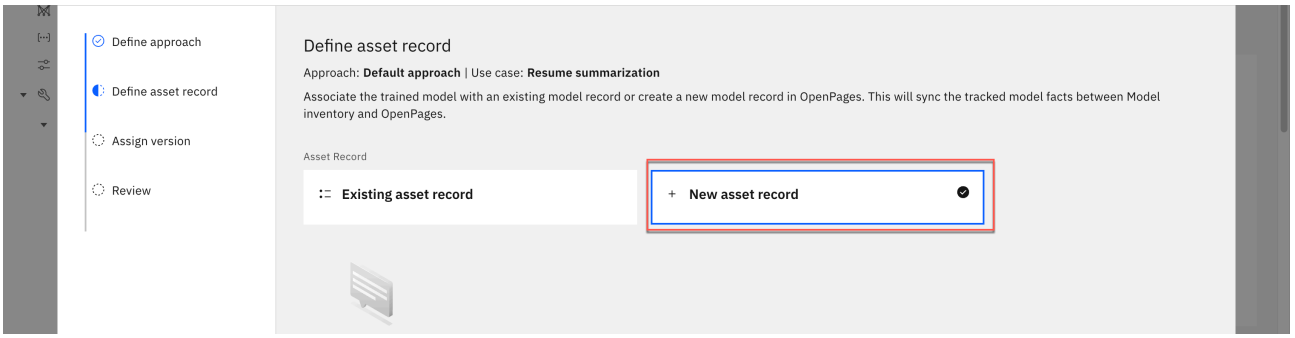


2. Click on the Track in AI use case button. This time, the Track in AI use case window opens to the Define approach screen.



3. Click on the Next button to accept the default approach. The Define asset record screen opens.

4. Click on the New asset record tile to specify that you would like to create a new model in the governance console inventory.

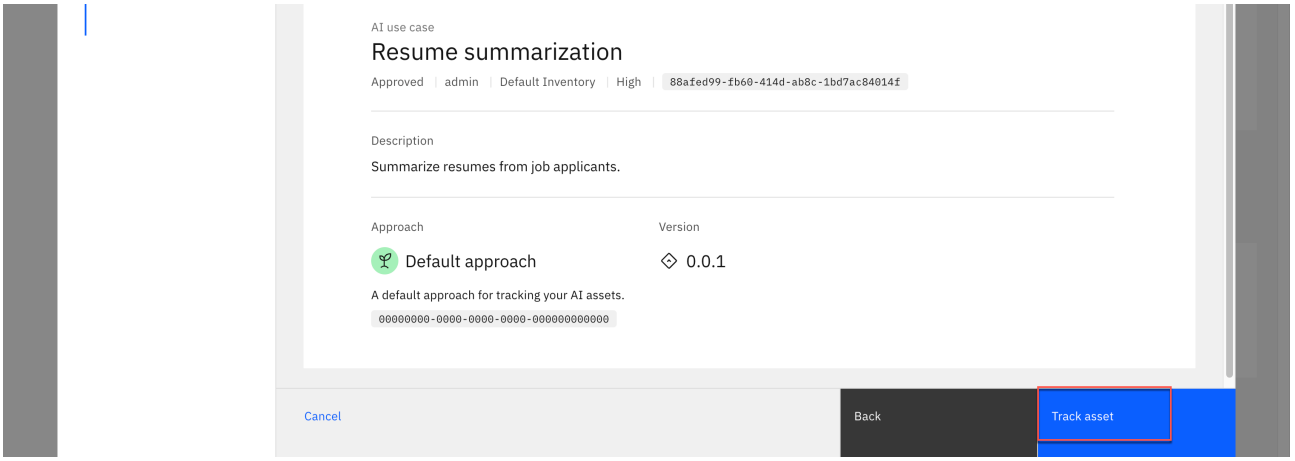


5. Click on the Next button. The Assign version screen opens.

6. Click on the Next button. The Review screen opens.

> ⓘ Note the warning at the top of the screen; once you begin tracking the template in an AI use case, you can no longer edit it.

7. Click on the Track asset button to enable tracking for the model.



> ⚠ Occasionally, slow network conditions may result in an error message at this point telling you that the model is already being tracked. In this case, the tracking request has typically succeeded. Clicking the Cancel button to return to the Factsheet and then refreshing the page will show the model as being tracked within the use case.

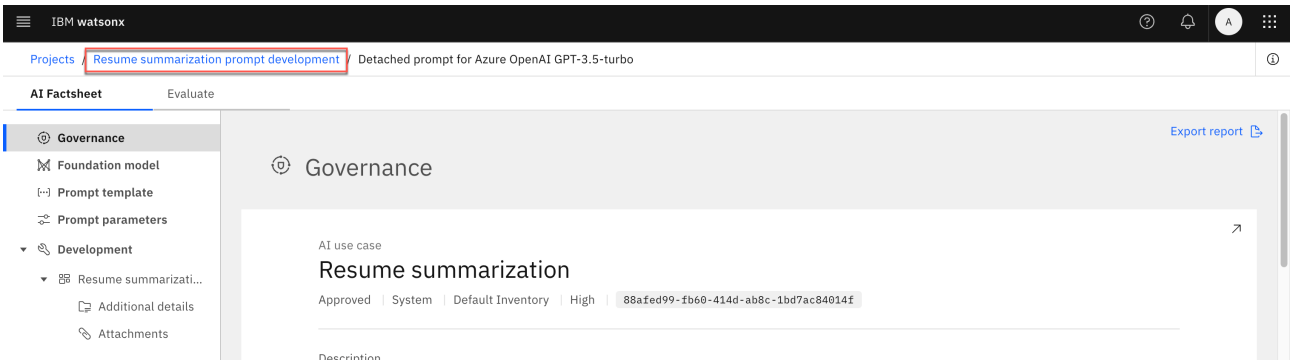
8. Take a moment to review the Factsheet. Note that it contains metadata on the type of model, provider, task, and prompt.

