Lean AI Solutions Hackathon, with IBM Granite Models

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.

The latest IBM Granite 3.3 represents another expansion of IBM Granite's multimodal footprint. Headlined by Granite Speech 8B, our first official speech-to-text model, Granite 3.3 marks the beginning of our explorations into audio capabilities. Alongside the recent addition of vision and reasoning capabilities, IBM continues to grow the versatility of the Granite series across the enterprise use cases that customers and the open source community need most. Learn more about Granite 3.3.

The hackathon challenge 2
A note on data sets before you begin 2
Get started with IBM Granite models 3
Option 1: Open-source platforms 3
System requirements 3
IBM Granite documentation 3
Chat template 3
Granite cookbooks 3
Granite Workshop 4
BeeAI Agentic Framework 4
Other Granite quick start samples 4
Option 2: IBM watsonx.ai 5
Note on IBM Cloud service usage 5
Note on available services 5
Access your IBM Cloud account 6
Access Prompt Lab on watsonx.ai platform 8
Work with the watsonx.ai Prompt Lab 9
Programmatic access (API endpoint) 11
Quick start hands-on exercises 15
Save your Prompt Lab session 15
Save your work on watsonx.ai 16
Appendix: Example use cases 17
Make work less boring with IBM Granite models 17
Decent work and economic growth challenge, brought to you by Call for Code 17

The hackathon challenge

In this hackathon, you will **build an AI-driven solution** that harnesses the power of IBM Granite models for **one** of the following hackathon themes:

- **Make work less boring**
  This is your chance to show the world what “different” could look like when it comes to business. Use IBM Granite models and build a proof-of-concept, AI-driven solution that brings creativity into how businesses operate. Whether it's generating new ideas, designing creative content, predicting innovation or market trends, enhancing brainstorming processes, or nurturing creative strategy, you'll identify an activity that can be improved and build a solution that can empower businesses to innovate faster and more creatively.
  Refer to make work less boring example use cases.

- **Decent work and economic growth challenge, brought to you by Call for Code**
  Economic growth and decent work can be a positive force for the whole planet. We're calling on you to build a proof-of-concept, AI-driven solution using IBM Granite models that can address an issue that falls under the United Nations Sustainable Development Goal (SDG) 8: Decent Work and Economic Growth.
  Refer to decent work and economic growth example use cases.

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 use IBM Granite models to build their innovative AI solution by accessing them on one of the following:

1. Open-source platforms
2. IBM watsonx.ai

Option 1: Open-source platforms

To access IBM Granite models through open-source platforms and run them locally, you have the following open-source 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 for agents and other use cases 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 cookbooks

Try sample use case of Granite cookbooks to understand and use different Granite models for your solution use case. Refer Granite cookbooks page.

Granite Workshop

Try the Granite Workshop to get hands-on experience for a few use cases that demonstrates the value of generative AI using Granite models. 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 Agentic Framework

BeeAI is an open-source agentic 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 Granite 3 series.

- Get started with your language of choice
  - BeeAI Framework for Python
    - Installation
    - Try quick example
  - BeeAI Framework for Typescript
    - Installation
    - Try quick example
  - Head to BeeAI GitHub for more documentation and examples.

- New to building AI agents? Try our hands-on labs to ramp up!
  - Get familiar with the basics such as 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.

Other Granite quick start samples

- Build an AI-powered multimodal RAG system with Docling and Granite
- IBM Granite 3.2: Writing / Debugging code with AI in seconds
- Build an AI research agent for image analysis with Granite 3.2 Reasoning and Vision models

Option 2: IBM watsonx.ai

You can access IBM Granite models and use the watsonx.ai SaaS runtime environment for model inferencing, 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 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 **deactivated within an hour**.

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 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.

