Build a generative AI application for a use case supporting productivity with IBM watsonx.ai, featuring IBM Granite

IBM watsonx.ai is a powerful platform that brings together the capabilities of IBM's Watson services, including natural language processing, machine learning, and automation, to create intelligent solutions tailored to your business needs. You can utilize watsonx generative AI technology to improve productivity by quickly extracting knowledge and insights from vast amounts of data across multiple documents.

The Prompt Lab is a powerful tool within watsonx.ai that enables you to interactively explore the capabilities of AI models and generate human-like text or code snippets. Here are some ways the watsonx.ai Prompt Lab can enhance day-to-day tasks:

- **Interactive exploration**: The Prompt Lab allows you to experiment with different AI models and parameters, giving you a deeper understanding of how they work and how they can be applied to your specific use cases.
- **Human-like text generation**: You can use the Prompt Lab to generate human-like text, for things like responses to customer inquiries or draft emails, saving you time and effort.
- **Code snippet generation**: The Prompt Lab can also generate code snippets for various programming languages, enabling you to automate tasks and integrate AI models into your workflows more efficiently.
- **Training assistance**: The Prompt Lab can help you train AI models by providing feedback on your input and suggesting improvements, making it easier to optimize your models for your specific needs.

Table of contents

The hackathon challenge

In this hackathon, you must use **watsonx.ai Prompt Lab** to get creative and build an intelligent solution that can significantly improve productivity and streamline an everyday activity.

Participants have the flexibility to determine the business or societal process they want to improve and can import their own datasets (please see the Data sets section).

Below are some example use cases to help you get started:

- **Customer service**: Empower customers to find solutions with easy, compelling experiences and automate answers with high accuracy.
- **Marketing**: Increase personalization and improve efficiency across the content supply chain to reduce content creation costs.
- **Communications**: Summarize key points from calls and discussions to share with user community to improve engagement.
- **Content creation**: Enhance digital sports viewing with auto-generative spoken AI commentary and scale live viewing experiences cost-effectively.
- **Knowledge worker**: Enable higher value work, improve decision making, increase productivity to reduce time of text reading and analysis work.
- **HR automation**: Reduce manual work and automate recruiting sourcing and nurturing job candidates to reduce employee mobility processing time.
- **Training and skills development**: Summarize long training documents before investing time to read and determine if the education will meet learning needs.
- **Supply chain**: Automate source-to-pay processes, reduce resource needs, and improve cycle times to reduce cost per invoice.
- **Planning and analysis**: Make smarter decisions and focus on higher-value tasks with automated workflows and AI for faster process planning data.
- **Regulatory compliance**: Support compliance based on requirements/risks, and proactively respond to regulatory changes to reduce time spent responding to issues.

- **App modernization, migration**: Generate code, and tune code generation response in real time to deliver faster development output.
- **IT automation**: Identify deployment issues, avoid incidents, and optimize application demand to supply to reduce mean time to repair (MTTR).
- **AIOps**: Assure continuous, cost-effective performance and connectivity across applications to reduce application support tickets.
- **Data platform engineering**: Redesign the approach for data integration using generative AI to reduce data integration time
- **Threat management**: Reduce incident response times from hours to minutes or seconds to contain potential threats faster.
- **Asset management**: Optimize critical asset performance and operations while delivering sustainable outcomes to reduce unplanned downtime.
- **Environmental intelligence**: Provide intelligence to proactively manage the impact of severe weather and climate to increase manufacturing output.
- **Agriculture**: Create marketing and sales campaigns based on crop output to improve produce sales.

Although, the watsonx.ai Prompt Lab offers multiple foundation model options to work, participants are encouraged to use the IBM Granite series models. IBM Granite models are a family of artificial intelligence (AI) models built for business, to help drive trust and scalability in AI-driven applications.

Before you begin

IBM Cloud account set up

Follow the instructions to set up your IBM Cloud account for the hackathon. Each participant will be provisioned with an individual IBM Cloud account for the hackathon. You will not be able to add your teammate to your account and have to collaborate offline/virtually to build your solution together on either of your assigned IBM Cloud accounts.

Note on IBM Cloud service usage

For this hackathon IBM is providing you an additional $100 credit to use towards IBM Cloud services, above the free limit that comes with the IBM Cloud account that is being provisioned for you. 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%. 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 bring their own model or fine tune models. These features are out of scope for this hackathon.

