

Generative AI Hackathon with IBM Granite

[IBM Granite](#) is a family of AI models purpose-built for business, engineered from the ground up to ensure trust and scalability in AI-driven applications. These enterprise-ready models deliver exceptional performance against safety benchmarks and across a wide range of enterprise tasks from cybersecurity to RAG.

IBM Granite 3.1 is the latest update to our [Granite series](#) of open, performant, enterprise-optimized language models. This suite of improvements, additions and new capabilities focuses primarily on augmenting performance, accuracy and accountability in essential enterprise use cases like tool use, retrieval augmented generation (RAG) and scalable [agentic AI workflows](#).

In this hackathon, you will build an AI-driven solution that harnesses the power of IBM Granite models to help businesses scale faster and operate smarter. Drive efficiency, automation, and innovation into business operations, from streamlining workflows to enabling seamless process integration.

A note on data sets before you begin

Participants are required to bring their own datasets to build the solution aligning to your use case. As you collect data for your project, you'll want to use best practices. Here are helpful tips:

- Teams are responsible for ensuring data is compliant.
- Data from public websites may be used, if the terms allow for commercial use, but please keep a list of the websites you use.
- Do not use data or assets containing company confidential data, or any other data without permission from the data owner. Teams are responsible for getting approval.
- Do not use any client data.
- Do not use any data containing personal information (PI).
- Do not use data obtained from social media.

Get started with IBM Granite models

Participants can access and use IBM Granite models to build their innovative solution by accessing them on:

1. [Open-source platforms](#)
2. [IBM Cloud watsonx.ai](#)

Option 1: Open-source platforms

To access IBM Granite models through open-source platforms and run them locally you have the following options:

- [Download Granite on Hugging Face](#)
- [Run locally with Ollama](#)
 - You have to download and install Ollama on your local machine to use the IBM Granite models.
 - [Mac](#)
 - [Windows](#)
 - [Linux](#)
 - [Granite models supported on Ollama](#)

System requirements

To run Granite models locally, it is recommended to use a machine with **at least 32 GB RAM and a GPU processor**. While running the models on a lower size RAM and CPU is possible, it may result in **slower performance**.

IBM Granite documentation

Refer to [IBM Granite documentation](#) to explore all the [Granite models](#) and [recipes](#) to help you get started.

Chat template

To obtain the best performance from Granite models, we recommend using the official chat template. Refer to the [chat template documentation](#).

Granite Workshop

Try the [Granite Workshop](#) to get hands-on lab exercises for a few use cases that demonstrates the value of generative AI. By the end of this workshop, you will be able to:

- Summarize a text document using [text summarization](#)
- Generate specific information from a large document using the [RAG](#) technique
- Predict future trends using [time series forecasting](#)
- Generate programming code ([Bash](#)) by prompting a code model

BeeAI Framework (recommended framework)

[BeeAI](#) is an open-source framework for building production-ready **AI agents**. The framework is specifically optimized to help you build powerful AI agents with smaller models such as the Granite3 series.

- **Get started with your language of choice**
 - Python → “***pip install beeai-framework***” and try [this example using Granite3.1-8B running in ollama](#).
 - Typescript → “***npm install beeai-framework***” and try [this example using Granite3.1-8B running in ollama](#).
 - Head to github.com/i-am-bee/beeai-framework for more documentation and examples.
- **New to building AI agents?** Try our [hands-on labs](#) to get you ramped up!
 - Get familiar with the [basics](#) such as PromptTemplates, Messages, Memory and how to setup and generate output using a ChatModel.
 - Try out the [Granite-powered ReActAgent](#) and connect it existing tools, or create your own.
 - [Advanced] Solve complex use cases by [building an AI agent as a workflow](#). This low-level implementation gives you the most control and flexibility to define your single agent or multi-agent implementation.

Option 2: IBM Cloud watsonx.ai

To access IBM Granite models and use the SaaS runtime environment for model inference on IBM Cloud watsonx.ai, participants must have requested an IBM Cloud account during hackathon registration. If you did not request an IBM Cloud account during registration, you will only be able to access and use IBM Granite models through open-source platforms.

