



# **Lab Guide: Getting Started with Generative AI in watsonx.ai**

Author: Elena Lowery, elowery@us.ibm.com



## Contents

<b>Overview .....</b>	<b>3</b>
<b>Required software, access, and files .....</b>	<b>3</b>
<b>Generative AI and Large Language Models .....</b>	<b>4</b>
<b>Part 1: Understand LLM capabilities.....</b>	<b>6</b>
<b>Part 2: Integrate LLMs with applications .....</b>	<b>31</b>
<b>Conclusion .....</b>	<b>40</b>

## Overview

In this lab you will learn how to implement generative AI use cases in **watsonx.ai**. *Watsonx.ai* is an AI platform which can be used to implement both traditional machine learning use cases and use cases that utilize *Large Language Models* (LLMs).

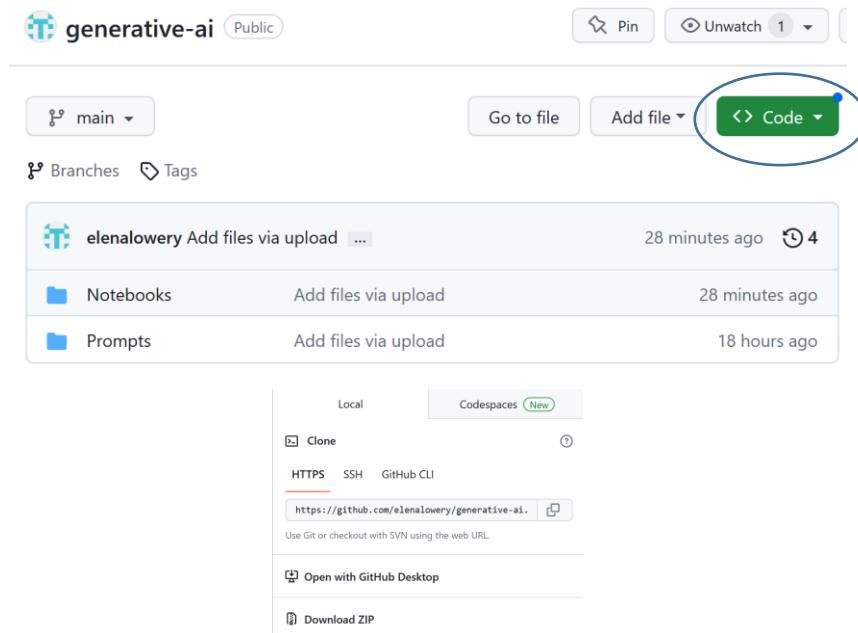
We will take a closer look at the following use cases:

- Generation
- Summarization
- Classification

*Note: LLMs are a type of a foundation model. In IBM tools and documentation, the terms LLM and foundation models are used interchangeably.*

## Required software, access, and files

- To complete this lab, you will need access to *watsonx.ai*. You can get access by signing up for an [IBM Cloud account](#) and provisioning the *watsonx.ai* service.
- Recommended, but not required: A Python IDE with Python 3.10 or 3.11 environment (*Visual Studio Code* or *PyCharm*). If you don't have a Python IDE, you can test integration with LLMs in *watsonx.ai* Notebooks.
- You will also need to download and unzip this GitHub repository: <https://github.com/elenalowery/generative-ai>
- Click the **Code** button and select **Download ZIP**



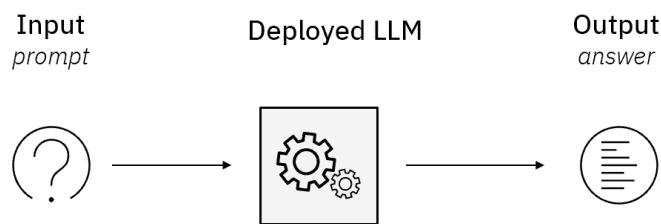
- Unzip the downloaded zip file. In the lab we will refer to this folder as the *git repo* folder.

**Important note:** Some screenshots in the lab may be slightly different from the product. If you have questions, please ask your workshop instructor.