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
- Cloud Object Storage service

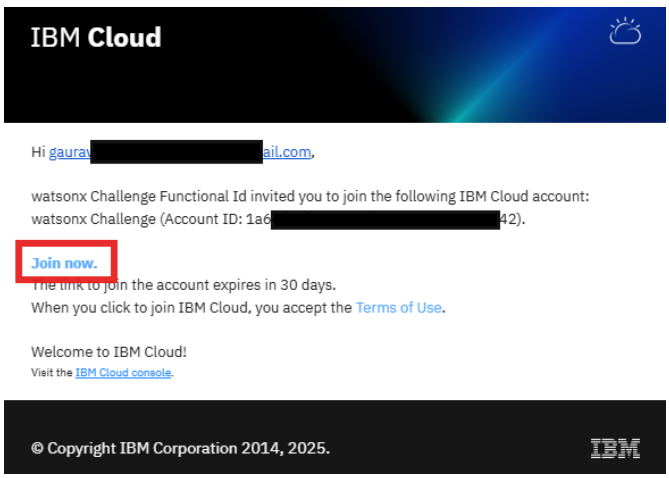
You are required to use ONLY IBM Granite models exclusively on watsonx.ai. Using any other models will negatively impact the judgement of your submission and are out of scope for this hackathon.

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.

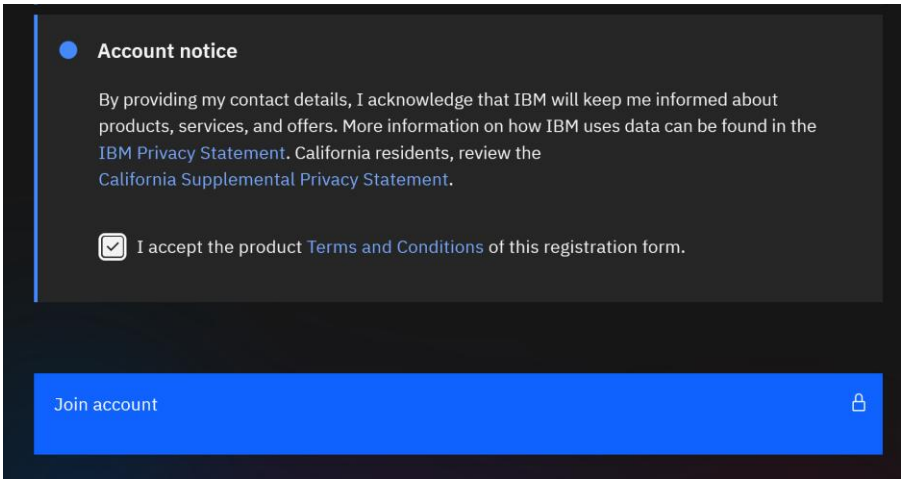
Access your IBM Cloud account

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 below steps to access your team's cloud account:

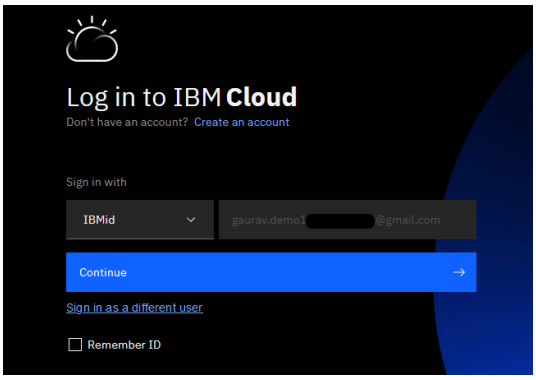
1. Check the email inbox you used to register for the hackathon and open the email you received from the IBM Cloud team about joining your cloud account. Please check your junk/spam folders if you are not able to find the email in your inbox. You can also quickly search for "IBM Cloud" to locate the email.
2. Click the **Join Now** button seen in that email. A new browser tab will open with the cloud account sign up page.