Note on IBM Cloud service usage

For this hackathon IBM is providing you an additional \$100 credit to use towards IBM watsonx.ai platform. This should be more than enough for you to design and create a very compelling submission for this hackathon. You will receive periodic emails that alert you to how much of your total services credits you have consumed. Email notifications will be sent at 25%, 50%, 80% usage, and your account will be deactivated within an hour once you have used 100%. The periodic email notifications are sent every one hour and there could be chances of you exhausting all your credits within that hour.

Please plan to use the services efficiently and back up your work accordingly. Refer [tips to work efficiently on watsonx.ai platform](#) and [saving your work](#).

Note on available services

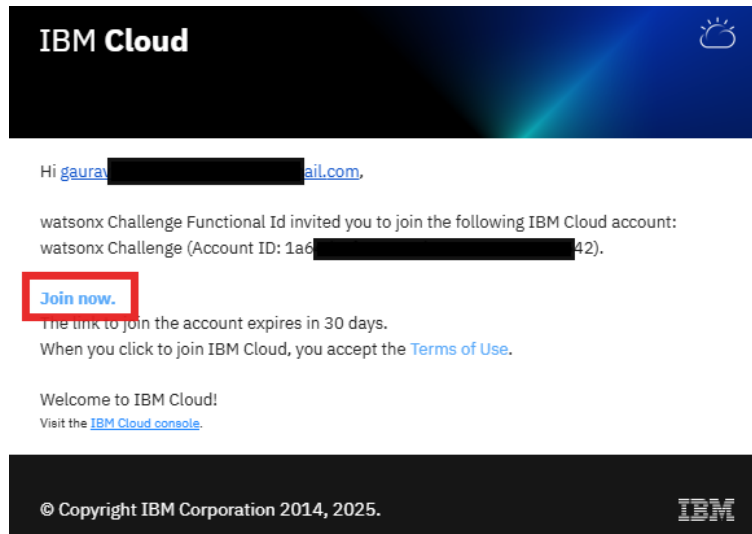
The IBM Cloud and the watsonx.ai platform are pre-configured with all the required services to complete the hackathon. If you notice a permission/access issue for any service or the cloud catalog, then they are not required/available for this hackathon.

Participants will not be able to use the Agent Studio (Beta) and bring their own model or fine tune models. These features are out of scope for this hackathon.

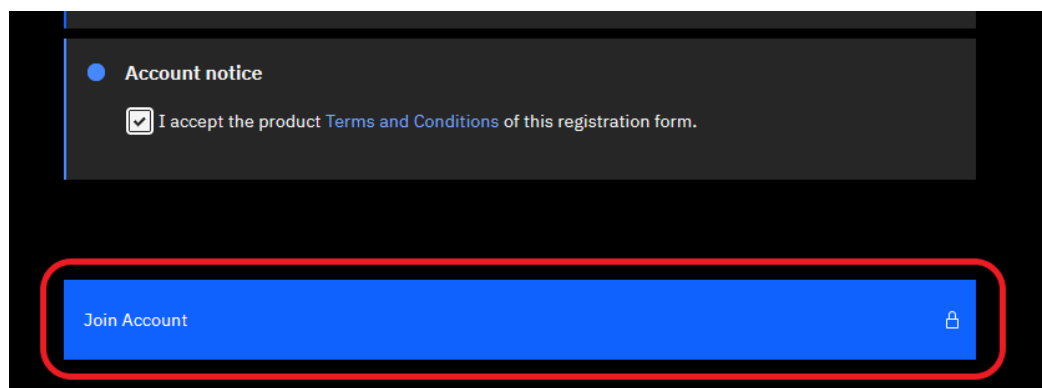
1. Access your IBM Cloud account

- a. Go to your hackathon registered email inbox and open the email you received from the IBM Cloud team about joining your cloud account. You should have received an email invitation from the IBM Cloud team. Please check your junk/spam folders if you are not able to find the email in your inbox. You can also quickly search for “IBM Cloud” to locate the email.