## Generative AI and Large Language Models

Generative AI is a new domain in AI which allows users to interact with models using natural language. When working with Large Language Models (LLMs), a user sends requests (*prompts*) to a model, and the model generates a response. To an end user generative AI may look like a chatbot or a search engine, but implementation of generative AI is different from legacy chatbots that rely on hardcoded business rules and search engines that use indexing.

*Example: Text generation*



**Figure 1: Prompt Lab in watsonx.ai**

Structured    Freeform    Model: flan-t5-xxl-11b ▾ | </> ▾

Generate a 5 sentence marketing message for a company with the given characteristics.

Characteristics:

- Company - Golden Bank
- Offer includes - no fees, 2% interest rate, no minimum balance
- Tone - informative
- Response requested - click the link
- End date - July 15

Prompt

Generated email:

We've got great news for you! Golden Bank is now offering Online Banking -- a better tool to manage your money. It's no-fee, and it offers a higher interest rate on new balances. Best of all, there is no minimum balance required. If you're interested, click on the link below to sign up.

Output

Unlike traditional machine learning models, which always require training, LLMs are pretrained on a very large dataset. There are dozens of LLMs, which are developed by different companies. Some companies contribute their models to open source, and many of them are available on the LLM community site [Hugging Face](#). It's up to the LLM developer to publish information about the model and the dataset that the model has been trained on. For example, see the [model card](#) for one of the popular open source models, *Mixtral-8x7B-Instruct-v0.1*. In general, all LLMs are trained on publicly available data. IBM is one of the few companies that publishes detailed information about the data that was used to train the model. This information can be found in [documentation](#) and [Research Papers](#) published by IBM.

You can think of the data that the model has been trained on as its "knowledge". For example, if the model was not trained on a dataset that contained *2022 Soccer World Cup* results, it will not be able to generate valid/correct answers related to this event. Of course, this applies to all business use cases in which we need models to interact with proprietary enterprise data. We will explain how to solve this problem later in the lab.



Two more factors influence LLM capabilities: size and instruction-tuning. Larger models have been trained on more data and have more parameters. In the context of LLMs, the number of parameters refers to the number of adjustable weights in the neural network that the model uses to generate responses. Parameters are the internal variables of the model that are learned during the training process and represent the knowledge the model has acquired.

While it may seem obvious that a larger model will produce better results, in a production implementation we may need to consider smaller models that meet our use case requirements because of the hosting and inferencing cost.

Instruction-tuned models are models that have been specifically trained for tasks such as classification or summarization. See [IBM documentation](#) for other considerations when choosing a model.

As we work through the lab, we will introduce a few more important LLM concepts.

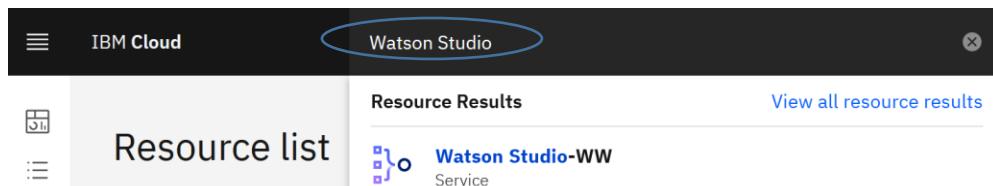
## Part 1: Understand LLM capabilities

In this lab, we will use the **Prompt Lab** in *watsonx.ai* to interact with LLMs included in the platform.

Typically, users (prompt engineers or data scientists) have three goals in this phase of the LLM lifecycle:

