IBM **TechXchange**

Hackathon for Progress with RAG and IBM watsonx.ai

Hackathon for Progress with RAG and IBM watsonx.ai

IBM watsonx.ai is a powerful AI platform, with robust data processing and intelligent system creation, designed to streamline model inference and facilitate the development of advanced Retrieval-Augmented Generation (RAG) solutions. With watsonx.ai, you can build intelligent systems that deliver personalized experiences, optimize processes, and drive innovation across various domains.

Retrieval augmented generation (RAG) is an architecture for optimizing the performance of an artificial intelligence (AI) model by connecting it with external knowledge bases. RAG helps large language models (LLMs) deliver more relevant responses at a higher quality. Learn more about RAG.

RAG adds power to generative AI models by interspersing real-time data retrieval, ensuring that the retrieval process produces a more accurate and timely output. However, RAG models come in different forms, suited primarily for different applications. Explore various RAG techniques along with how they work, strengths, and limitations of each RAG type and their usability in various use cases.

Contents

The hackathon challenge

In this hackathon, you will use watsonx.ai and retrieval augmented generation (RAG) to **build a proof-of-concept, AI solution** for **one** of the following hackathon themes:

- **Advance the future of customer experience**
  Ready to make customer interactions smarter, faster, and more personal? Develop a proof-of-concept using RAG and watsonx.ai to tackle real-world challenges like building problem-solving virtual assistants that leave customers delighted, creating personalized marketing strategies, or analyzing customer feedback to uncover hidden gems.
  Refer to Advance the future of customer experience use cases.

- **Climate Challenge, brought to you by Call for Code**
  Take urgent action to combat climate change and its impacts. We're calling on you to build a proof-of-concept, RAG AI-driven solution using watsonx.ai that can address a pressing climate challenge.
  Refer to Climate Challenge use cases.

Participants are expected to create a **custom retrieval augmented generation (RAG) AI solution** using **IBM watsonx.ai** as a model inference provider.

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 the best practices. Here are some 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 watsonx.ai

To access IBM watsonx.ai and use it for the hackathon, participants must be registered for the hackathon and have access to the hackathon site. Once you have access to the hackathon site, follow the instructions on the "**Complete the hackathon**" page to request an IBM Cloud account for your team to use the watsonx.ai platform to build your RAG AI solution.

Note on IBM Cloud service usage

For this hackathon, **$100 credits** will be automatically applied on the provisioned **IBM watsonx.ai platform**. This should be sufficient for designing and creating a compelling submission.

You will receive periodic email notifications about your **credit consumption** at the following usage levels: **25%, 50%,** and **80%**. Once you reach **100% usage**, your account will be **suspended**. You can appeal the suspension by completing the form shared in the account suspension notification email.

Please note that these email notifications are sent **once per hour**, so there is a possibility that you may **exhaust all your credits before receiving an alert**.

Please plan to use the watsonx.ai efficiently and back up your work accordingly. Refer **tips to work efficiently on watsonx.ai platform** **(Tokens and CUH explained)** and **saving your work**.

Important:

- **Foundation model inferencing** consumes tokens, which are measured as Resource Units (RUs). **1,000 tokens = 1 RU,** and each RU costs **$0.0001 USD.**
  Learn more about tokens and tokenization.

- If you are using **Jupyter Notebook editor on watsonx.ai,** consider selecting **a lower runtime environment** to avoid high resource consumption and quickly depleting your credits. Notebook runtimes are billed based on **Capacity Unit Hours (CUH)** at a rate of **$1.02 USD per CUH.**
  Learn more about capacity unit hours and watsonx.ai Studio pricing plans.

Note on available services

The IBM Cloud and the watsonx.ai platform are **pre-configured with only the services required** 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.

These features/capabilities are out of scope for this hackathon:

- Agent Studio (Beta)

- Deploy on IBM Cloud/watsonx.ai (including Deployment space)
- Bring your own model
- Fine tuning models
- AutoAI pipeline
- SPSS Modeler
- Federated Learning
- Cloud Object Storage service

DO NOT USE the below listed models as they are out of scope for the hackathon and can negatively impact the judgment of your project submission.

- llama-3-405b-instruct
- mistral-large
- mixtral-8x7b-instruct-v01
- mistral-small-3-1-24b-instruct-2503
- pixtral-12b

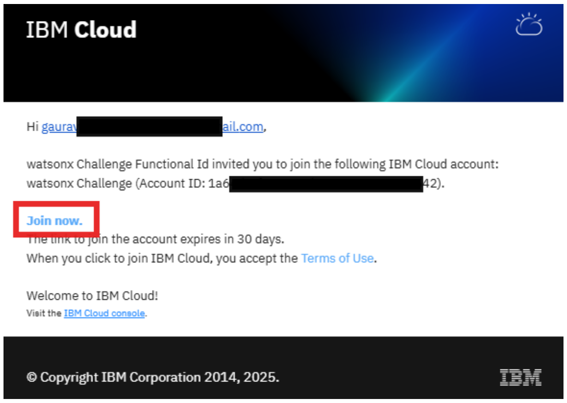
The hackathon provisioned IBM Cloud account will be deactivated after the completion of the hackathon. Please plan to save your work at the end of the hackathon.