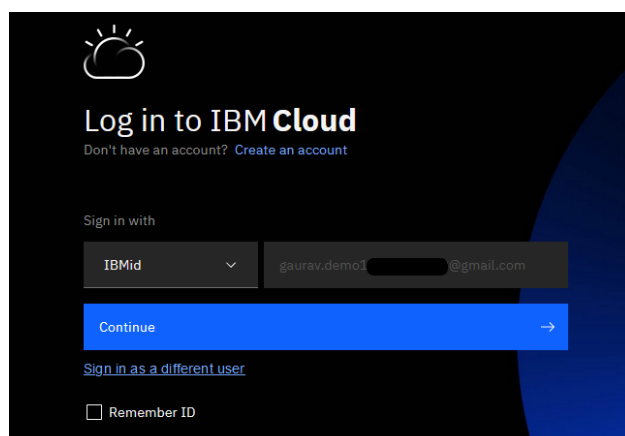
- b. Click the **Join Now** button seen in that email.



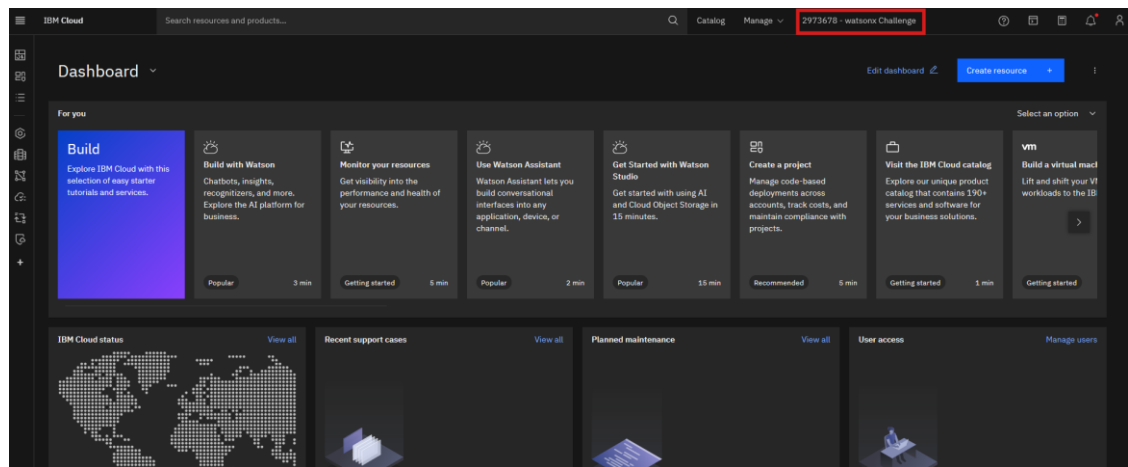
- c. A new browser tab will open with the cloud account sign up page. If you are joining an IBM Cloud account for the first time using your hackathon registered email, you will be asked to enter a new password and your personal information. Read and accept the Account notice and click the **Join Account** button.



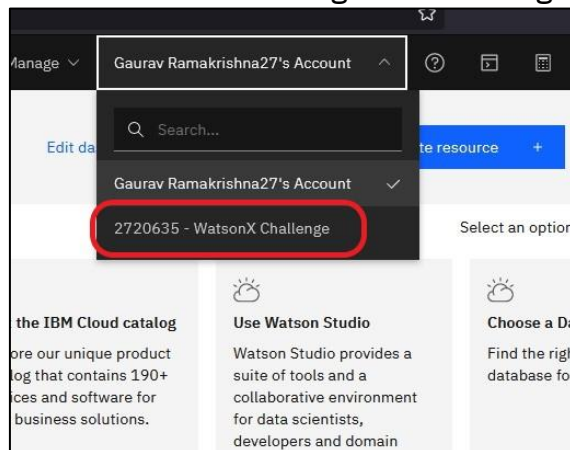
- d. Complete the authentication process by clicking the **Continue** button.



- e. After you authenticate successfully, you will be taken to the IBM Cloud dashboard.