The model has been created in the project, and is being tracked as part of a use case. Next, you will deploy the model to a space for evaluation.

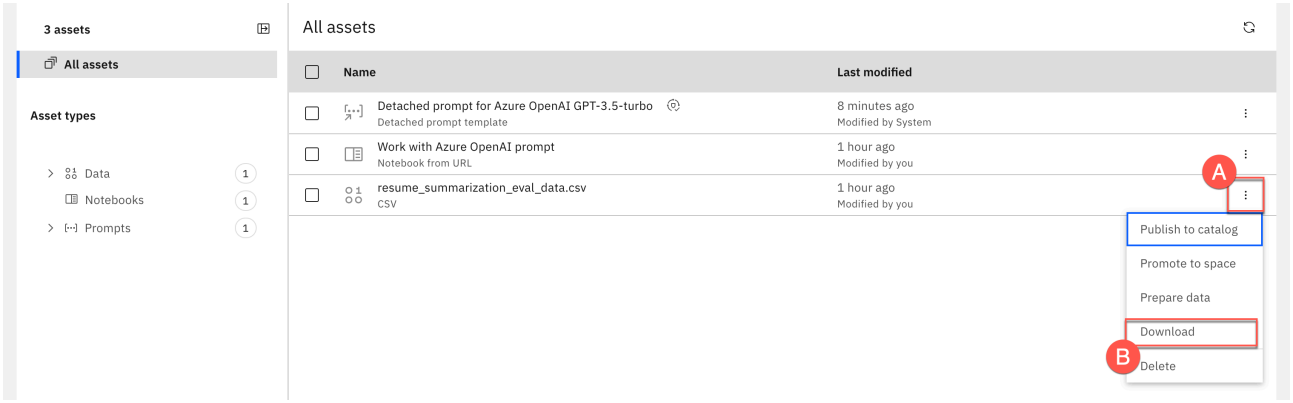
4. Deploy the model to a space

In this step, you will download the output to use in an evaluation, and promote the model to a space.

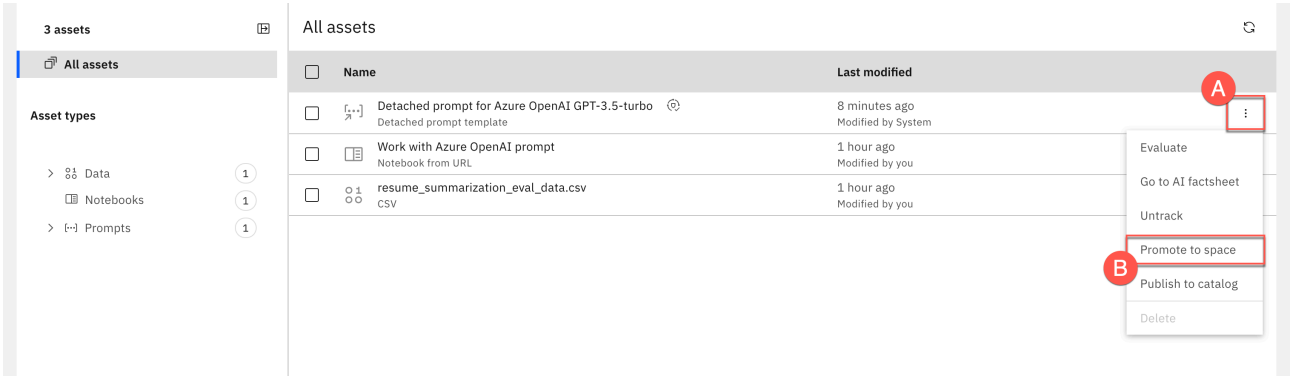
1. Click on your project name from the breadcrumb trail at the top of the Factsheet. The project screen opens.