Access your IBM Cloud account

Once your team has been provisioned an IBM Cloud account, all team members will receive an email invite to join the cloud account. Follow the steps below to access your team's cloud account:

1. Check the email inbox you used to register for the hackathon and open the email you received from the IBM Cloud team about joining your cloud account. 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.

2. Click the **Join Now** button seen in that email. A new browser tab will open with the cloud account sign up page.



IBM Cloud

Hi gaura[redacted]ail.com,

watsonx Challenge Functional Id invited you to join the following IBM Cloud account:
watsonx Challenge (Account ID: 1a6[redacted]42).

Join now.

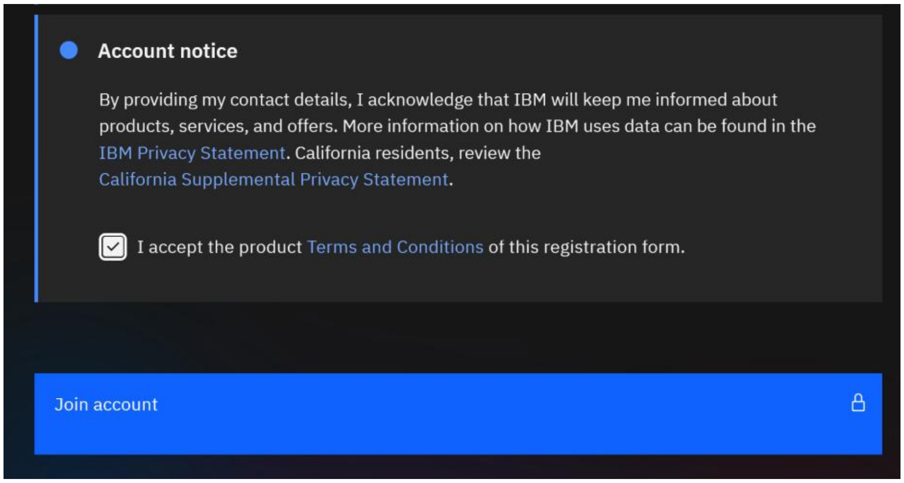
The link to join the account expires in 30 days.
When you click to join IBM Cloud, you accept the Terms of Use.

Welcome to IBM Cloud!
Visit the IBM Cloud console.

© Copyright IBM Corporation 2014, 2025.

3. Review your account and personal information.

4. Read and accept the Account notice and click the **Join Account** button.

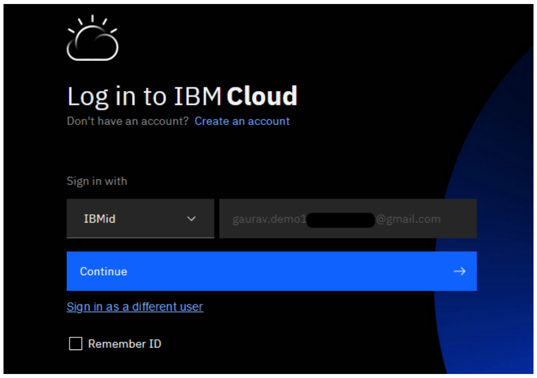


● **Account notice**

By providing my contact details, I acknowledge that IBM will keep me informed about products, services, and offers. More information on how IBM uses data can be found in the IBM Privacy Statement. California residents, review the California Supplemental Privacy Statement.

☑ I accept the product Terms and Conditions of this registration form.