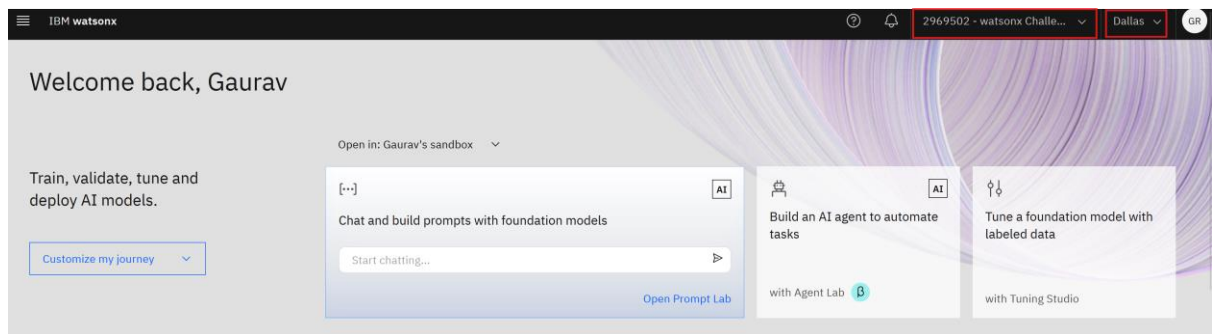
Important note: If you have an existing personal IBM Cloud account, sometimes you will be directed to your personal account. In this case, please switch your account to the **xxxxxxx - watsonx Challenge** account. Refer to the below image on switching accounts.



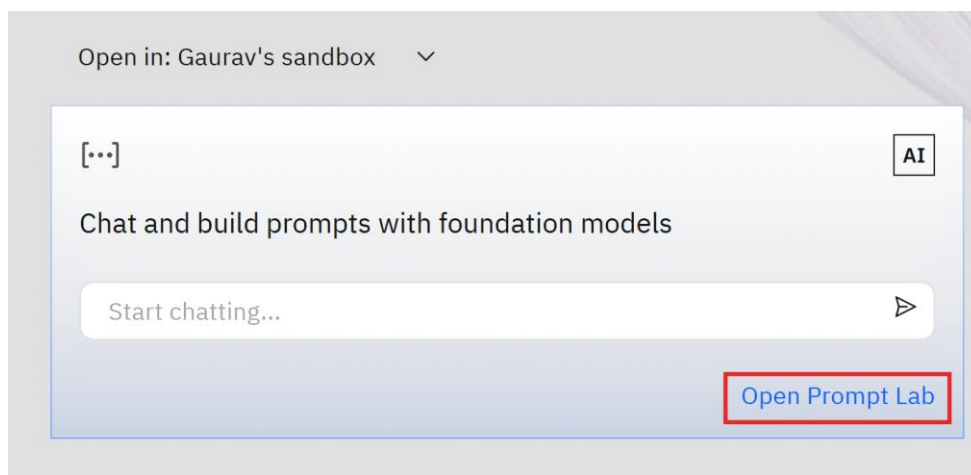
2. Access Prompt Lab on watsonx.ai platform

After successfully joining the IBM Cloud account, you can now access the Prompt Lab on watsonx.ai platform and use the Granite model to build your innovation solution.

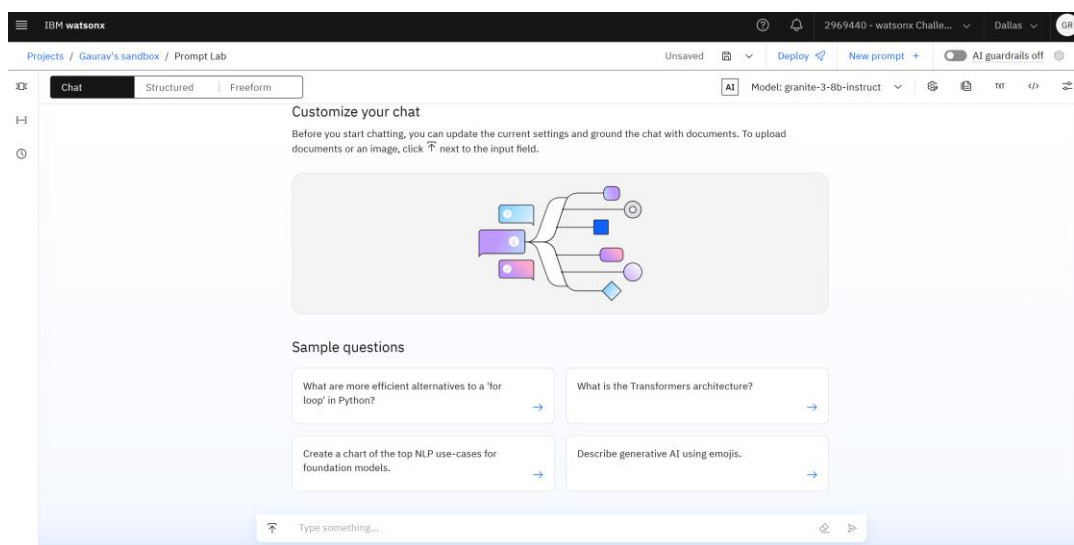
- Log in to the watsonx.ai platform (<https://datapatform.cloud.ibm.com/wx/home?context=wx>) with the email you used to access your IBM Cloud account.
- After successful authentication, you will see the watsonx.ai dashboard. Ensure the name of the account is **“xxxxxxx – watsonx Challenge”** and the region is **“Dallas”**.