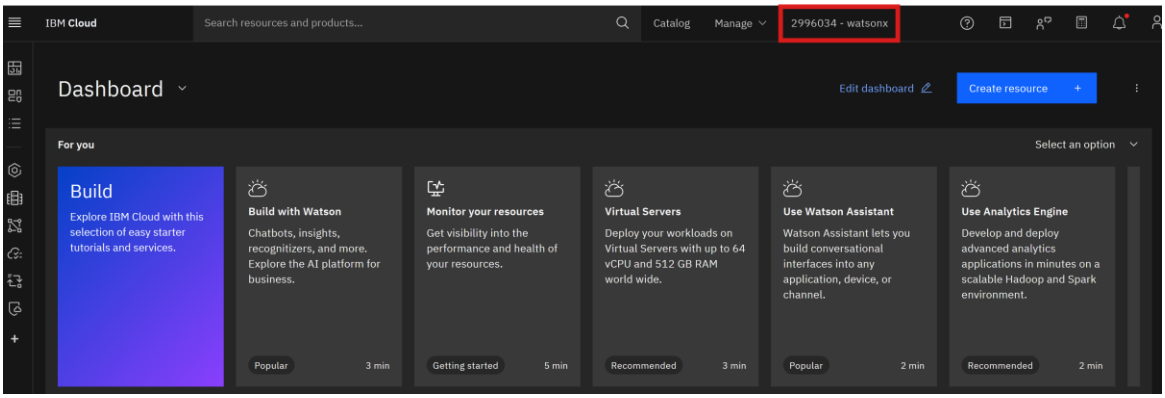
3. Review your account and personal information.
4. Read and accept the Account notice and click the **Join Account** button.



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

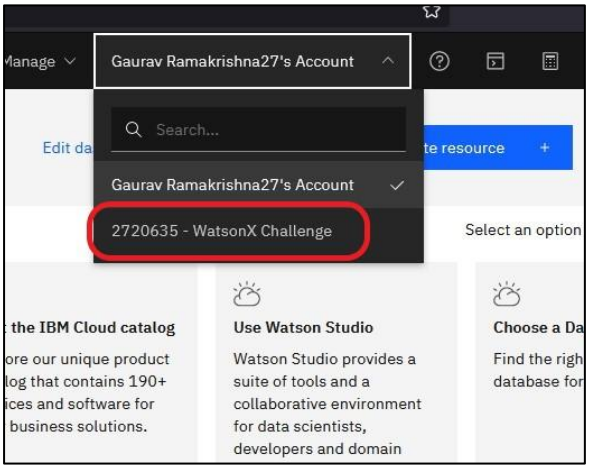


After you authenticate successfully, you will be taken to the IBM Cloud dashboard.



Important note: If you have an existing personal IBM Cloud account for the same email/IBMid, sometimes you will be directed to your personal account.

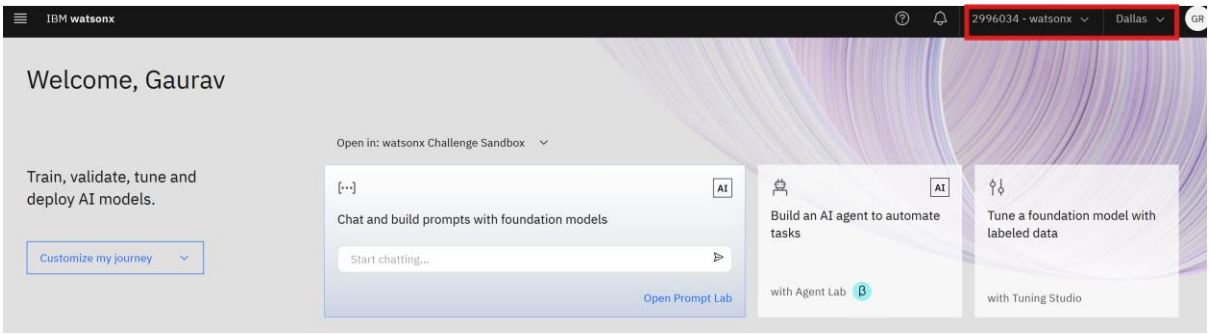
In this case, please switch your account to the **xxxxxxx - watsonx** account. Select your account drop-down at top-right of the dashboard and select watsonx account. Refer to the below image on switching accounts in your cloud dashboard.