5. Complete the authentication process by clicking the Continue button.



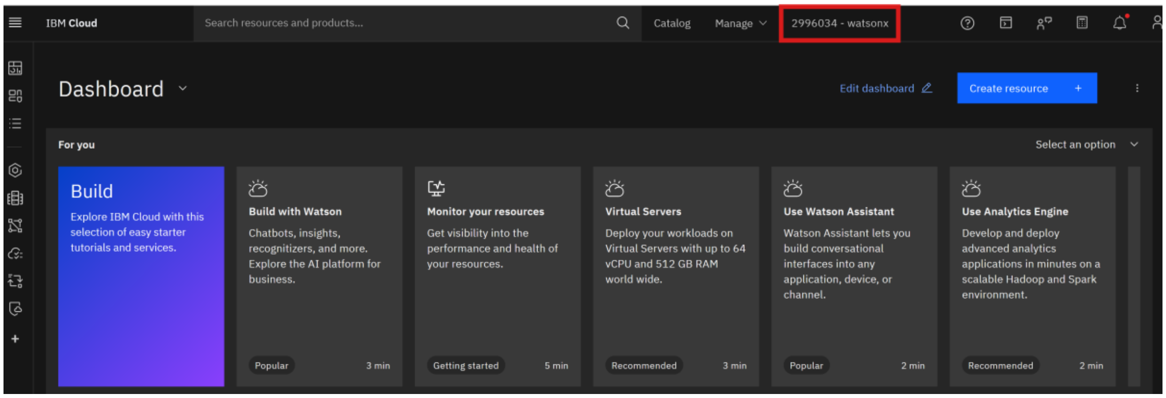
Log in to IBM **Cloud**

Don't have an account? Create an account

Sign in with

IBMid: gaurav.demo1[redacted]@gmail.com

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



Explore IBM Cloud with this selection of easy starter tutorials and services.

Chatbots, insights, recognizers, and more. Explore the AI platform for business.

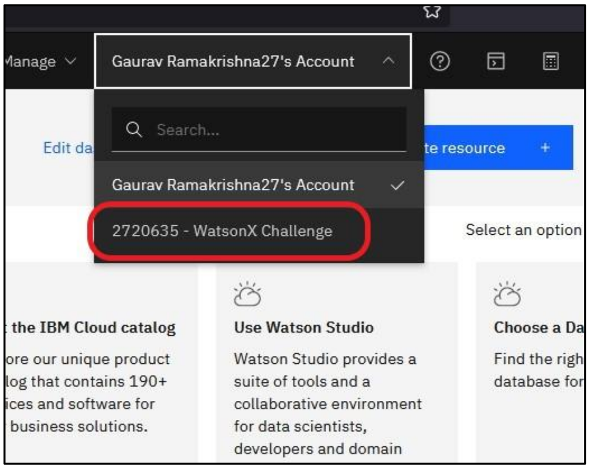
Get visibility into the performance and health of your resources.

Deploy your workloads on Virtual Servers with up to 64 vCPU and 512 GB RAM world wide.

Watson Assistant lets you build conversational interfaces into any application, device, or channel.

Develop and deploy advanced analytics applications in minutes on a scalable Hadoop and Spark environment.

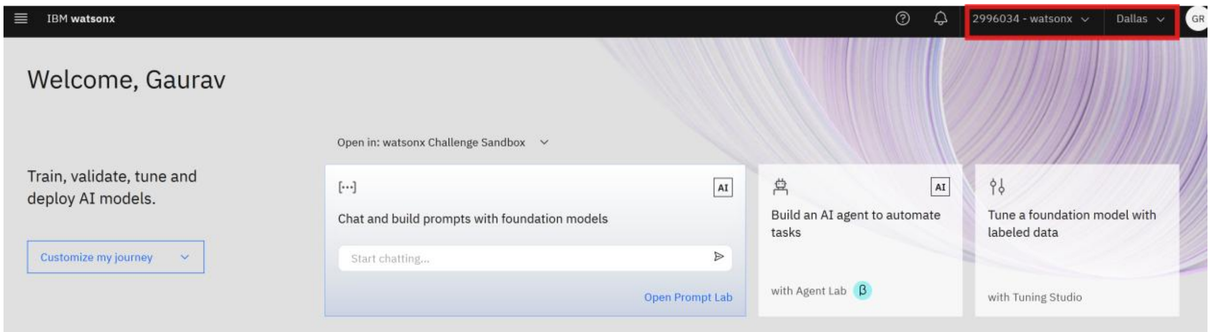
7. If you have an existing personal IBM Cloud account for the same email/IBMid, sometimes you will be directed to your personal account. In this case, please switch your account to the **xxxxxxx - watsonx** account. Select your account drop-down at top-right of the dashboard and select watsonx account. Refer to the below image on switching accounts in your cloud dashboard.



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 to work with the AI models supported on the platform and build your solution.

1. Log in to the watsonx.ai platform (https://dataplatform.cloud.ibm.com/wx/home?context=wx) with the email you used to access your IBM Cloud account.
2. After successful authentication, you will see "Welcome to watsonx". You can either take the tour or skip it.
3. Next, you will see the watsonx.ai dashboard. Ensure the name of the account is "**xxxxxxx – watsonx**" and the region is "**Dallas**".



Welcome, Gaurav

Train, validate, tune and deploy AI models.