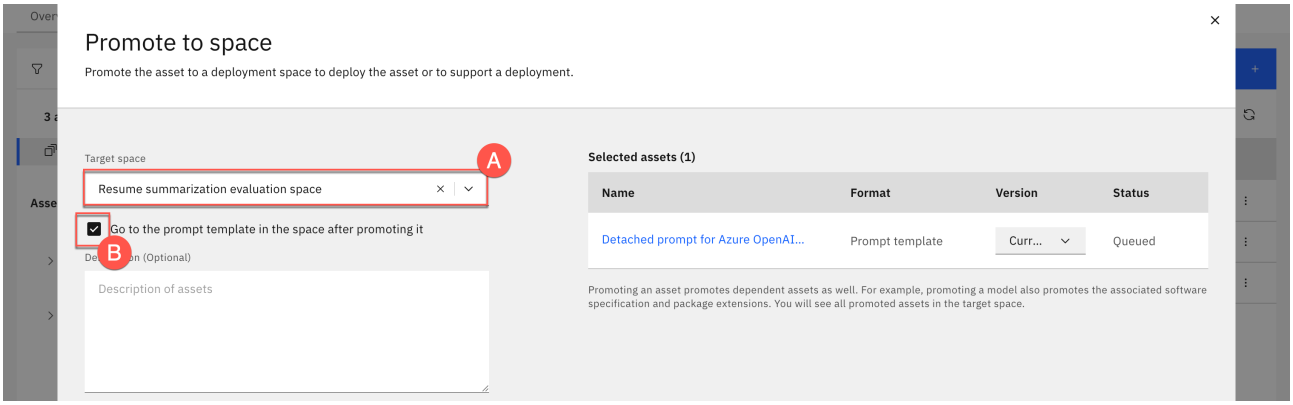
2. From the Assets tab, click on the three vertical dots to the right of the resume_summarization_eval_data.csv file (A) to open the context menu. Click on the Download menu item (B) to download the file to your machine.



3. From the Assets tab, click on the three vertical dots to the right of the Detached prompt template... (A) to open the context menu. Click on the Promote to space menu item (B). The Promote to space window opens.

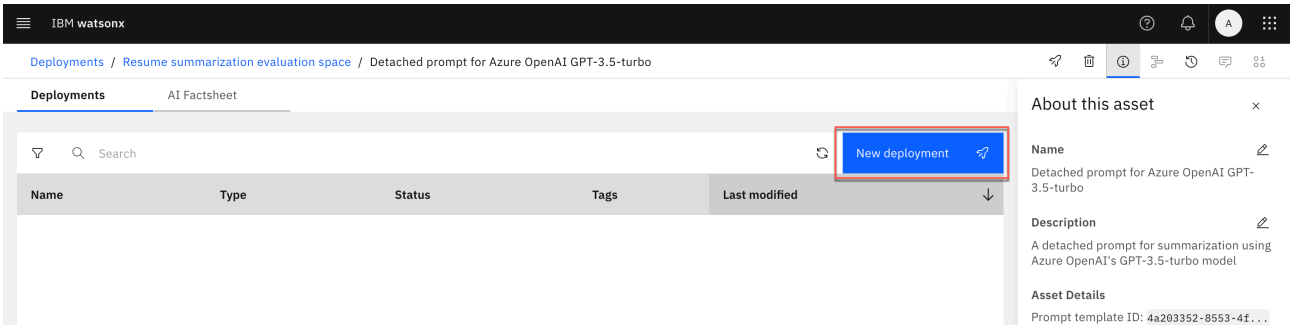


4. Click on the Target space dropdown and select the space you created in the previous step (A). Check the box to the left of Go to the space after promoting the assets (B).