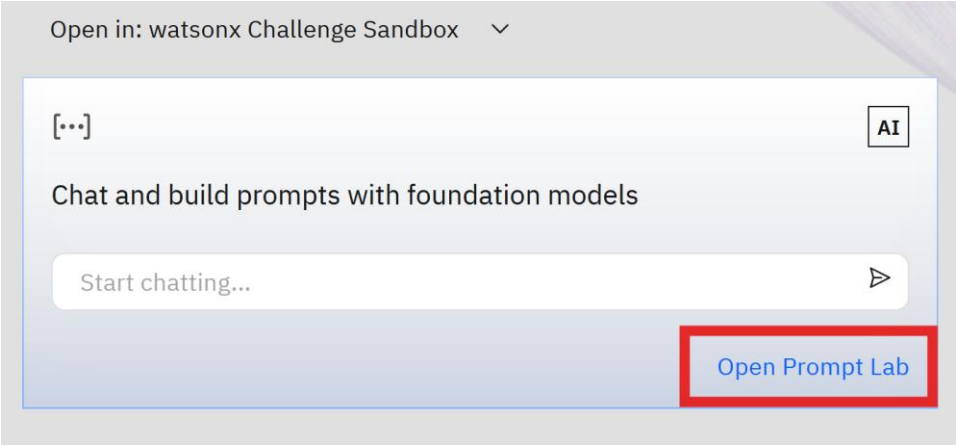
Access Prompt Lab on watsonx.ai platform

After successfully joining the IBM Cloud account, you can now access the Prompt Lab on watsonx.ai platform to work with the Granite models supported on the platform and build your solution.

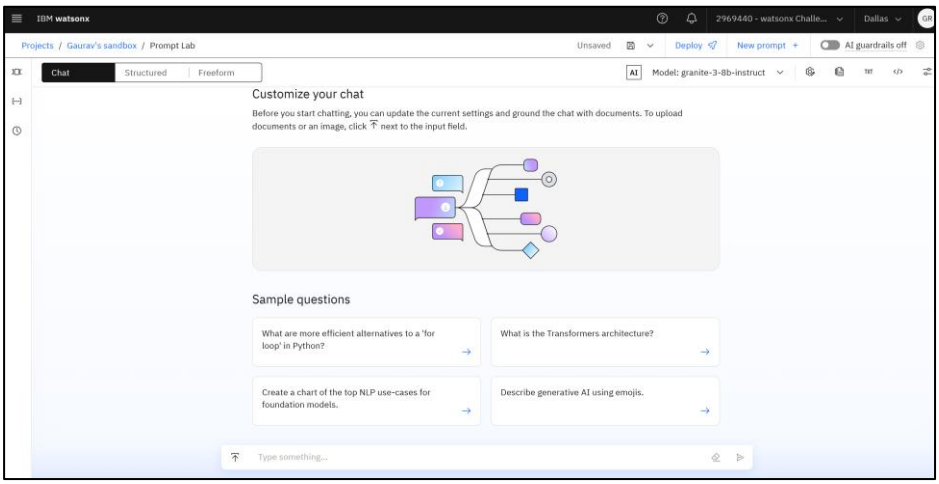
1. Log in to the watsonx.ai platform (https://dataplatform.cloud.ibm.com/wx/home?context=wx) with the email you used to access your IBM Cloud account.
2. After successful authentication, you will see “Welcome to watsonx”. You can either take the tour or skip it.
3. Next, you will see the watsonx.ai dashboard.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



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



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.



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 IBM Granite models to build your innovative solution using Prompt Lab.

Prompt Lab editor

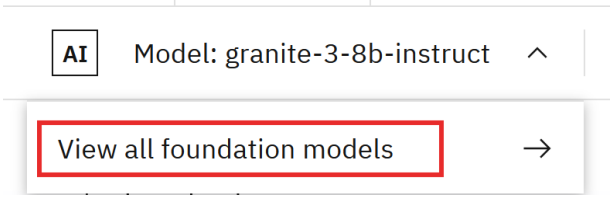
In the Prompt Lab, you can experiment with prompting different foundation models, explore sample prompts, as well as save and share your best promptsThe Prompt Lab editor is a great place to experiment and iterate with your prompts. Try the quick start lab.