Data sets

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:

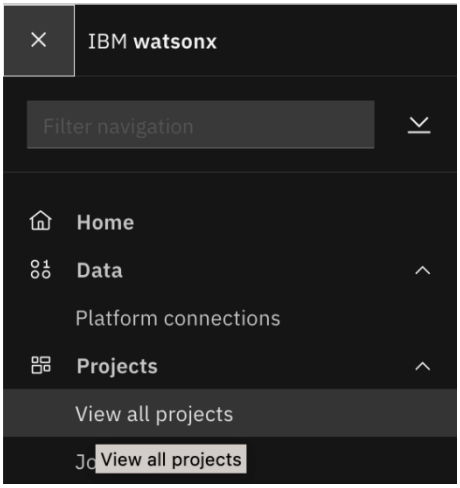
- Teams are responsible for ensuring data is compliant.
- Data from public websites may be used, if the terms allow for commercial use, but please keep a list of the websites you use.
- Do not use data or assets containing company confidential data, or any other data without permission from the data 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 the watsonx.ai Prompt Lab

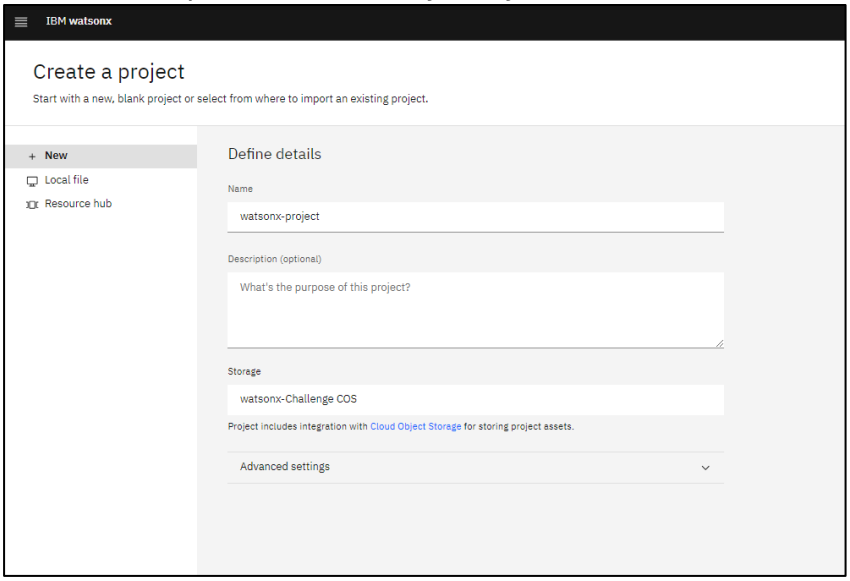
1. Create a project

Before you start using Prompt Lab, you will first need to create a project.

1. Log in to the watsonx.ai platform (https://dataplatform.cloud.ibm.com/wx/home?context=wx) with the email you use to access your IBM Cloud account.
2. From the navigation menu at the top-left corner, select **View all projects**.



3. Select **New Project > Create an empty project**.
4. Give the project any name. In the **Storage** panel, “watsonx – Challenge COS” will be pre-selected for you by default.