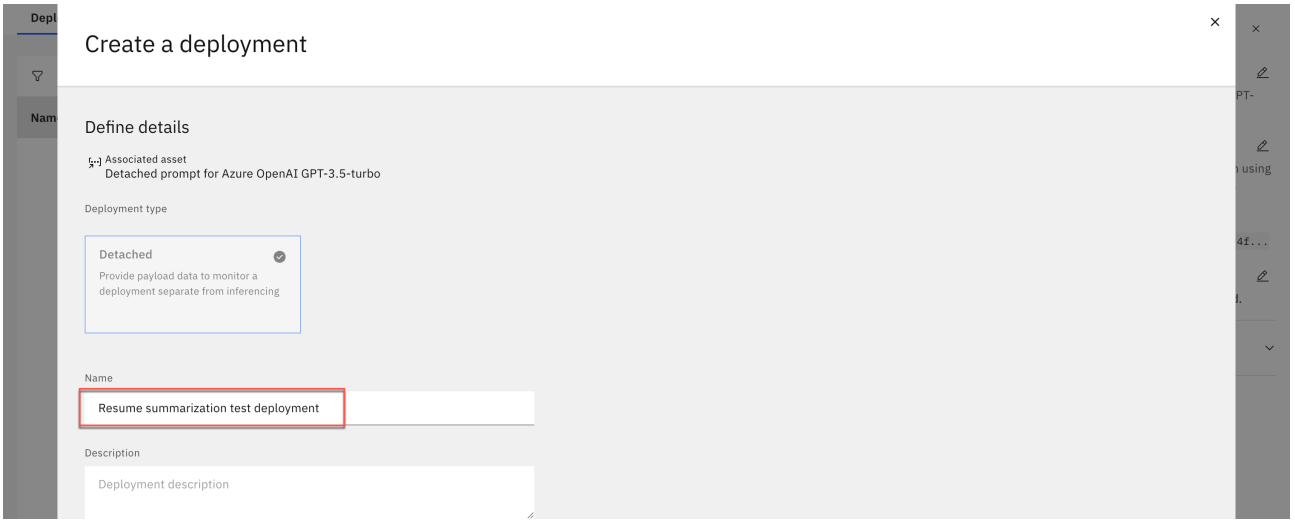


5. Click on the Promote button to promote the prompt template. The prompt will be created in the new space.

6. Click on the New deployment button. The Create a deployment screen opens.



7. Give your deployment a Name.



8. Click on the Create button to create the deployment.

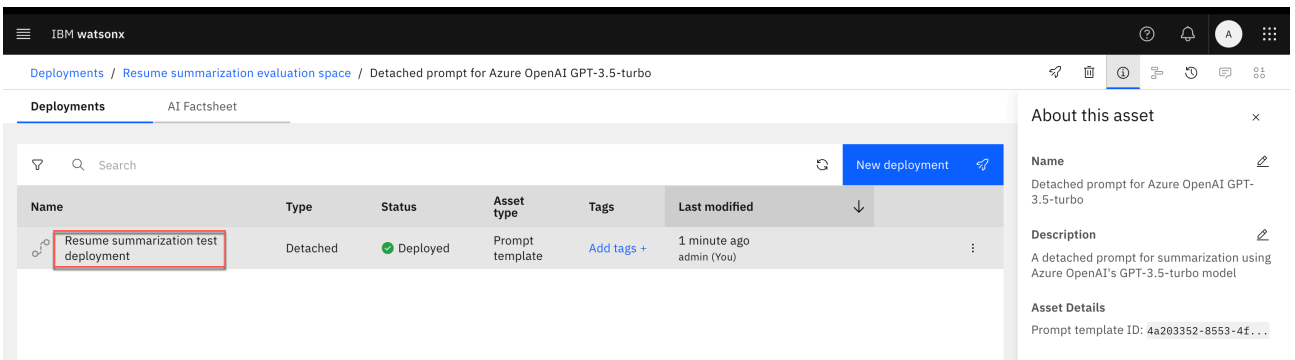
The prompt is now available as a REST endpoint. It can also be evaluated.

Evaluate the prompt

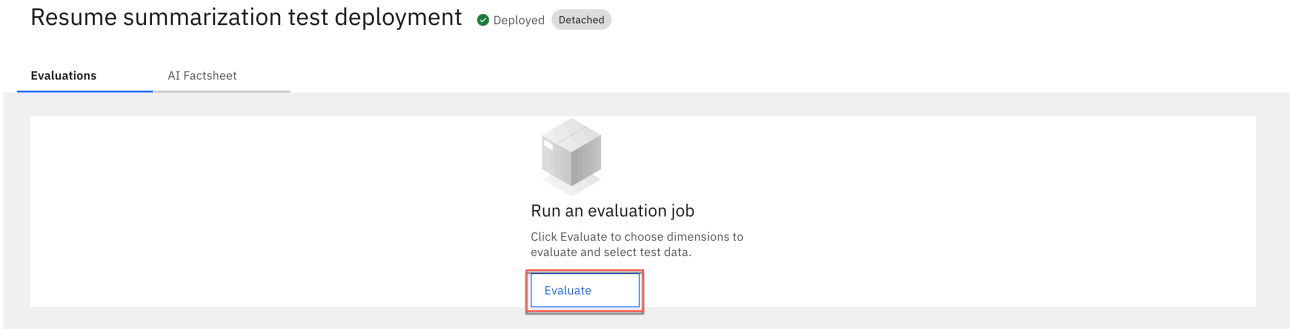
1. Perform a quality evaluation

In this step, you will evaluate the model for quality.

1. Click on the link for your newly-created deployment. The deployment summary screen opens.

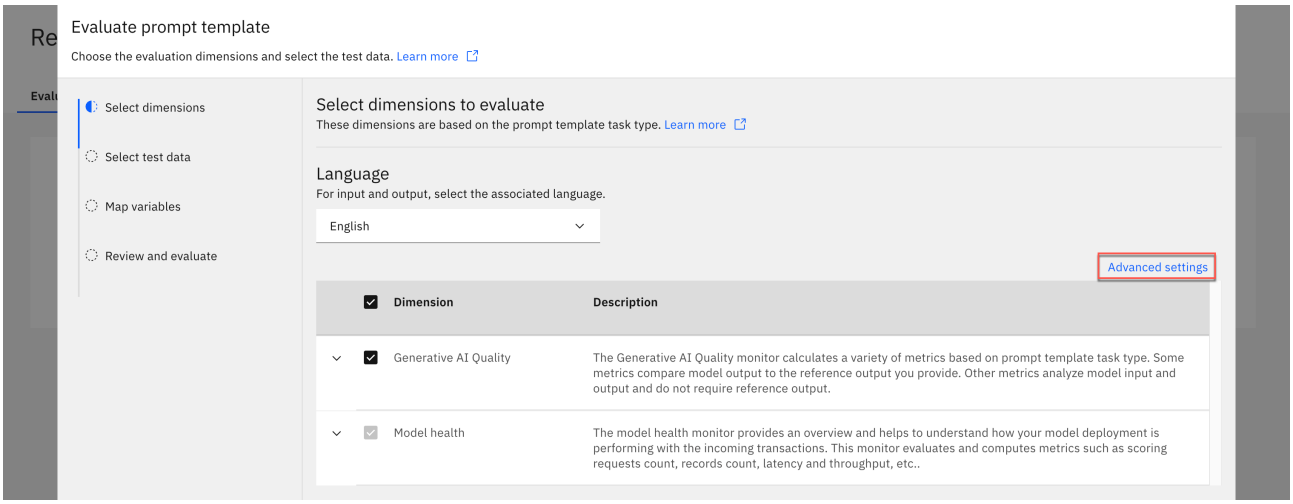


2. From the Evaluations tab, click on the Evaluate button. The Associate a service instance popup appears.



3. Click on the Associate a service instance button to associate a machine learning service with the space. The Evaluate a prompt template window opens. By default, the prompt will be evaluated for generative AI quality and model health. However, you can configure the acceptable thresholds for these metrics.