- c. Select the **“Open Prompt Lab”** button on the “Chat and build prompts with foundational models” widget.



- d. Welcome to Prompt Lab tour will be displayed. You can take the tour to get a quick introduction or skip it.
- e. The Prompt Lab Editor opens with a chat window to get you started with the prompt session.



3. Work with the Prompt Lab

The watsonx.ai Prompt Lab is an easy-to-use prompt engineering interface where you can experiment prompting different IBM Granite foundation models, explore sample prompts, tune model parameters, integrate applications with an API endpoint, and save and share your best prompts.

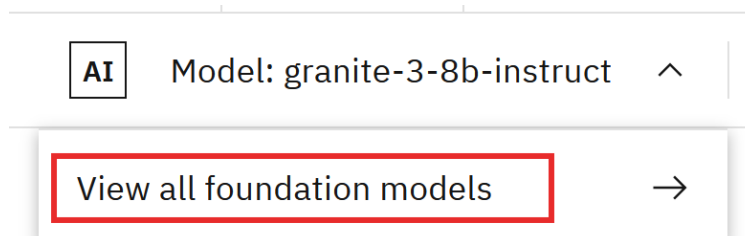
[Take a tour of the Prompt Lab](#) and try the [interactive demo](#).

You can access and use IBM Granite models to build your innovative solution using Prompt Lab.

a. Selecting an IBM Granite model

A **granite-3-8b-instruct** model will be pre-selected by default in the Prompt Lab. You can either use the same model or change to a different Granite model. To select a different Granite model:










- i. Select the AI Model drop-down menu at the top-right of the editor and select **View all foundation models**.



- ii. The **Select a foundational model** widget will appear. Clear all the filters and enter “granite” in the search bar. All the granite series models will be displayed. You can select any granite model tile to learn about the model and use it.

Select a foundation model

To choose a model, review characteristics such as tasks that models perform. Compare model benchmarks with scores in the range 0–100. Higher scores are better.

All models		Model benchmarks	
▼ 🔍 granite		Want to bring your own model?	
 granite-3-8b-instruct The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for... Provider: IBM Type: Provided mo...	 granite-13b-instruct-v2 The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for... Provider: IBM Type: Provided mo...	 granite-20b-code-instruct The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for... Provider: IBM Type: Provided mo...	 granite-20b-multilingual The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for... Provider: IBM Type: Provided mo...
 granite-3-2b-instruct The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for... Provider: IBM Type: Provided mo...	 granite-34b-code-instruct The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for... Provider: IBM Type: Provided mo...	 granite-3b-code-instruct The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for... Provider: IBM Type: Provided mo...	 granite-8b-code-instruct The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for... Provider: IBM Type: Provided mo...
		 granite-guardian-3-2b The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for... Provider: IBM Type: Provided mo...	

b. Programmatic access (API endpoint)

To prompt a Granite model programmatically, you can view and copy the prompt code by selecting the **View code** icon `</>` at the top-right of the editor.



The prompt code is available as a Curl, Node.js and Python.