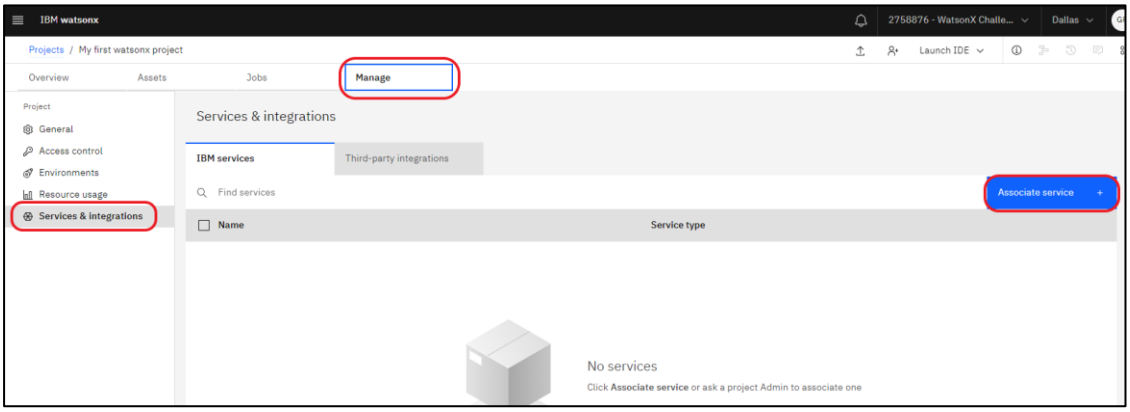


5. Click **Create**. You will be taken to the project’s overview dashboard.

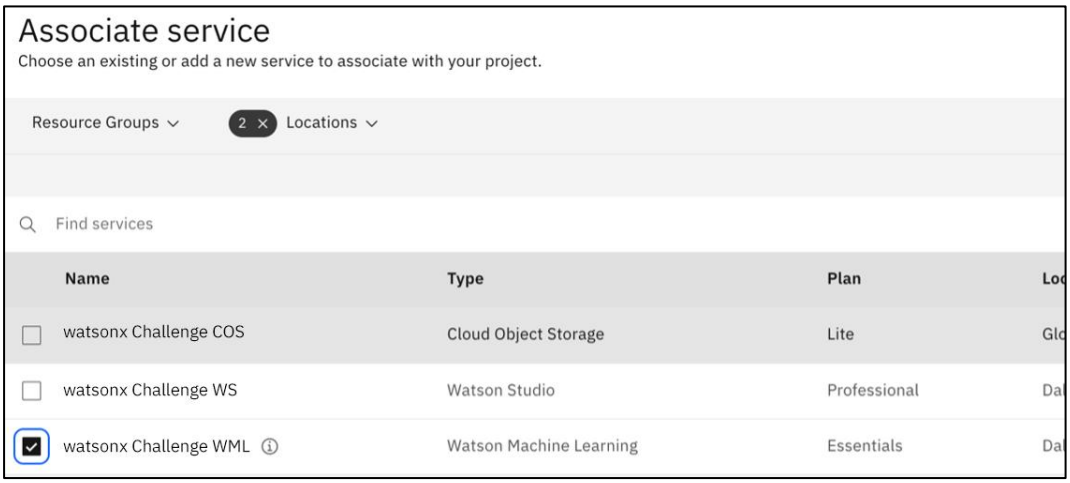
2. Associate a Machine Learning service

You must associate a machine learning service for your project to successfully work with the Prompt Lab.

1. Select the **Manage** tab on your project dashboard, select the **Services & integrations** on the left pane, and click **Associate service +**.



2. From the list of services, select the "**watsonx-Challenge WML**" Watson Machine Learning instance and then click **Associate** button to complete associating the machine learning service.

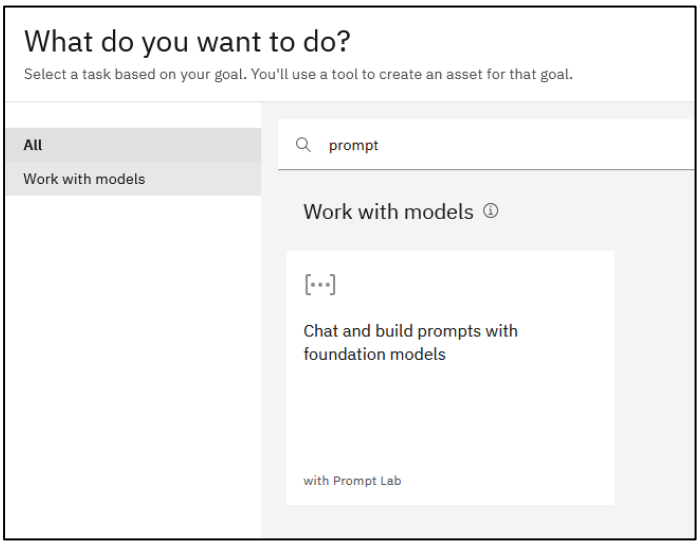


3. Access the Prompt Lab

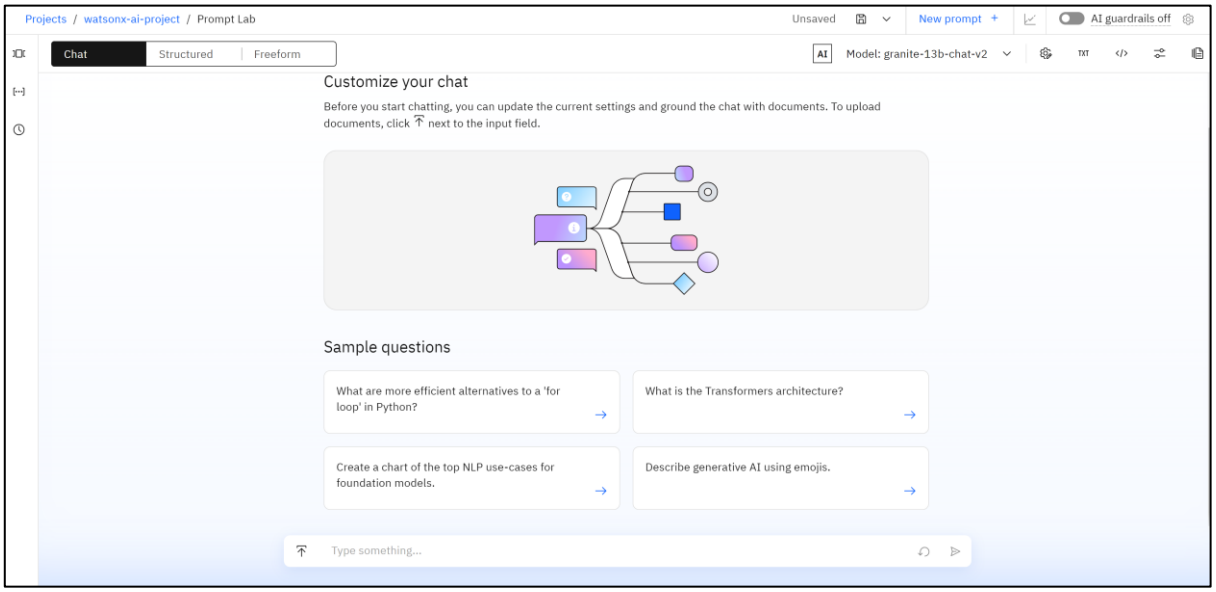
Now you can access the Prompt Lab and start building your innovative solution.