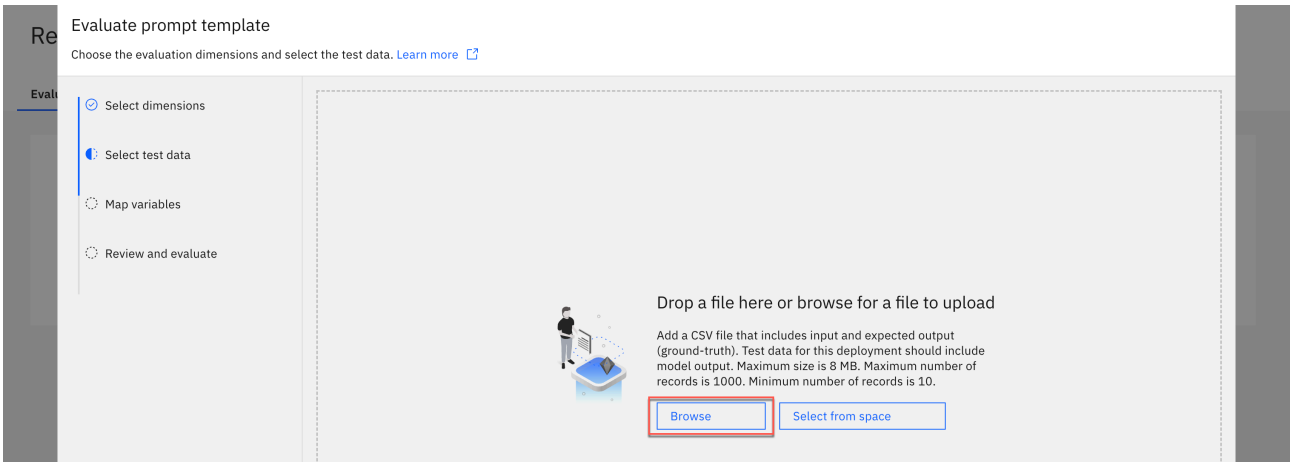
4. Click on the Advanced settings button. The configuration window for the evaluation metrics opens.



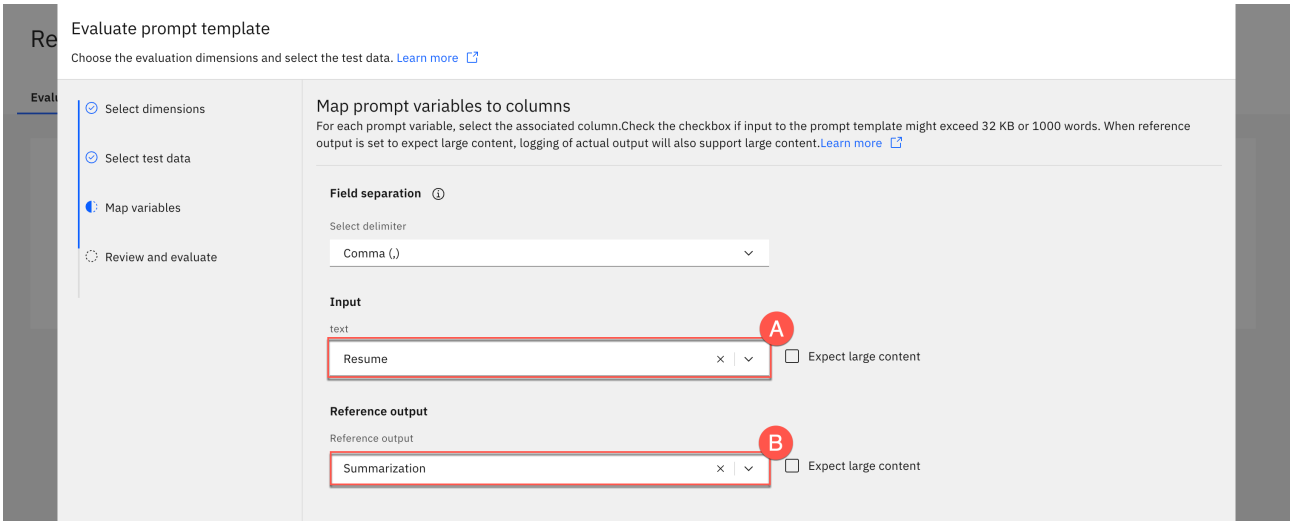
5. Take a moment to review the different thresholds for quality and model health on this screen. When you are finished, click the Save button if you made any changes, or click the Cancel button to return to the Select dimensions to evaluate screen.

6. Click on the Next button to advance to the Select test data screen.

7. Drag and drop the *resume_summarization_eval_test_data.csv* file you downloaded from your project in the previous step into the appropriate area on the screen, or click the Browse button and browse to the file. If you were unable to generate the file, you can download a version of it from GitHub. The Map prompt variables to columns window opens when the file finishes uploading.



8. Click on the text dropdown in the Input section and select Resume (A). Click on the Reference output dropdown and select Summarization (B).