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

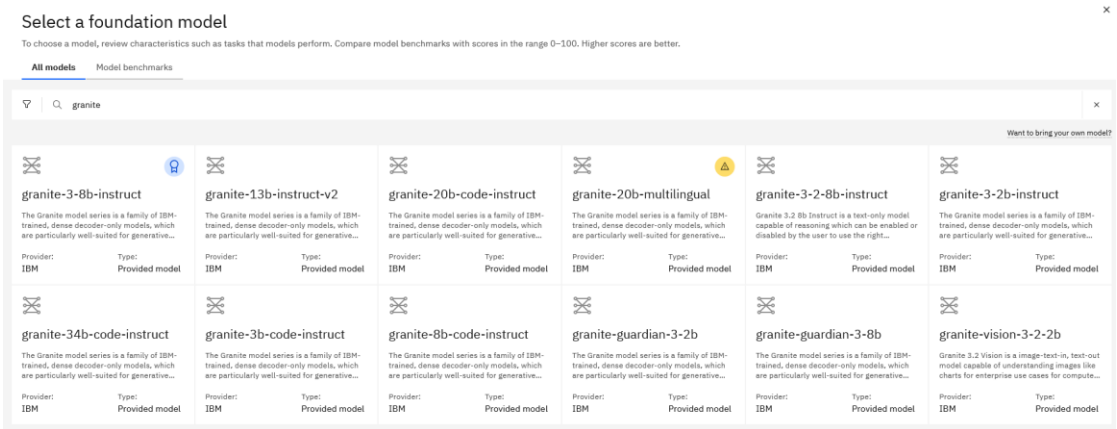
Selecting an IBM Granite model

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 Granite model. To select a different Granite 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 all the filters and enter "granite" in the search bar. All the granite series models will be displayed. You can select a model tile to learn about the model and use it.



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 model is not always better. Learn why smaller models are better and cost effective.

Important: You are required to use **ONLY IBM Granite models** exclusively on watsonx.ai. Using any other models will negatively impact the judgement of your submission and are out of scope for this hackathon.

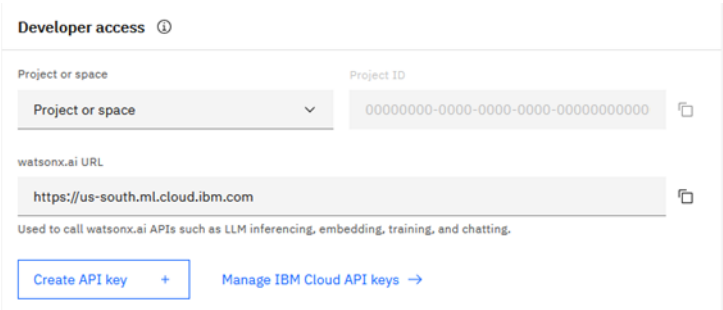
Programmatic access (API endpoint)

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

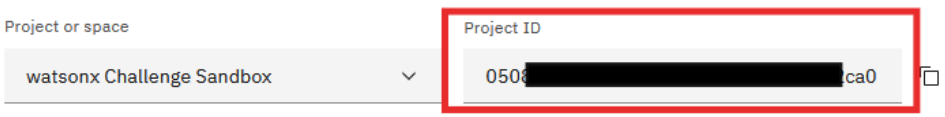
Developer access information

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

1. Go to watsonx.ai home page.
2. Scroll down to the "**Developer access**" section.

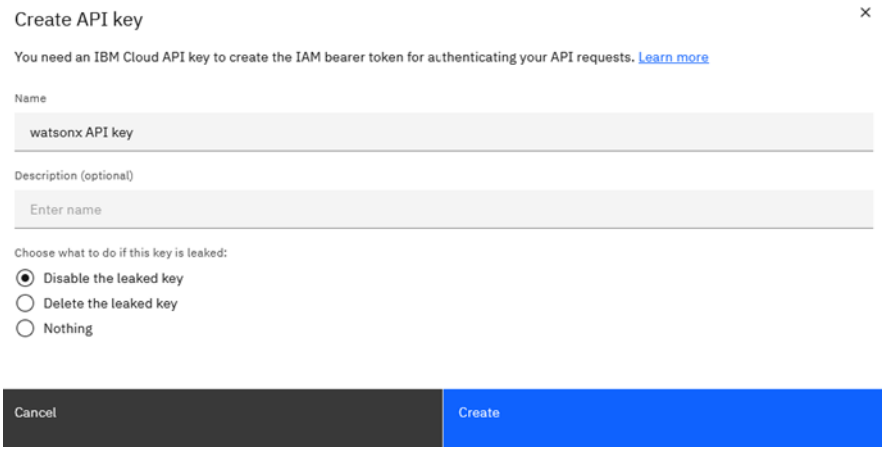


3. Select the "**Project or space**" drop-down and select the "**watsonx Challenge Sandbox**" option. A **project ID** will be displayed.
   Note: A space ID is not required as it is out of scope for the hackathon.