View code

Create a personal API key, and use it to create temporary access tokens. [Learn more.](#)

```

Curl  Node.js  Python

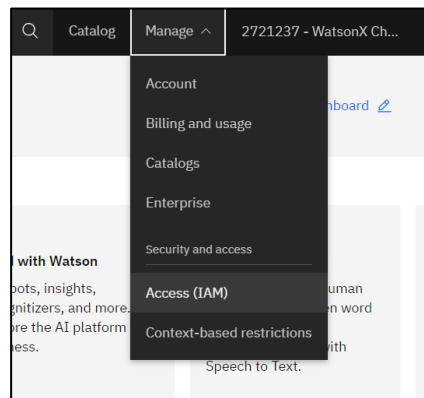
curl "https://us-south.ml.cloud.ibm.com/ml/v1/text/generation?version=2023-05-29" \
-H 'Content-Type: application/json' \
-H 'Accept: application/json' \
-H 'Authorization: Bearer ${YOUR_ACCESS_TOKEN}' \
-d '{
  "input": "<[start_of_role]>system<[end_of_role]>You are Granite, an AI language model developed by IBM in 2024. You are a cautious assistant. You carefully follow instructions. You are helpful and harmless and you follow ethical guidelines and promote positive behavior.<[end_of_text]>\n<[start_of_role]>assistant<[end_of_role]>".
  "parameters": {
    "decoding_method": "greedy",
    "max_new_tokens": 900,
    "min_new_tokens": 0,
    "stop_sequences": [],
    "repetition_penalty": 1
  },
  "model_id": "ibm/granite-3-8b-instruct",
  "project_id": "30w"
}'

```

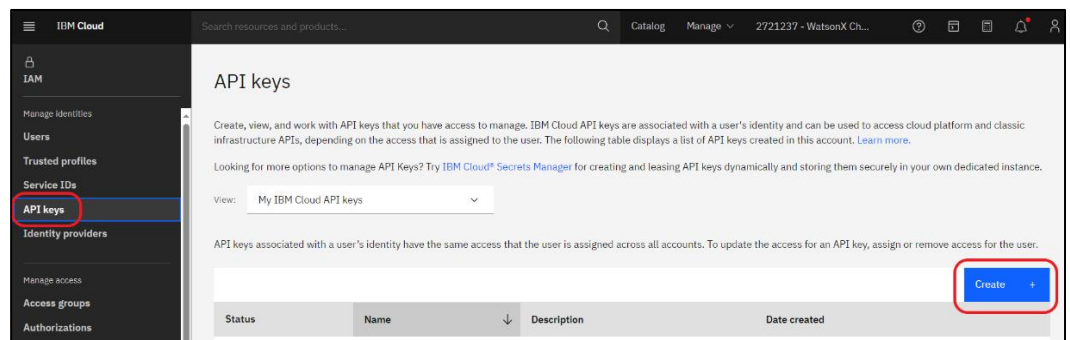
You will require an IAM access token to authorize the prompt code and need to replace **`${YOUR_ACCESS_TOKEN}`** placeholder in the prompt code. You can create an IAM access token using the IBM Cloud API key.

i. Create an IBM Cloud API key

- a. In your [IBM Cloud account dashboard](#), select **Manage > Access (IAM)** at the top of the dashboard.



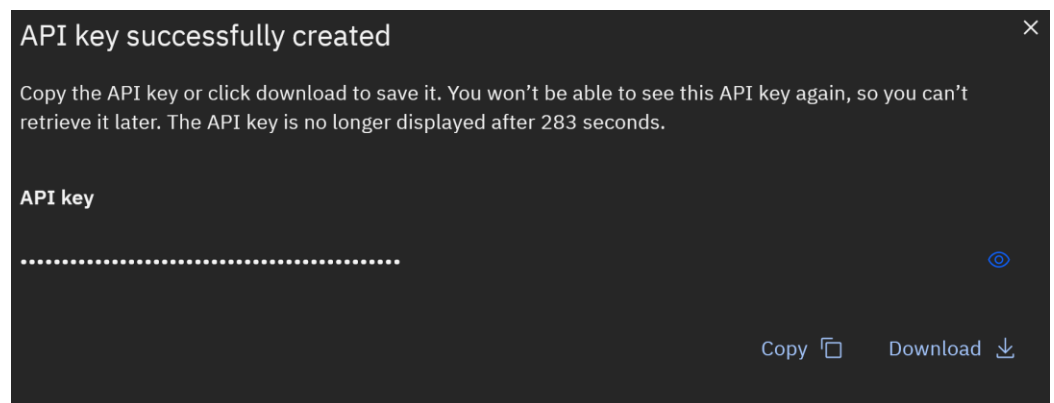
- b. Select **API keys** on the left pane and click the **Create +** button.