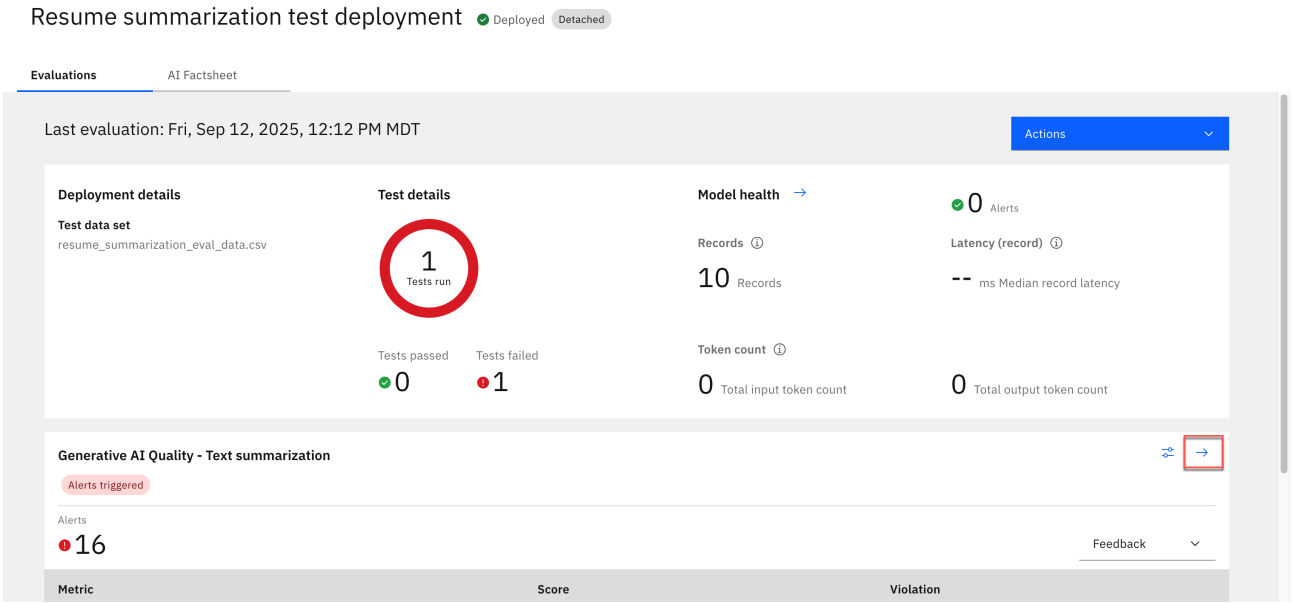


10. Click on the Next button. The Review window opens.

11. Click on the Evaluate button to run the evaluation, which can take several minutes to complete.

> ⚠ The evaluation may fail due to slow network conditions. These failures can frequently be fixed by re-running the evaluation with the same file.

12. Click on the arrow icon in the Generative AI Quality - Text summarization section to open an expanded view of the metrics.



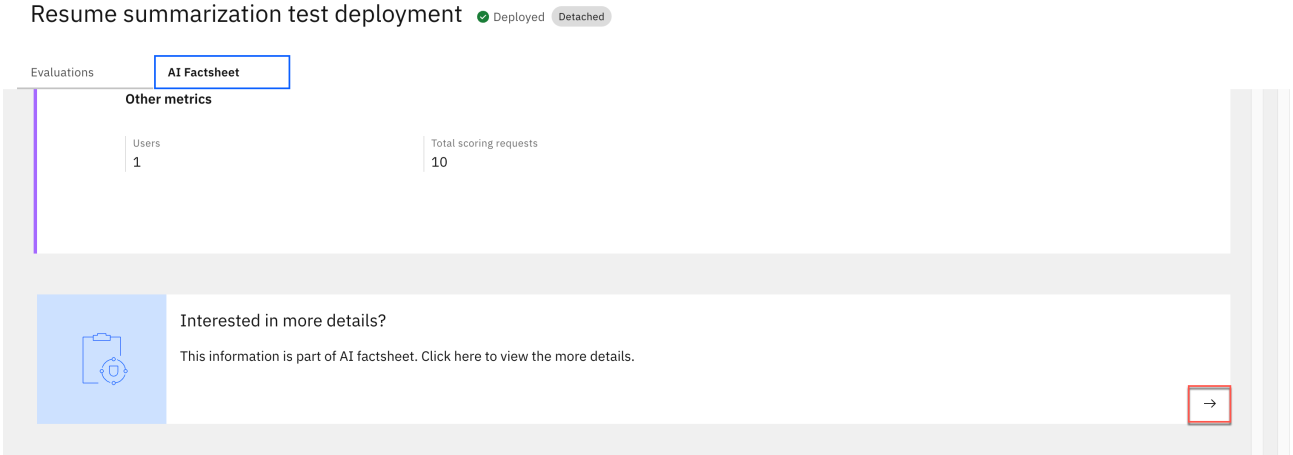
13. Take a moment to review the metrics that have been calculated. For more information on the individual metrics, see the watsonx.governance documentation.

14. Click on the AI Factsheet tab, and note that the model's Factsheet now contains the model's metadata as well as the evaluation results.