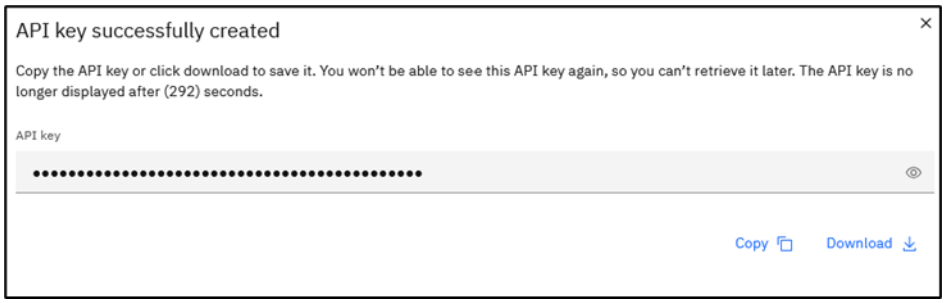


4. A default **watsonx.ai endpoint URL** will be displayed for the Dallas region. Ensure the region is always set to **Dallas** at the top right of the watsonx.ai home page.

5. 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.



6. 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.



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.

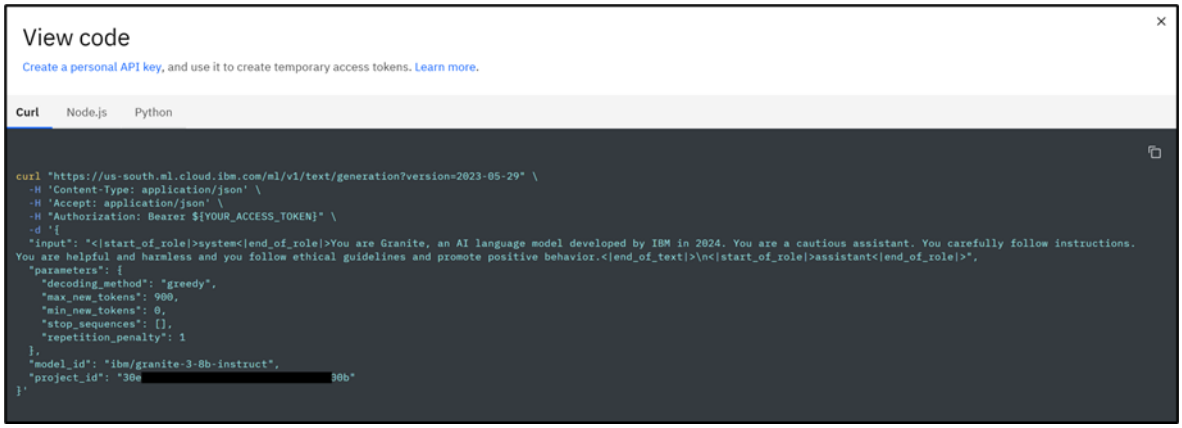
- Chat
- Tool calling
- Text generation
- Time series
- Text rerank
- Embeddings
- Text extraction

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.