1. Select the **Assets** tab, click **New asset +**, and search for "prompt".

2. Select the “**Chat and build prompts with foundational models**” tile.



3. The Prompt Lab Editor opens with a chat window to get you started with the prompt session.



4. Working with the Prompt Lab

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

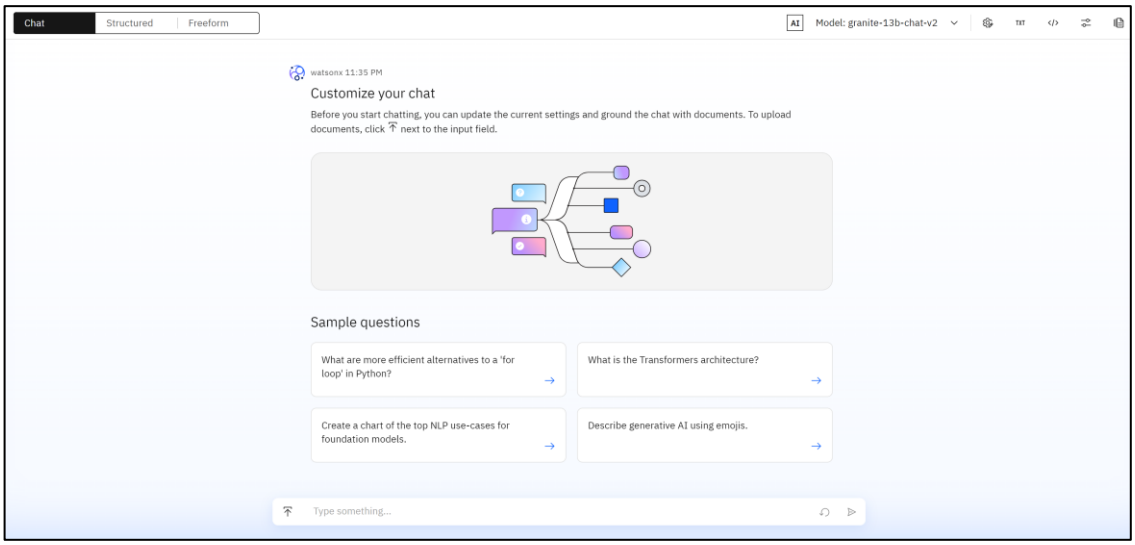
Take a tour of the Prompt Lab.

a. Create and run a prompt

Prompt Lab comes with 3 modes of editor options to enter your prompts and generate a response.

1. Chat mode

You can chat with the foundation model to see how the model handles dialog or question-answering tasks.



Start the chat by submitting a query or request for the foundation model to answer. Alternatively, you can click a quick start sample to submit to the model. Quick start samples are sent to the Llama foundation model. If you want to work with the Granite foundation model or a different model, add your own prompt text.

Take a tour of working with the Chat mode in Prompt Lab video or try the using Chat mode to prompt a foundational model guided exercise.

Clear chat:

Click **Clear chat** icon to stop and make changes. You cannot make changes while a chat is in progress.

System prompt:

Predefined text called a *system prompt* is included at the start of the chat to establish ground rules for the conversation. To review and customize the text, click the **Edit system prompt** icon located at the top right of the chat editor.

Prompt text:

You can click the **View full prompt text** icon located at the top right of the chat editor, to see the full prompt text that will be submitted to the foundation model.

Grounding foundational model:

To help the foundation model to return current, factual and relevant output, ground the model with relevant information. This pattern, which is known as retrieval-augmented generation (RAG), is especially helpful in question-answering scenarios where you want the foundation model to generate predictable, accurate answers. For more information about RAG, see Retrieval-augmented generation (RAG).

Add documents to ground your foundation model in data from a vector index. If you upload documents, an in-memory vector index is created to search your documents.

Click the **Upload documents** icon , and then choose **Add documents**.