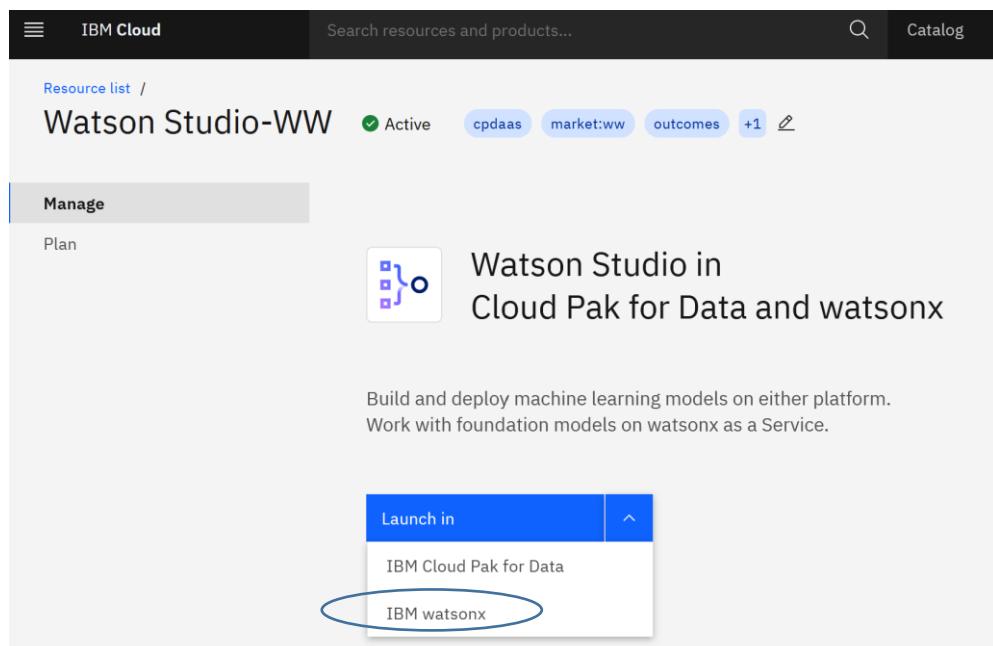
- Find if LLMs can be used for the proposed use case
- Identify the best model and parameters
- Create prompts for the use case.

1. Log in to your **IBM Cloud account**. Your lab instructor will provide the URL and userid/password.
2. If you are not already in *watsonx.ai* service, you can navigate to it from the **IBM Cloud Resources** page.
  - Search for *Watson Studio* in the search box on top of the page.

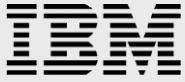


The screenshot shows the IBM Cloud Resources interface. At the top, there's a search bar with "Watson Studio" typed into it. Below the search bar, the "Resource list" is displayed, showing one result: "Watson Studio-WW Service".

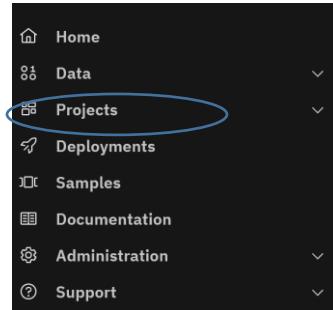
- On the *Watson Studio* service page, make sure to select *Launch in IBM watsonx*.



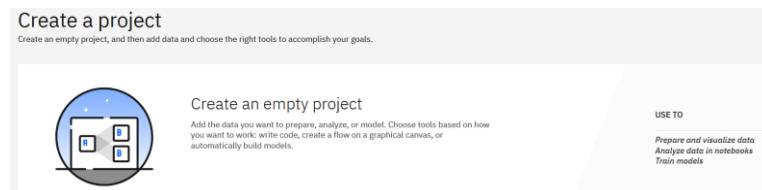
The screenshot shows the "Watson Studio-WW" service page in the IBM Cloud interface. The "Manage" tab is selected. In the center, there's a summary card for "Watson Studio in Cloud Pak for Data and watsonx". Below the card, a description reads: "Build and deploy machine learning models on either platform. Work with foundation models on watsonx as a Service." At the bottom, there's a "Launch in" button with a dropdown menu. The menu shows two options: "IBM Cloud Pak for Data" and "IBM watsonx", with "IBM watsonx" highlighted by a blue oval.



3. From the main menu in the top left corner select **Projects -> View All Projects**.



4. Click on the **New Project** button. Select **Empty Project** and add your initials to the project name. For example, *LLM-workshop-YI*.