- c. Enter a name, select **Yes** to “Session creation” and click the **Create** button.

A screenshot of the 'Create IBM Cloud API key' dialog box. The 'Name' field is filled with 'watsonx-ai-key'. Below it is a 'Description (optional)' field. The 'Leaked action' section has three radio button options: 'Disable the leaked key' (selected), 'Delete the leaked key', and 'Nothing'. The 'Session creation' section has two radio button options: 'Yes' (selected) and 'No'. At the bottom, there are 'Cancel' and 'Create' buttons. The 'Create' button is highlighted.

- d. An API key will be created. Download and save the file in a secure path in your system.



ii. Generate IAM Access Token

Programmatically generate an IAM access token with the IBM Cloud API key using the following cURL command:

```
curl -X POST 'https://iam.cloud.ibm.com/identity/token' -H  
'Content-Type: application/x-www-form-urlencoded' -d  
'grant_type=urn:ibm:params:oauth:grant-type:apikey&apikey=MY_APIKEY'
```

- **curl -X POST** → Specifies an HTTP **POST** request.
- **URL ("https://iam.cloud.ibm.com/identity/token")** → The endpoint to request an authentication token from IBM Cloud.
- **-H "Content-Type: application/x-www-form-urlencoded"** → Sets the request header to indicate that the data is sent in **form-encoded format**.
- **-d (Data Payload)** → Sends the required data:
 - **grant_type=urn:ibm:params:oauth:grant-type:apikey**
→ Specifies the OAuth grant type as API Key.

- **apikey=MY_IBM_CLOUD_API_KEY** → Replace MY_IBM_CLOUD_API_KEY with your actual IBM Cloud API key.

Expected Response:

```
{
  "access_token": "eyJhbGciOiJIUz.....sgrKIi8hdFs",
  "refresh_token": "SPrXw5tBE3.....KBQ+luWQVY=",
  "token_type": "Bearer",
  "expires_in": 3600,
  "expiration": 1473188353
}
```

The Prompt Lab graphical interface is a great place to experiment and iterate with your prompts. However, you can also prompt foundation models in watsonx.ai programmatically by using the Python library or REST API. For details, see [Coding generative AI solutions](#).

c. Creating and running a prompt

You can utilize the watsonx.ai Prompt Lab editor to build your AI solution. Explore the three modes of the Prompt Lab editor:

- [Chat](#)
- [Structured](#)
- [Freeform](#)

d. Quick start labs

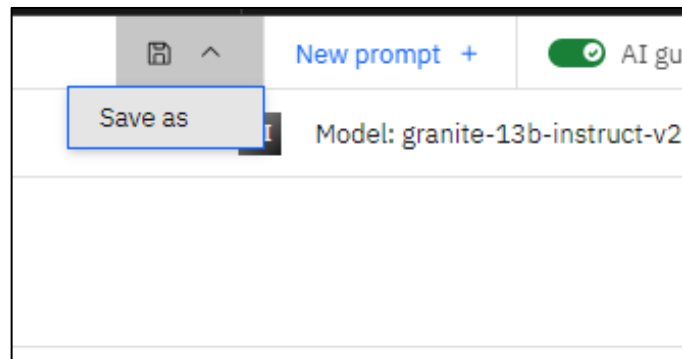
Explore the quick start labs:

- [Prompting a foundational model with Prompt Lab](#)
- [Prompting a foundational model with the RAG pattern](#)