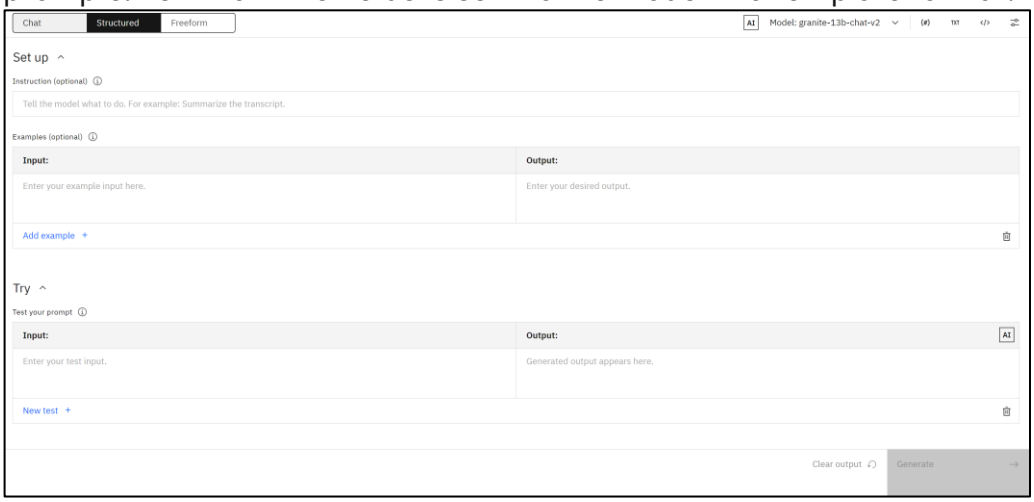
Try the using Chat mode to ground the foundational model with information guided exercise to learn how to upload documents to ground the foundational model. For more information, see Grounding foundation model prompts in contextual information.

You can also add relevant data from a third-party vector store. Click the **Grounding with documents** icon located at the top right of the chat editor, and then select the vector index.

For more information, see Adding grounding data for generative AI models.

2. Structured mode

Structured mode is designed to help new users create effective prompts. Text from the fields is sent to the model in a template format.



You add parts of your prompt into the appropriate fields:

- **Instruction**: Add an instruction if it makes sense for your use case. An instruction is an imperative statement, such as “Summarize the following article.”

- **Examples**: Add one or more pairs of examples that contain the input and the corresponding output that you want. Providing a few example input-and-output pairs in your prompt is called few-shot prompting.

  If you need a specific prefix to the input or the output, you can replace the default labels, “Input:” or “Output:”, with the labels you want to use. For example, you might replace the default labels with custom labels that were used in training data when a foundation model was prompt-tuned.

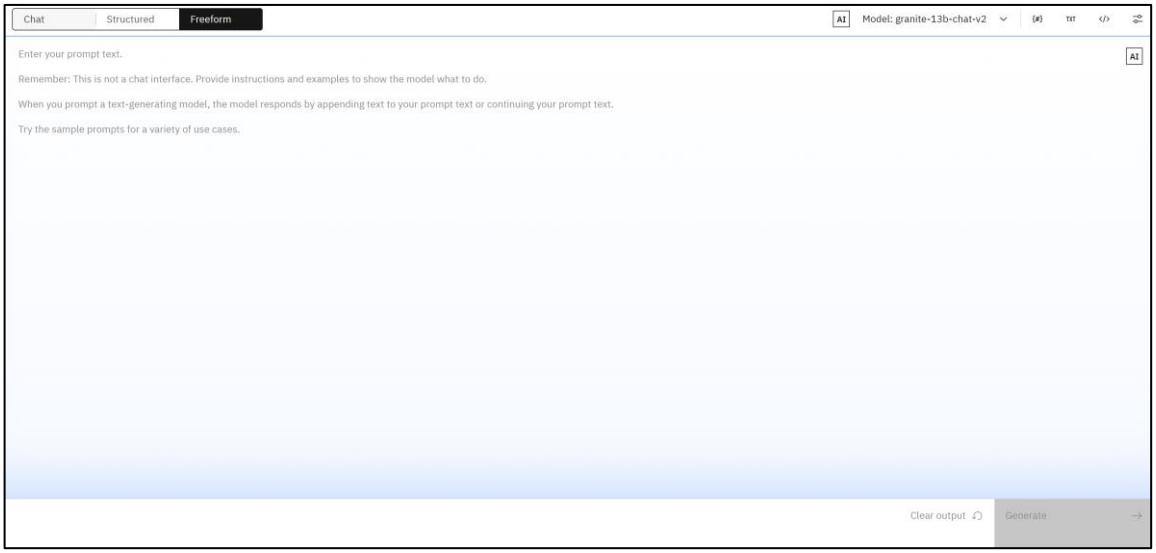
  A space is added between the example label and the example text.

- **Test your input**: In the Try area, enter the final input of your prompt.

Explore and try the sample prompts available on the Prompt Lab.

3. Freeform mode

You add your prompt in plain text. Your prompt text is sent to the model exactly as you typed it.



Freeform mode is a good choice when you want to submit structured input and know how to format the prompt.