### New project

Define details

Name  
LLM-Workshop-YI

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

5. Switch to the **Services and Integrations** tab, then click **Associate Service**.

Projects / LLM-Workshop-YI

Overview Assets Jobs Manage

Project General Access control Environments Resource usage Services & integrations

Services & integrations

IBM services (1) Third-party integrations

Find services Name Service type

Associate service +



6. Select the displayed **Machine Learning** service and click **Associate**.

*Note: if the service is already associated, click **Cancel** and proceed to the next step.*

### Associate service

Choose an existing or add a new service to associate with your project.

Name	Type	Plan	Location	Status	Group
Machine Learning-fz	Watson Machine Learning	v2 Standard	Dallas	Not associated	default

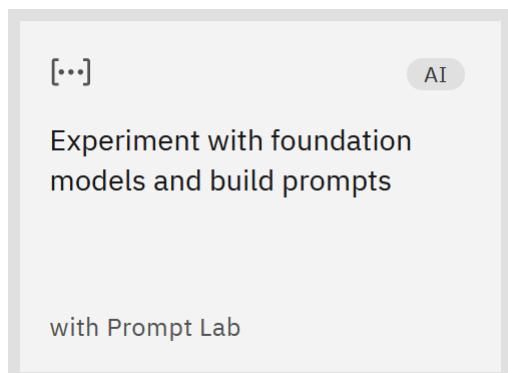
Cancel      Associate

7. Switch to the **Assets** tab, then click the **New asset** button.

All assets
All assets

Import assets      New asset

8. Click on the **Experiment with foundation models...** tile.



Before we start experimenting with prompting, let's review some key concepts about **model selection**.

As you try different models, you will notice that some models return better results with zero-shot prompting (providing instructions without examples) than others. Usually, models that have gone through *fine-tuning*, *instruction-tuning*, and *RLHF* generate significantly better output.

- *Fine-tuning* means that the original LLM was trained on high-quality labeled data for specific use cases. For example, if our goal for the model was to "Act as an IT architect" when generating output, during the fine-tuning process we provided labeled data examples of a writing style for IT architecture.
- If the model goes through the instruction-tuning process, then it will be able to generate output without explicit instructions, such as "*Can you answer this question?*" The model will understand that you are asking a question from the context and sentence structure.
- *RLHF (Reinforcement Learning from Human Feedback)* is a technique that is used to improve model output based on feedback provided by testers, usually domain experts (for example, lawyers for tasks related to generation of legal documents). Before the model is released, it is updated based on testing results.

While all vendors can say that their model has been fine-tuned, instruction-tuned, and has gone through RLHF, the industry benchmarks for LLMs are not mature. Even benchmarks that may eventually become the industry standard (for example, *TruthfulQA*) test only certain aspects of model output.

A potential solution to this issue is the research done by the broader LLM community. The LLM community is very active; information about the quality of models usually becomes widely known through various community resources such as articles, blogs, and YouTube videos. This is especially true for open source vs. proprietary models. For example, search for "llama vs. ChatGPT" and review results. You can also review the leaderboard on Hugging Face, keeping in mind that you will need to understand the evaluation criteria used for the leaderboard (see the About page of the leaderboard).

At this time, *llama-70b-2-chat* model, is one of the best models for zero-shot prompting. While it may seem like an obvious choice to always use *llama-70b-2-chat* in watsonx.ai, it may not be possible for several reasons:

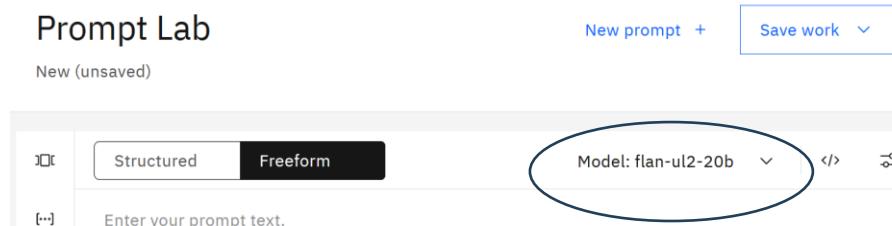
- Model availability in the data center (due to resources or licensing)
- Inference cost
- Hosting cost (for on-premises or hybrid cloud deployments).