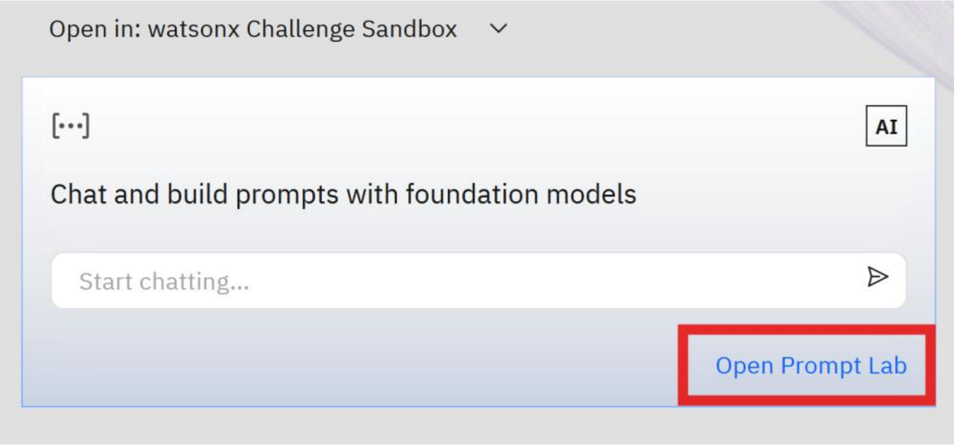
Open in: watsonx Challenge Sandbox

Chat and build prompts with foundation models

Build an AI agent to automate tasks

Tune a foundation model with labeled data

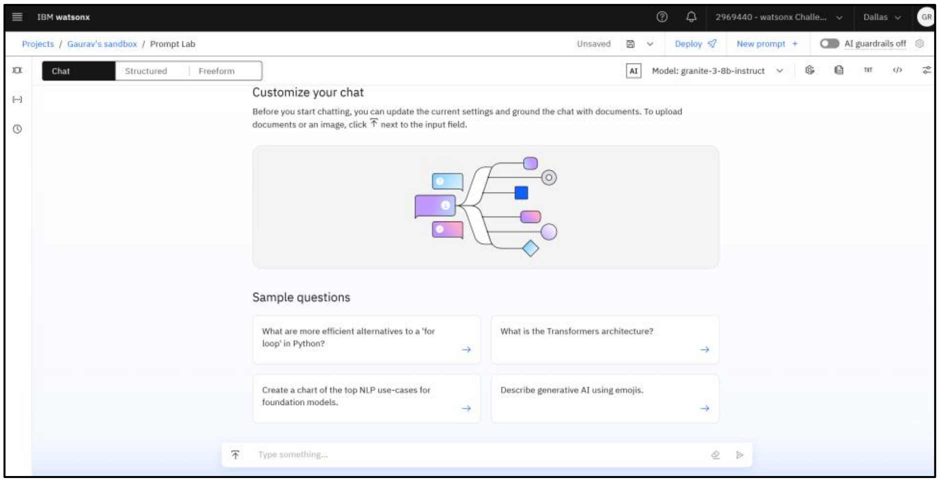
4. Select the “**Open Prompt Lab**” button on the “Chat and build prompts with foundational models” widget.



Open in: watsonx Challenge Sandbox

Chat and build prompts with foundation models

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



Customize your chat

Before you start chatting, you can update the current settings and ground the chat with documents. To upload documents or an image, click ⤒ next to the input field.

Sample questions

What are more efficient alternatives to a 'for loop' in Python? →

What is the Transformers architecture? →

Create a chart of the top NLP use-cases for foundation models. →

Describe generative AI using emojis. →

Work with the watsonx.ai Prompt Lab

The watsonx.ai Prompt Lab is an easy-to-use prompt engineering interface where you can experiment prompting different AI 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 the AI models to build your innovative solution using Prompt Lab.

Prompt Lab editor

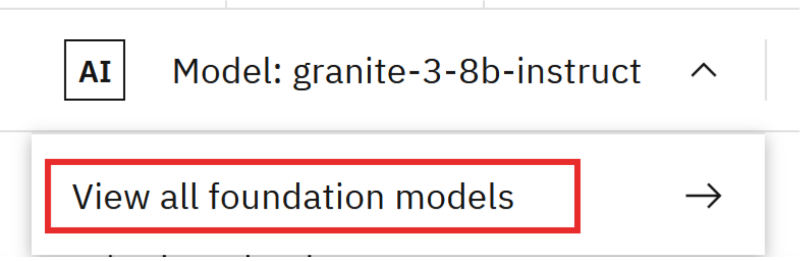
In the Prompt Lab, you can experiment with prompting different foundation models, explore sample prompts, as well as save and share your best prompts. The Prompt Lab editor is a great place to experiment and iterate with your prompts. Try the quick start lab.

However, you can also prompt foundation models in watsonx.ai programmatically. Refer to "Programmatic access (API/SDK)" section.