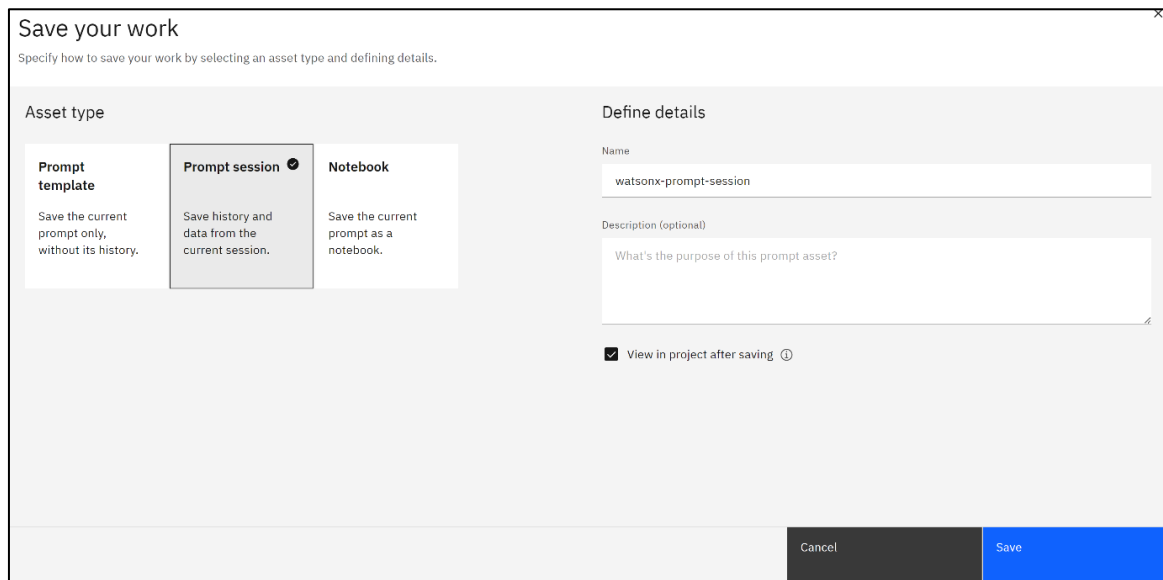
e. Save your prompt

You can save your Prompt Lab session for later use.

1. On the top of the Prompt Lab, select the **Save work** dropdown button and then select the **Save as** option.



2. A **Save your work** widget will appear. Select **Prompt session** under the **Asset type** option. Enter a **name** and check the **View in project after saving** option under the **Define details** section. Finally, click the **Save** button.



Save your work

Specify how to save your work by selecting an asset type and defining details.

Asset type

- Prompt template**
Save the current prompt only, without its history.
- Prompt session** (selected)
Save history and data from the current session.
- Notebook**
Save the current prompt as a notebook.

Define details

Name
watsonx-prompt-session

Description (optional)
What's the purpose of this prompt asset?

☒ View in project after saving ⓘ

Cancel Save

3. Once you save, you will see the saved work under the **Assets** tab.
4. You can also save your work as:
 - a. **Prompt template** to save only the current prompt without its history and selecting a **Task** suitable for your prompting.
 - b. **Notebook** to continue working on your prompting on a Jupyter Notebook environment. Prior knowledge of notebooks and Python programming language would be helpful to work with Jupyter notebook. [Read more about notebooks.](#)

f. **Save your work**

Make sure to save any work you want to retain for your records. IBM Cloud accounts will be disabled at the end of the hackathon. Follow the below steps to save your work:

1. To save the Prompt Lab work, click on the 'Save' icon in the top menu bar (under the Bell icon)
2. Click 'Save As'
3. Select 'Notebook' as the Asset type on the next screen
4. Enter a name and description for the Notebook, and then click 'Save'
5. Go to your project's 'Overview' tab
6. Click on the 'Export or import project' drop down below the Bell icon in the top menu bar
7. Click on 'Export project' > this will open 'Export project to desktop' screen
8. Select all Notebook assets shown in your project (Work saved as Project template or Project session cannot be exported) and click 'Export' on the bottom right of the screen
9. Then next screen will ask for confirmation that all sensitive information has been removed
10. Click on 'Continue export'
11. The download (.zip) will be initiated and file will be saved on your computer