It may be possible to achieve similar results with other models or with smaller versions of *llama* by using few-shot prompting or fine-tuning, which is why it is important to experiment with multiple models and understand prompt tuning and model tuning techniques.

*Note: Instructions in this lab are written for the *llama*, *granite-13b-chat-v2*, *flan*, and *mixtral* models, which are available in all IBM Cloud data centers where *watsonx.ai* is hosted as well as in the on-premises version of *watsonx.ai*. We encourage you to try other models (for example, *codellama*), if they are available in your workshop environment.*

9. After the **Prompt Lab** UI opens, switch to the **Freeform** tab.

Select the *flan-ul2-20b* model. We will review model settings later in the lab.

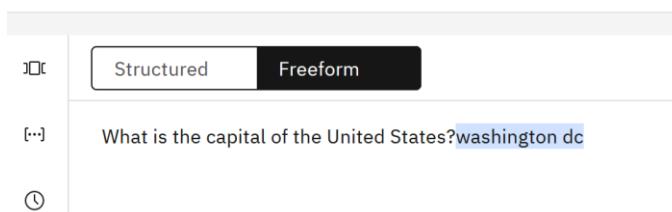


Since most LLMs, including the selected *flan* model, were trained on publicly available data, we can ask it some general questions.

10. Type in the question: "What is the capital of the United States?" and click **Generate**. The generated answer is highlighted in blue.

### Prompt Lab

New (unsaved)



We got the answer to our question without instructing the model to do it because the *flan* model was *instruction-tuned* to answer questions. Google, the creator of this model, published the training instructions that were used for the model in this [git repository](#). As you can see in documentation, the instructions are often shown in a "technical format", but they are still helpful for understanding the best prompting options for this model.



Scroll down to the *natural questions* section of the git page. Here we can see the various phrases we can use with the model when asking questions.

```
"natural_questions": [
    ("Question: {question}?\\nAnswer:", "{answer}"),
    ("{question}", "{answer}"),
    ("Answer the following question:\\n\\n{question}", "{answer}"),
    ("Answer this question:\\n\\n{question}?", "{answer}"),
    ("Please answer this question: {question}", "{answer}"),
    ("Answer the question...{question}?", "{answer}"),
    ("What is the answer to this question? {question}", "{answer}"),
    ("Can you tell me the answer to {question}?", "{answer}"),
    ("Next question: {question}", "{answer}"),
    ("Q: {question} A:", "{answer}"),
],
],
```

Next, we will ask a different question: "When was Washington, DC founded?"

## Prompt Lab

New prompt +

Save work ▾

New (unsaved)



Structured

Freeform

Model: flan-ul2-20b ▾

</>



[...] When was Washington, DC founded?1791

Double check if this is a correct answer by doing a traditional Internet search. You will find out that the correct answer is *July 16<sup>th</sup>, 1790*.

Next, switch the model to *granite-13b-chat-v2*, then *llama2-70b-chat* and ask the question again. This time we get the correct answer.

Chat

Structured

Freeform

AI

Model: granite-13b-chat-v2 ▾

{#} </>

When was Washington, DC founded?

Washington, DC was founded on July 16, 1790. It was established as the capital of the United States, replacing Philadelphia. The district is named after George Washington, the first President of the United States and one of the Founding Fathers. The creation of Washington, DC was part of the Residence Act of 1790, which also established the territory and individual land lot sales of the District of Columbia. The district is unique as it is not a state and has a federal government that is separate from the states.

Chat

Structured

Freeform

AI

Model: llama2-70b-chat ▾

{#} </>

When was Washington, DC founded?

Washington, DC was founded in 1790. The capital of the United States was previously located in Philadelphia, Pennsylvania, but a new capital city was desired to be built along the Potomac River. The Residence Act of 1790 authorized the construction of a new federal capital along the Potomac River, and President George Washington appointed a commission to oversee the project. The city was designed by French engineer Pierre Charles L'Enfant, who envisioned a grand city with wide avenues, public parks, and monumental buildings. The city was named after George Washington, and the cornerstone of the Capitol Building was laid in 1793. The city has since grown and developed into the vibrant and historic city it is today.