Selecting an AI model

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

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

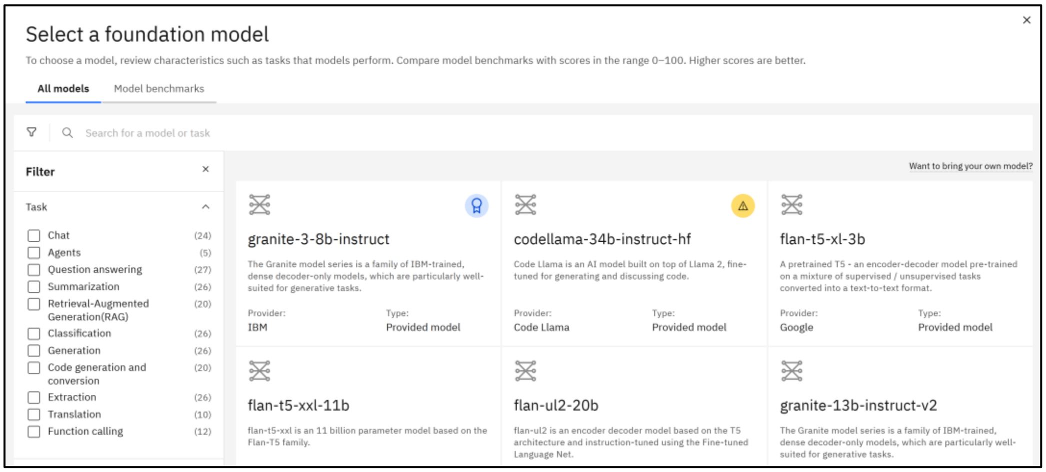


2. The **Select a foundation model** widget will appear. Clear the filters to see all the available models. You can use the filters to choose the right model for your solution building. You can select a model tile to learn about the model and use it.

Important: **DO NOT USE the below listed models as they are out of scope for the hackathon and can negatively impact the judgment of your project submission.**

- llama-3-405b-instruct

- mistral-large
- mixtral-8x7b-instruct-v01
- mistral-small-3-1-24b-instruct-2503
- pixtral-12b



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

Search for a model or task

Filter

Task

- □ Chat (24)
- □ Agents (5)
- □ Question answering (27)
- □ Summarization (26)
- □ Retrieval-Augmented Generation(RAG) (20)
- □ Classification (26)
- □ Generation (26)
- □ Code generation and conversion (20)
- □ Extraction (26)
- □ Translation (10)
- □ Function calling (12)

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 generative tasks.

Provider: IBM Type: Provided model

codellama-34b-instruct-hf

Code Llama is an AI model built on top of Llama 2, fine-tuned for generating and discussing code.

Provider: Code Llama Type: Provided model

flan-t5-xl-3b

A pretrained T5 - an encoder-decoder model pre-trained on a mixture of supervised / unsupervised tasks converted into a text-to-text format.

Provider: Google Type: Provided model

flan-t5-xxl-11b

flan-t5-xxl is an 11 billion parameter model based on the Flan-T5 family.

flan-ul2-20b

flan-ul2 is an encoder decoder model based on the T5 architecture and instruction-tuned using the Fine-tuned Language Net.

granite-13b-instruct-v2

The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for generative tasks.

To understand how models can address your use case, including information on model modalities, supported languages, tuning, and indemnification, see our product documentation on choosing a model.

Note: Bigger models are not always better. Learn why smaller models can be better and more cost effective.

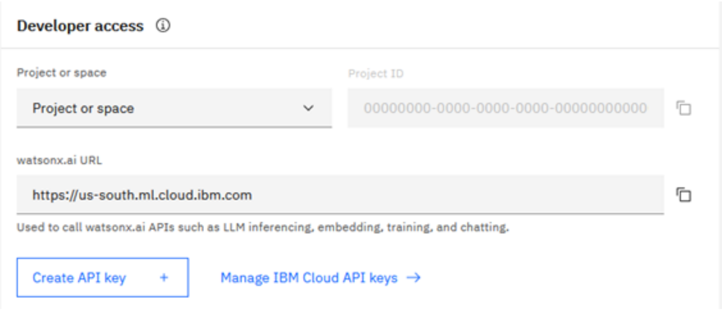
Programmatic access (API/SDK)

You can inference the watsonx.ai models with API or SDK requests.

Developer access information

To use the supported watsonx.ai APIs/SDKs, you will need three values: a **project ID**, an **endpoint URL** and an **API key**.

- Go to watsonx.ai home page.
- Scroll down to the “**Developer access**” section.



Developer access

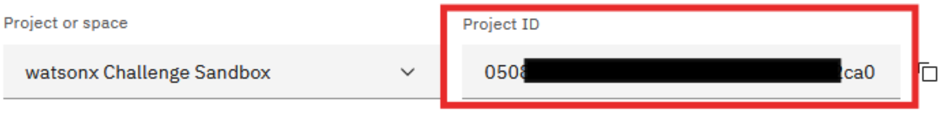
Project or space: Project or space

Project ID: 00000000-0000-0000-0000-00000000000

watsonx.ai URL: https://us-south.ml.cloud.ibm.com

Used to call watsonx.ai APIs such as LLM inferencing, embedding, training, and chatting.