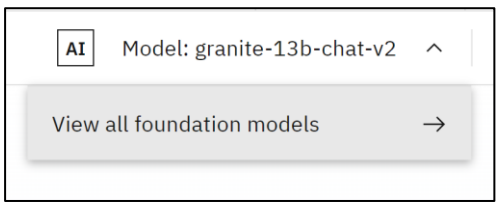
b. Model and prompt configuration options

1. Model choices

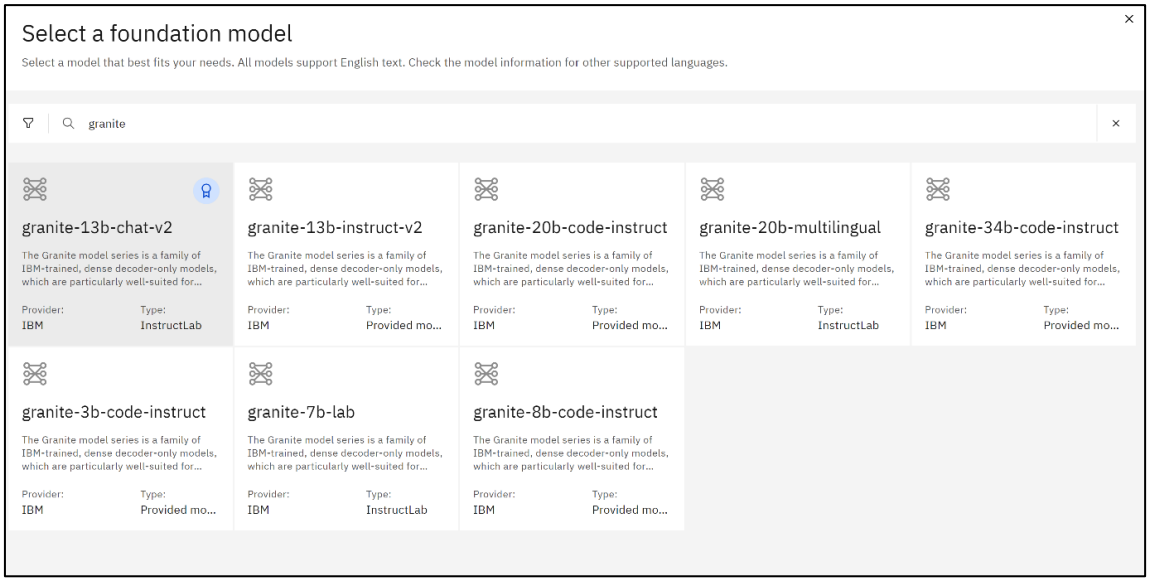
In the Prompt Lab, you can submit your prompt to any of the models that are supported by watsonx.ai. You can choose recently-used models from the drop-down list. Or you can click **View all foundation models** to view all the supported models, filter them by task, and read high-level information about the models.

To view and select different IBM 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 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 for your prompting session.

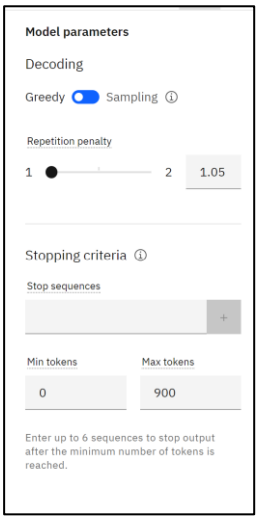


2. Model parameters

To control how the model generates output in response to your prompt, you can specify decoding parameters and stopping criteria. To edit the model parameters:

1. Select the model parameters icon at the top-right of the editor.

2. Default model parameters configuration will be displayed to make required tuning.

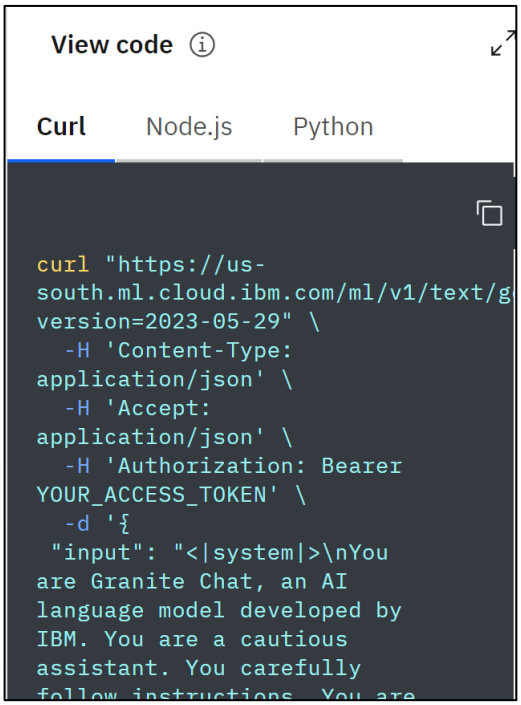


For more information, see Model parameters for prompting.

3. Prompt code (API endpoint)

If you want to run the prompt 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.



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 that you created at the beginning while setting up your IBM Cloud account.

1. Follow Step 2 in the IBM Cloud setup instructions to create an IBM Cloud API key.
2. Programmatically generate an IAM access token with the IBM Cloud API key following the instructions.

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.

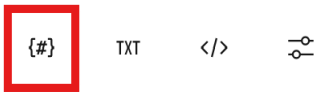
4. Prompt variables

To add flexibility to your prompts, you can define prompt variables. A prompt variable is a placeholder keyword that you include in the static text of your prompt at creation time and replace with text dynamically at run time.