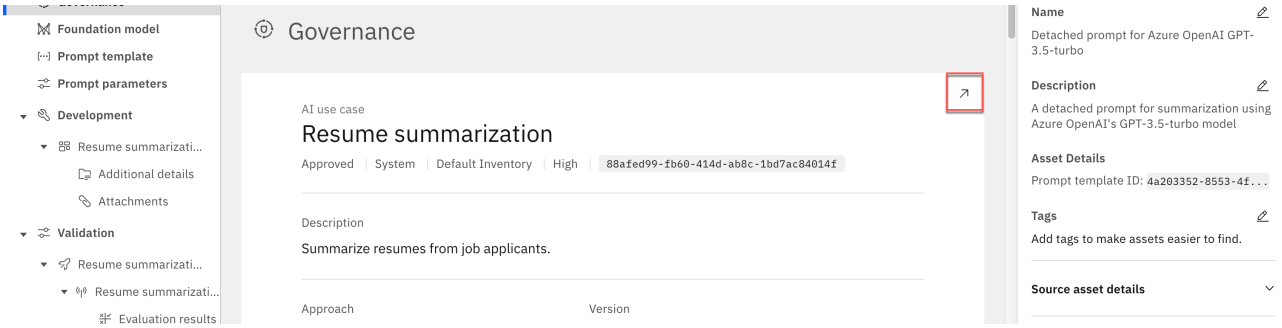
2. View the metrics in the governance console

Now that the metrics have been calculated, they can be viewed in the governance console. The watsonx service automatically updates the model's records in the governance console with the metrics information, allowing stakeholders to be sure that they are viewing the latest data.

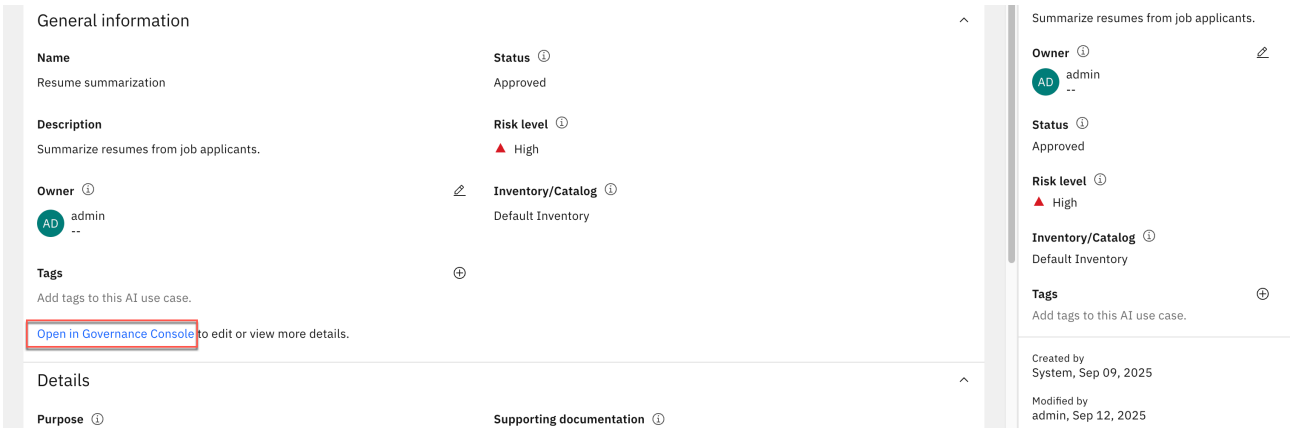
1. Scroll to the bottom of the Factsheet and click on the arrow button in the More details section. A more detailed version of the AI Factsheet opens, showing the model's position in the lifecycle, links to the development project, deployment spaces, and more.



2. From the Governance section, click on the View details button for the use case. The AI use case screen opens.



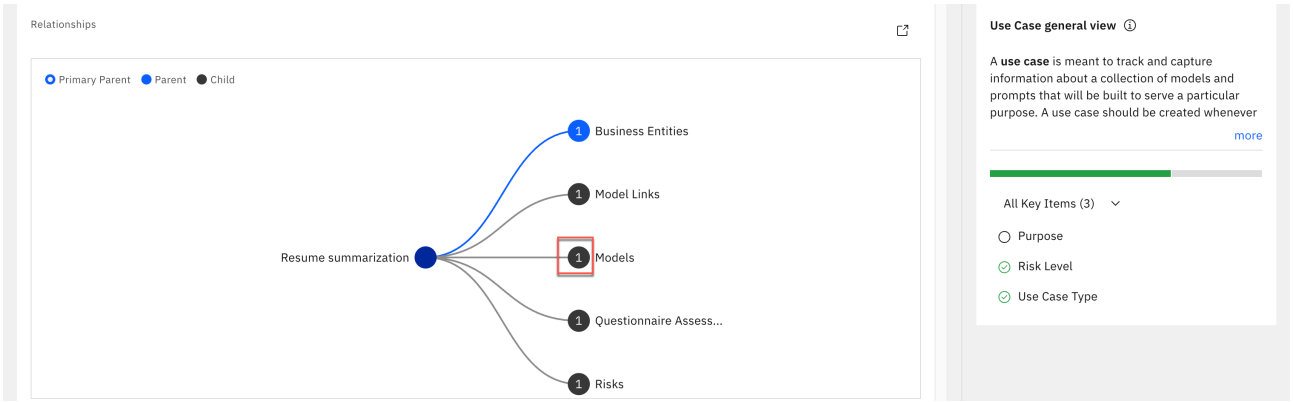
3. Scroll down to the General information section and click on the Open in Governance Console link. The watsonx governance console opens in a new tab and loads the model use case entry.