In general, this was not a simple question because at the time of writing conflicting dates are listed on the *Library of Congress* website and *Wikipedia*. *Library of Congress* is a more



credible source, and in this example the *Wikipedia* page, which may have been used for model training of the *flan* model, has the incorrect date.

We provided this example to highlight the fact that the primary usage of LLMs should not be *general knowledge question and answer*. The quality of LLM output depends on the knowledge base that it was trained on. If we asked another question, it's possible that *flan* would outperform other models.

***We should think of LLMs as an "engine" that can work with unstructured data rather than a "knowledge base".***

When you first start working with LLMs, you may think that some models are not returning the correct response because of the prompt format. Let's test this theory with the *flan* model.

Enter this prompt in the **Prompt Lab**:

*Answer the question provided in "".*

*Question: ""When was Washington, DC founded?""*

*Answer:*

Let's review why we constructed the prompt in this format:

- Triple single quotes ('') are often used to identify a question or text that we want the LLM to use. You can choose other characters, but avoid " (double quotes) because they may already be in the provided text
- Notice that we provided the word "Answer:" at the end. Remember that LLMs *generate the next probable word*, and providing the word "Answer" is a "hint" for the model.

Structured      Freeform

Answer the question provided in "".

Question: ""When was Washington, DC founded?""

Answer:

1791

Unfortunately, we did not get a more accurate result from the *flan* model.

We will try one more approach, this time with a different prompt, which you can copy from *Washington\_DC\_prompts.txt* in the *lab repo/Prompts* folder.

11. Copy *Prompt 1 (flan)* from the .txt file and paste it to the **Prompt Lab**.

## Prompt Lab

New prompt +

Save work

New (unsaved)

The screenshot shows the Watsonx.ai Prompt Lab interface. At the top, there are tabs for 'Structured' and 'Freeform', with 'Structured' selected. To the right, it says 'Model: flan-ul2-20b'. Below the tabs, there's a prompt template:

- [...] Using the following paragraph, answer the question provided in "".
- (① Paragraph:  
Washington, D.C. was founded on July 16, 1790. Washington, D.C. is a unique and historical place among American cities because it was completely planned for the national capital and needed to be distinct from the states. President George Washington chose the specific site (e) along the Potomac and Anacostia Rivers.  
The Serial Set contains a lot of information about the plan for the city of Washington, including maps, bills, and illustrations. In this Story Map, you can see original plans and pictures of Washington, D.C.)
- Question: ""When was Washington, DC founded?""
- Answer:July 16, 1790

While this example may seem simple because we provide the answer in our prompt, it demonstrates one of the key use cases for LLMs, which is called *Retrieval Augmented Generation (RAG)*. With RAG, we ask LLMs to answer questions or generate content based on the information that we provide. In our example we hardcoded content in the prompt, but it is also possible to implement RAG with automatic information retrieval from various knowledge bases, such as Websites, documents, emails, etc. In this case, the main feature of an LLM that we are interested in is “understanding” and not “knowledge”.

We used a simple example to ask a trivia question, but think about topics that may be relevant to your business for which “general information” may also exist, for example:

- What are the steps to get a driver’s license?
- What are the steps for submitting a car insurance claim?
- How can I close a credit card?
- How can I improve my credit score?
- Will an airline reimburse me for a canceled flight?

Most LLMs will be able to answer these questions because they were trained on data available on the Internet, but if you want the correct answer to your question, in most cases you will need to use *RAG*, i.e. provide information from your company’s data sources.

Watsonx.ai supports several implementations of RAG. We will cover it in more detail in one of the other labs.

Next, we’ll test prompts that generate output.

12. Click on the model parameters icon in the top right corner of the **Prompt Lab**.

If you would like to learn more about each input in the Model parameters panel, you can review [documentation](#).

Change the **Max tokens** to **500**. When LLMs process instructions and generate output, they convert words to tokens (a sequence of characters). While there isn’t a static ratio



for letter to token conversion, we can use  $10 \text{ words} = 15 \text{ to } 20 \text{ tokens}$  as a rule of thumb for conversion.