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,
        "scope": “ibm openid”
}
```

  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 usage of old Granite model versions or non-Granite model. You can replace them with newer versions of Granite for better performance and output. To check the latest supported Granite models on watsonx.ai, either follow selecting an IBM Granite model on Prompt Lab or refer to supported Granite models on watsonx.ai.

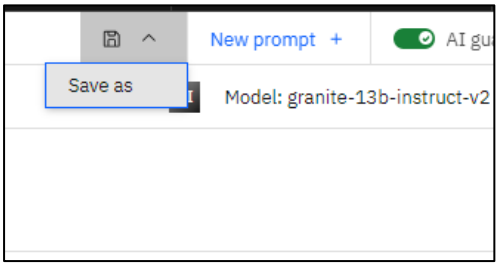
Sample exercises:

1. Build an AI stylist with IBM Granite using watsonx.ai
2. Getting Started with Time Series Models on IBM watsonx.ai
3. Build an agentic framework with CrewAI and IBM watsonx.ai

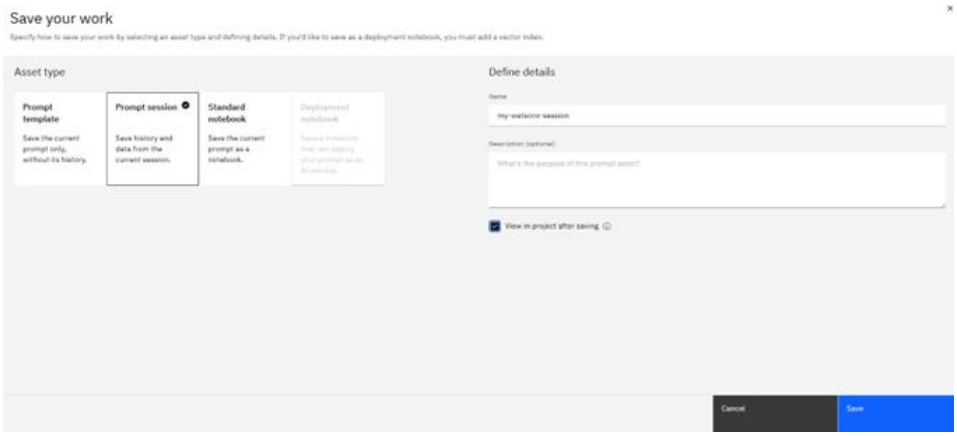
Save your Prompt Lab session

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



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 below steps 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 a several examples for how you could apply IBM Granite models to solve a specific issue within your chosen theme:

Make work less boring with IBM Granite models

- **Personalize learning and skill development**: Facilitate self-guided learning by using Granite LLMs as intelligent tutors or coaches to keep employees engaged. Create tailored learning paths and recommend relevant courses, articles, and tools to help employees acquire new skills.

- **Facilitate interactive collaboration and brainstorming**: Use Granite LLMs as smart collaborators during brainstorming sessions to generate fresh ideas, challenge assumptions, provide problem-solving suggestions, and synthesize input from team members to develop innovative solutions.

- **Enable intuitive communications**: Automate and improve communication tasks such as email responses, reminders, scheduling, and interactions with clients or teammates, making everyday communication more insightful.

- **Generate intelligent content for creative projects**: Assist with content creation by helping design or generate draft text, marketing materials, reports, or presentations, introducing fresh ideas and streamlining creative processes.

- **Automate and optimize personal tasks and workflows**: Automate routine tasks like data entry, report generation, and scheduling, freeing up time for more meaningful, engaging, and creative work.

Decent work and economic growth challenge, brought to you by Call for Code

- **Build AI-driven compliance, documentation, and workforce management**: Generate customized contracts, employment agreements, or legal documents to ensure workers' rights are protected and labor laws are adhered to.

- **Build AI-powered career guidance or job matching for young workers**: Leverage labor market data and individual profiles to offer personalized career recommendations, including matching youth with suitable training programs, certifications, or internships aligned with market demand and their skillsets.

- **Automate the creation of business plans or financial guidance for individuals**: Analyze market trends, identify funding sources, assist in creating investment pitches, or generate detailed business plans or risk assessments so that aspiring entrepreneurs, small business owners, or anyone looking to start a new venture could access professional-grade resources without needing specialized expertise.

- **Help businesses optimize and scale to create more job opportunities**: Improve operational efficiency so businesses can scale what they do, increase their production, and potentially hire more people for more dynamic, sustainable roles. Analyze operational data, identify inefficiencies, and provide recommendations to improve production processes, inventory management, or logistics. Identify opportunities for employees to train on new tools or systems, enhancing their skills and opening up opportunities for career advancement in more specialized roles.

- **Promote decent work for persons with disabilities**: Suggest inclusive employment programs that ensure accessible workplaces or anti-discrimination policies for persons with disabilities. Recommend opportunities for marginalized groups, promote workforce diversity, or drive innovation through inclusive practices.