4. Scroll down to the Performance Monitoring section of the page. This section contains an overview of the metrics generated by the evaluation you ran in the previous step. The Metrics in Breach table shows all the metrics whose values fell below the minimum acceptable thresholds.

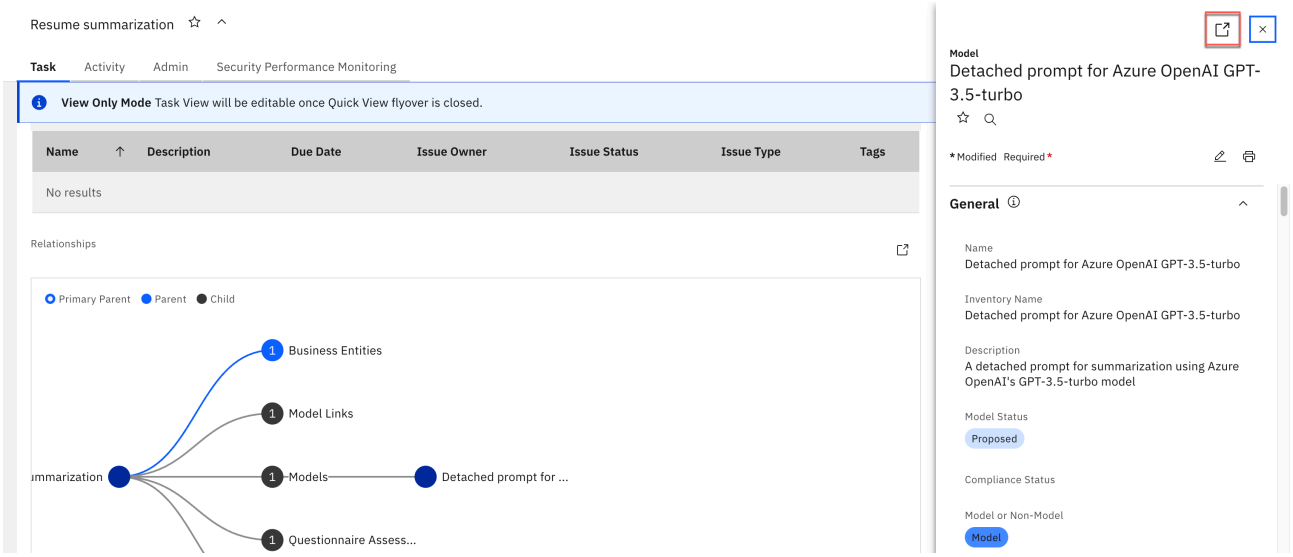
5. Scroll down to the Other Relationships section of the screen. Note that the Resume summarization parent node has one listed Model as child nodes.

6. Click on the circle for the Models node to expand it.



7. The resume summarization model you created and assigned to the use case is listed here. Click on it. The Model information panel opens on the right, showing the model details.

8. Click on the Open in tab button at the top right of the panel. The model will open in a new tab in the governance console.



9. Scroll down to the Associations section of the window and click on the Deployments tab. Note that the tab contains a link to the deployment of the model that you created in a previous step.

At this point in the lab, you have created a questionnaire and customized a governance workflow. Acting as a stakeholder, you have proposed a model for development, and gone through the governance process. You have then deployed and evaluated a third-party model. Throughout the entire process, you have seen how the model metrics and metadata are automatically tracked and surfaced in a variety of locations, allowing risk managers, AI engineers, and other business stakeholders to collaborate on implementing generative AI projects.

Conclusion

Congratulations, you have completed the Governing Generative Models hands-on lab. In this lab, you saw how the configurations you performed in previous labs control the approval process for a model use case. You then oversaw the model lifecycle, including metrics gathering, for a generative models on Microsoft Azure. You saw how the metrics evaluations of those models were automatically updated in multiple platforms, from Factsheets to the governance console, to provide the right information to the right stakeholder at the right time without any additional effort from data science teams, or any reliance on manual processes.

Your feedback is essential to the improvement of this course. Please feel free to provide that on the course page, or directly to the course author. Thank you for your time, and happy selling.

Troubleshooting

The following issues may appear as you run through the lab. This section will grow over time based on user feedback.

1. Governance console Save button disabled

When editing entities in the governance console, occasionally the Save button will be disabled. The most common cause is that some relevant information in the form is missing, which may or may not be called out in the progress panel on the right. Ensure that all required fields (denoted with a red asterisk) have been filled out.

2. Governance console errors

Occasionally, creating new entities or altering existing ones may result in network errors when attempting to save.

In most cases, re-trying the action will resolve the problem. In some cases when creating a new entity, you will receive an error stating that the entity already exists, in which case it likely saved successfully.

Typically, the object has been created successfully, but the action took longer than expected, which generated the failure message. In these cases, you can ignore the message and proceed. In rare cases, you will need to delete and then re-create the entity.

3. Requested operation could not be completed in the governance console

The most frequent cause of this error is incorrectly persisted browser session information when switching between the admin user and the created user in the governance console. For this reason, it is HIGHLY RECOMMENDED that you use your browser's private/incognito mode when signing in as the created user.