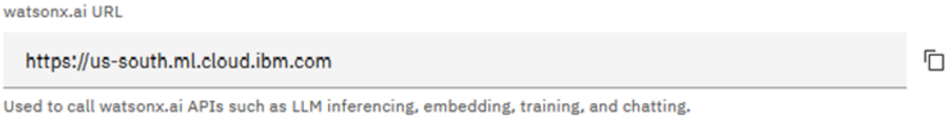
- Select the “**Project or space**” drop-down and select the “**watsonx Challenge Sandbox**” option. A **project ID** will be displayed.
  Note: A space ID is **not required** as it is out of scope for the hackathon.



Project or space: watsonx Challenge Sandbox

Project ID: 050[redacted]ca0

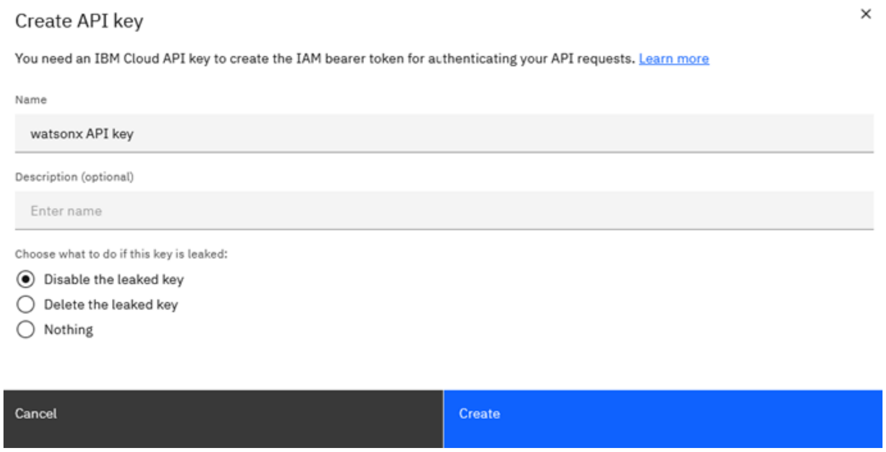
- A default **watsonx.ai endpoint URL** will be displayed for the Dallas region. Ensure the region is always set to **Dallas** at the top right of the watsonx.ai home page.



watsonx.ai URL: https://us-south.ml.cloud.ibm.com

Used to call watsonx.ai APIs such as LLM inferencing, embedding, training, and chatting.

- Select the “**Create API key**” button. A **Create API key** widget will be displayed. Enter a name, provide optional description and choose the “Disable the leaked key” option. Click the “**Create**” button.



Create API key

You need an IBM Cloud API key to create the IAM bearer token for authenticating your API requests. Learn more

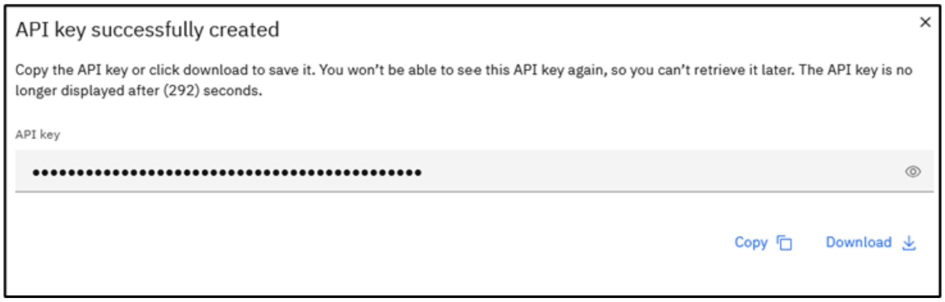
Name: watsonx API key

Description (optional): Enter name

Choose what to do if this key is leaked:

● Disable the leaked key
○ Delete the leaked key
○ Nothing

- An API key will be created successfully. Copy the API key and save it safely to use for calling the API/SDK. You can also download and save the file in a secure path in your system.



API key successfully created

Copy the API key or click download to save it. You won't be able to see this API key again, so you can't retrieve it later. The API key is no longer displayed after (292) seconds.

API key

••

watsonx.ai programmatic options

There are multiple options to help you get started using watsonx.ai APIs/SDKs.

Option 1: Prompt Code on Prompt Lab

Refer to the access prompt code instructions to learn how to quickly get access to the text generation API within the watsonx.ai Prompt Lab.

Option 2: Different watsonx.ai API capabilities

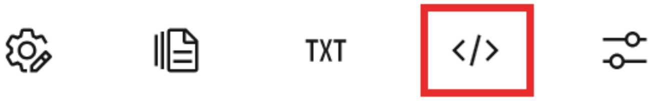
Explore and leverage different watsonx.ai API capabilities in your solution.