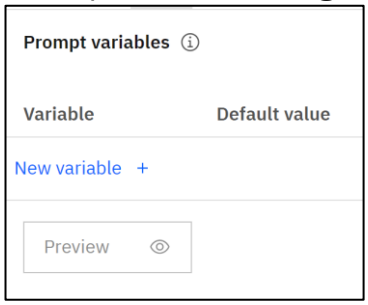
NOTE: You cannot define prompt variables in chat mode; this is supported only in the freeform and structured mode.

To add/edit the prompt variables:

1. Select the model parameters icon at top right of the editor.



2. Prompt variable configuration will be opened to add/edit variables.



For more information, see Building reusable prompts.

5. Prompt text

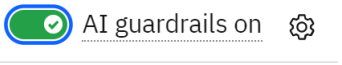
You might want to see the full prompt text that will be submitted to the foundation model in the following situations:

- When prompt variables are in use, to see resolved variable values in context.
- In chat mode, where the recommended prompt formats for different foundation models are applied automatically.
- In structured mode, where you add parts of the prompt into separate fields.

  Select the TXT icon at the top right of the editor to see your full prompt text.

6. AI guardrails

When you set the **AI guardrails** switcher to **On**, harmful language is automatically removed from the input prompt text and from the output that is generated by the model. Specifically, any sentence in the input or output that contains harmful language is replaced with a message that says that potentially harmful text was removed.



Note: This feature is supported for English-language models only. If you're working with a non-English foundation model, disable *AI guardrails*.

For more information, see Removing harmful content.

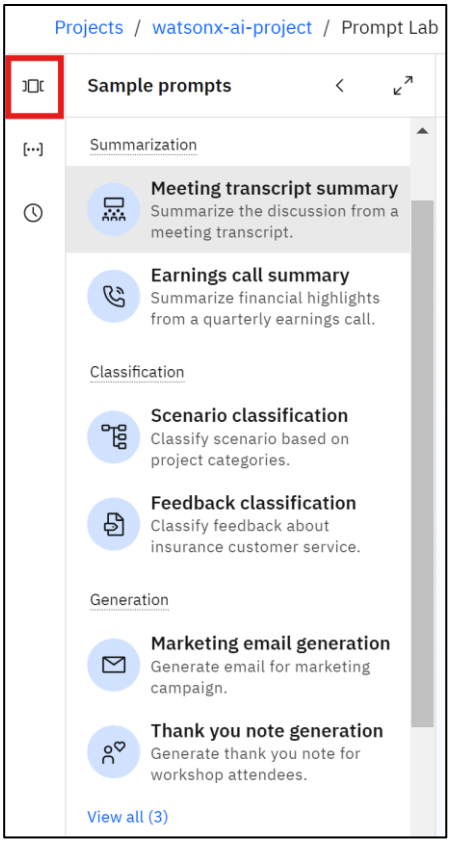
c. Sample prompts

A collection of sample prompts is available in the Prompt Lab. The samples demonstrate effective prompt text and model parameters for different tasks, including classification, extraction, content generation, question answering, and summarization.

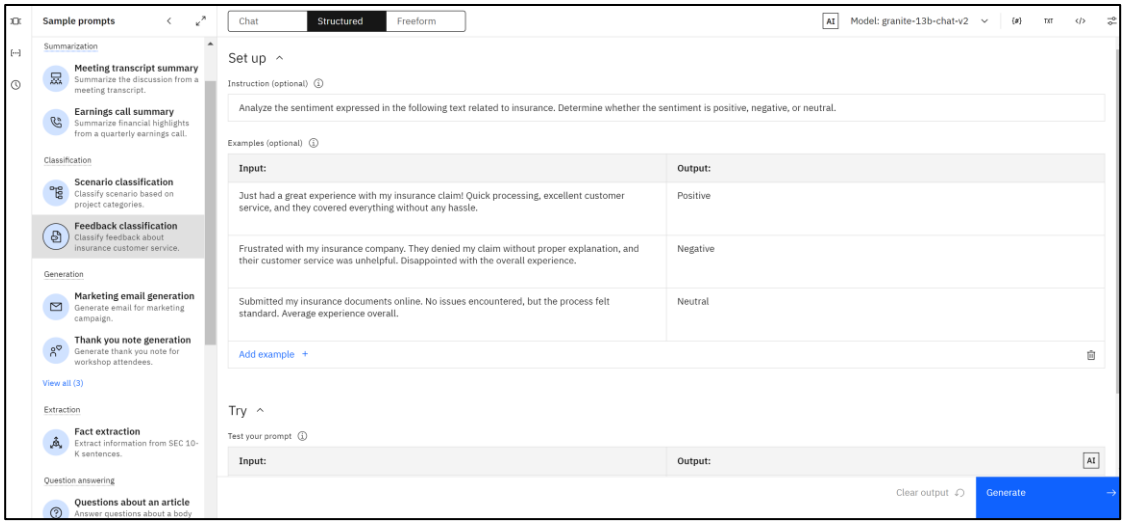
When you click a sample, the prompt text loads in the editor, an appropriate model is selected, and optimal parameters are configured automatically.