- Chat
- Tool calling
- Text generation
- Time series
- Text rerank
- Embeddings
- Text extraction

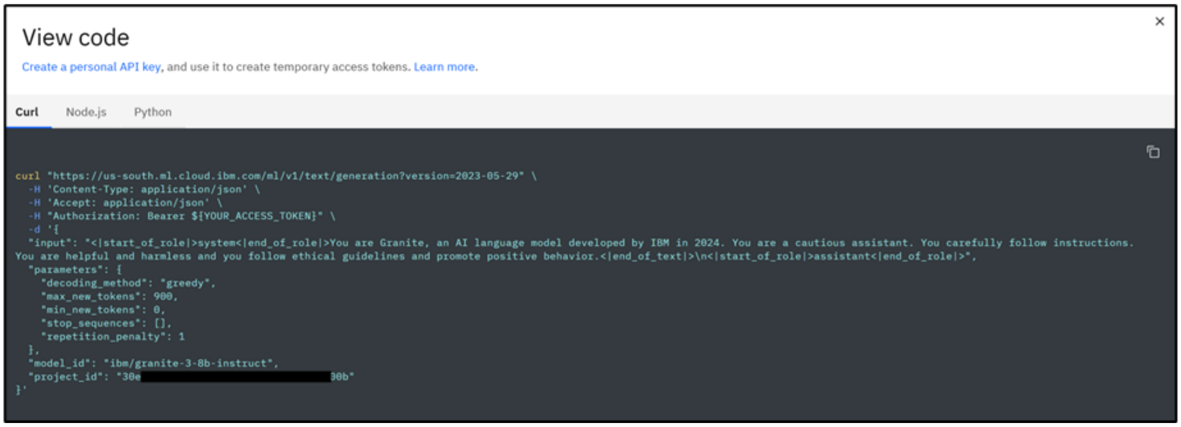
Refer supported API functionality by model here.

Access the prompt code (API) from Prompt Lab editor

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



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



Create a personal API key, and use it to create temporary access tokens. Learn more.

```
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": "30e[redacted]00b"
}'
```

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 an API key.

- **API key**

  Refer to Developer access information to get an API key.

- **Generate IAM Access Token**

  Programmatically generate an IAM access token with the 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": "not_supported",