To try a sample prompt:

1. Select the "**Sample prompts**" icon on the left pane of the editor. You can see the list of all the sample prompts for different type of tasks.



2. Select the "**Feedback classification**" sample prompt.



Notice the structured mode editor is auto populated with set up instruction and a few examples. This sample is a few-shot learning prompt, meaning the model is trained with a few examples to provide a context input and the response output format.

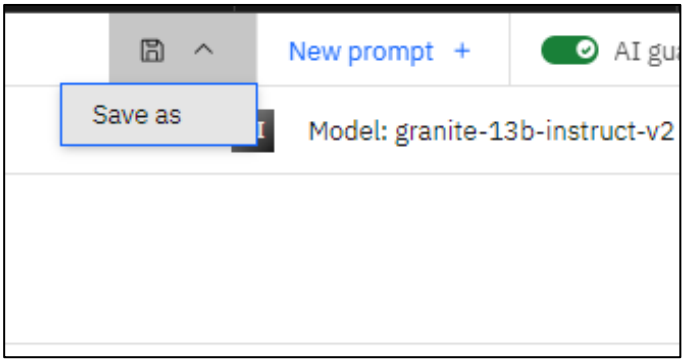
3. Change the AI model to "**granite-13b-instruct-v2**" by selecting Model drop down > **View all foundation models**.
4. Notice a test prompt is auto-populated in the **Try** section. Now click the **Generate** button. The output will be **positive**.

Now try yourself with different test input prompts and check the output classification response of the model. Add more examples to train the model and try modifying the model parameters to generate more accurate responses. Explore other sample prompts for different tasks.

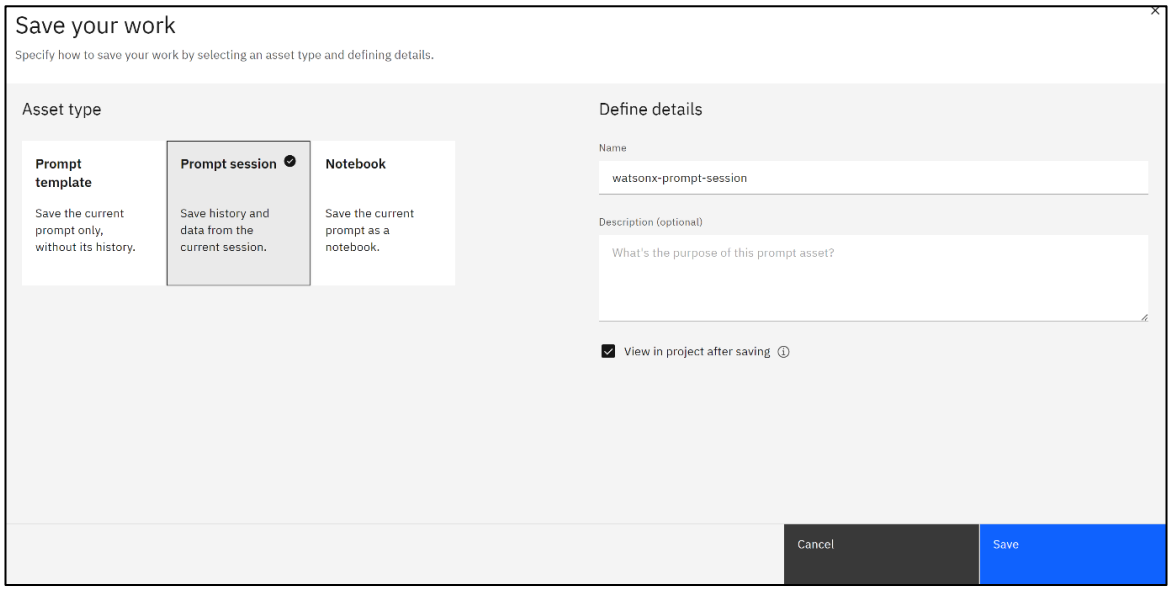
d. Saving your prompt

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.



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.

e. 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

Final submission

To compete in the hackathon, your team will need to submit the following deliverables:

- Video demonstration of the team's solution
- Written problem and solution statements, including how IBM watsonx would be utilized
- Description of productivity improvement
- Link/URL to publicly accessible example of your working solution or code (optional)

See the hackathon web site for more details.

Other useful resources

Refer to below resources to learn more about the watsonx.ai capabilities and leverage the technology to build your innovative solution.

- Prompt tips
- Prompt a foundation model using Prompt Lab
- Prompt a foundation model with the retrieval-augmented generation pattern
- Retrievel-augmented generation (RAG)
- Analyze data in a Jupyter notebook
- Adding Data to the project