"token_type": "Bearer",

"expires_in": 3600,

"expiration": 1473188353,
```

Note: An IAM token is valid for up to 60 minutes, and it is subject to change. When a token expires, you must generate a new one. Use the property "*expires_in*" for the expiration of the IAM token that you have just created.

Quick start hands-on exercises

Try the quick start exercises and notebooks for sample use cases to get started with using watsonx.ai.

Important notes:

- Refer to developer access information section to use watsonx.ai credentials as you try the exercises.
- Some of the exercises could include the usage of old AI model version. You can replace them with newer versions for better performance and output. To check the latest supported AI models on watsonx.ai, either follow select an AI model on Prompt Lab or refer to supported foundation models on watsonx.ai.
- The hackathon provisioned cloud accounts **do not support solution deployment.** You can run your solution deployment locally on your machine and showcase them in your submissions.
- **Foundation model inferencing** consumes tokens, which are measured as Resource Units (RUs). **1,000 tokens = 1 RU,** and each RU costs **$0.0001 USD.** Learn more about tokens and tokenization.
- If you are using **Jupyter Notebook editor on watsonx.ai,** consider selecting **a lower runtime environment** to avoid high resource consumption and quickly depleting your credits. Notebook runtimes are billed based on **Capacity Unit Hours (CUH)** at a rate of **$1.02 USD per CUH.** Learn more about capacity unit hours and watsonx.ai Studio pricing plans.

Sample exercises:

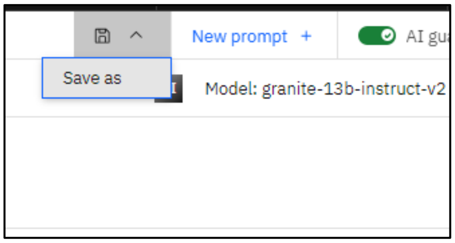
- Prompt a foundation model with the retrieval-augmented generation pattern
- Create a LangChain RAG system in Python with watsonx

- [Build a RAG application with watsonx.ai flows engine]()
- [Notebook: RAG: A simple introduction]()
- [Notebook: Use watsonx, Elasticsearch, and LangChain to answer questions (RAG)]()
- [Notebook: Use watsonx, Chroma, and LangChain to answer questions (RAG)]()
- [Notebook: Build a LangChain agentic RAG system using the Granite model in watsonx.ai]()

Save your Prompt Lab session

You can save your Prompt Lab editor session for later use.

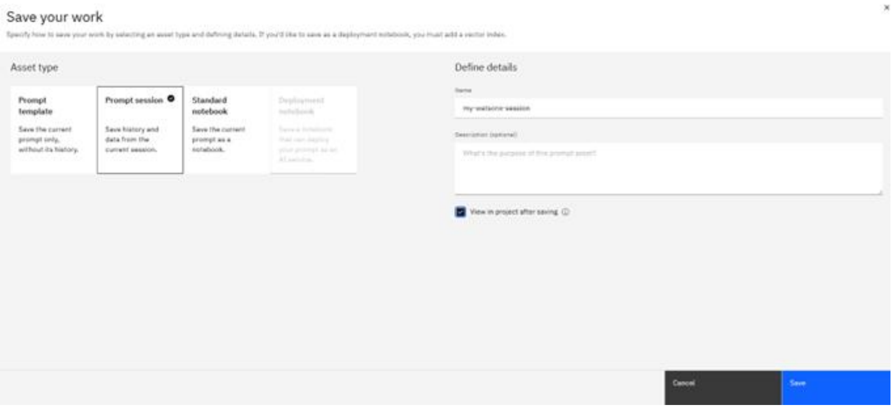
1. At the top of the Prompt Lab screen, select the **Save work** dropdown button and then select the **Save as** option.



2. A **Save your work** widget will appear. Select **Prompt session** under the **Asset type** option.

3. Enter a **name** and check the **View in project after saving** option under the **Define details** section.

4. Finally, click the **Save** button. Once you save, you will see the saved work under the **Assets** tab



Save your work

Specify how to save your work by selecting an asset type and defining details. If you'd like to save as a deployment notebook, you must add a vector index.

Asset type

- Prompt template: Save the current prompt only, without its history.
- Prompt session ●: Save history and data from the current session.
- Standard notebook: Save the current prompt as a notebook.
- Deployment notebook: Save a notebook that can deploy your prompt as an AI service.

Define details

Name: my-watsonx-session

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

☑ View in project after saving

You can also save your work as:

- **Prompt template** to save only the current prompt without its history and selecting a **Task** suitable for your prompting.

- **Notebook** to continue prompting on a Jupyter Notebook environment. Prior knowledge of notebooks and Python programming language would be helpful to work with a Jupyter notebook. Read more about notebooks.

Save your work on watsonx.ai

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

1. Go to your project's 'Overview' tab.
2. Select the 'Export or import project' drop down below the Bell icon in the top menu bar.
3. Click the 'Export project' option. This will open 'Export project to desktop' screen.
4. Select all the assets shown in your project (Work saved as Project session cannot be exported) and click 'Export' on the bottom-right of the screen.
5. The next screen will ask for confirmation that all sensitive information has been removed.
6. Click on 'Continue export'.
7. The download (zip) will be initiated and the file will be saved on your computer.

Appendix: Example use cases

You are not limited to these ideas, but here are several examples for how you could apply RAG and IBM watsonx.ai to solve a specific issue within your chosen theme:

Advance the future of customer experience

- **Personalized product recommendations**: Develop a RAG solution that analyzes customer browsing behavior and purchase history to generate personalized product recommendations, enhancing the shopping experience and increasing sales.
- **Intelligent customer support**: Create virtual assistants that use RAG to provide instant, accurate responses to customer inquiries, reducing wait times and improving satisfaction.
- **Customer sentiment analysis**: Implement a system that uses RAG to analyze customer reviews and social media posts, providing insights into customer sentiment and areas for improvement.
- **Dynamic content generation**: Use RAG to tailor personalized content for emails, newsletters, and websites based on customer preferences and behavior, making interactions more engaging.
- **Predictive customer service**: Develop models that predict customer issues before they arise, allowing businesses to proactively address concerns and enhance overall customer experience.

Coding Challenge, brought to you by Call for Code

- **Extreme weather event prediction**: Build a RAG solution that analyzes historical weather data and current conditions to predict extreme weather events, helping communities prepare and respond effectively.

- **Environmental impact analysis**: Create tools that use RAG to analyze environmental data, tracking progress towards sustainability goals and identifying areas for improvement.
- **Climate education platforms**: Develop interactive educational platforms that use RAG to provide information on climate change, its impacts, and actions individuals can take to mitigate it.
- **Carbon footprint monitoring**: Implement a system that uses RAG to monitor and analyze carbon footprints of businesses and individuals, suggesting ways to reduce emissions or promote transparency.
- **Climate refugee support**: Use RAG to create a multilingual chatbot that provides legal aid, relocation support, and emergency assistance for climate refugees displaced by events like rising sea levels or extreme weather.
- **Sustainable resource management**: Develop models that optimize the use of natural resources, promoting sustainability and reducing environmental impact through intelligent planning and management.