Use the *flan* model for your first test.

Model: flan-t5-xxl-11b

Model parameters

Stopping criteria ⓘ

Stop sequences

Min tokens: 0      Max tokens: 500

Enter up to 6 sequences to stop output after the minimum number of tokens is reached.

13. Change the prompt to write a paragraph:

*Write a paragraph about the capital of the United States.*

*Paragraph:*

Notice that our output is rather brief.

Structured      Freeform

Write a paragraph about the capital of the United States.

Paragraph: Washington, D.C. is the capital of the United States of America.

Next, we will try different model parameters and models to see if we can get better results.

14. In the model settings switch **Decoding** from *Greedy* to *Sampling*. Sampling will produce greater variability/creativity in generated content (see [documentation](#) for more information).

*Important: Make sure to delete the generated text after the word "Paragraph": before clicking **Generate** again because the model will continue generating after any given text in a prompt, which may result in repetition.*

Click the **Generate** button. It looks like we're not getting better results with this model, so let's try another one.

Structured    Freeform

Write a paragraph about the capital of the United States.

Paragraph: Washington, D.C. is the capital of the United States of America.

15. Test the same prompt with the *granite-13b-chat-v2* model. Delete the generated text and test again.

In our testing, we get better results with this model. Notice that every time you click **Generate** (after deleting the generated text), you get different results. We're seeing this because we set the *Decoding* option to *Sampling*. You can also try *Sampling* with different temperature (higher value will result in more variability).

Chat    Structured    Freeform

AI    Model: granite-13b-chat-v2    {#}    </>    ⚙

Write a paragraph about the capital of the United States.

Paragraph:  
The capital of the United States is Washington, D.C., a city located in the District of Columbia. Established in 1790, the city was named after George Washington, the first President of the United States. As the nation's capital, Washington, D.C., is home to numerous important institutions, including the White House, the U.S. Congress, and the Supreme Court. The city is also known for its rich cultural, educational, and political heritage, making it a significant tourist destination. With a population of over 700,000, Washington, D.C., is a vibrant and diverse city that offers a wide range of amenities and opportunities for its residents and visitors alike.

In addition to being the political heart of the United States, Washington, D.C., is also a hub for scientific and technological innovation. The city is home to numerous research institutions, such as the National Institutes of Health, the Food and Drug Administration, and the National Aeronautics and Space Administration (NASA's) Goddard Space Flight Center. These institutions contribute significantly to the city's economy and help to maintain its status as a leader in various fields, including medicine, healthcare, and space exploration.

Furthermore, Washington, D.C., is home to a diverse range of neighborhoods, each with its unique character and charm. From the trendy and artistic Shaw neighborhood to the historic and picturesque Georgetown, there is something for everyone in this city. The city's extensive public transportation system, including the Metro subway and bus system, makes it easy for residents and visitors to explore the different neighborhoods and attractions that Washington, D.C., has to offer.

Overall, Washington, D.C., is a city that represents the best of the United States, blending history, culture, politics, and innovation in a way that is both unique and inspiring. Whether you are a history buff, a culture vulture, or a science enthusiast, Washington, D.C., has something for you.

While it may seem unusual that the model generates a different output each time, it's what we are instructing the model to do by both giving it instructions ("write") and setting model parameters ("sampling"). We would not use the same instructions/parameters for a classification use case which needs to provide consistent output (for example, *positive*, *negative*, or *neutral* sentiment).

16. Finally, try the *llama* model. For "creative output" use cases, *llama* usually produces production-quality output.

While you can use the same prompt, we should also be familiar with *system prompt format* in *llama*. *llama* is one of the few models that accepts prompts in a specific format.



Enter this prompt (Prompt 2), which you can copy from the from *Washington\_DC\_prompts.txt* in the *lab repo /Prompts* folder.

```
<s>[INST] <<SYS>>  
You are a motivational speaker.  
<</SYS>>
```

*Please write a paragraph to motivate a tourist to visit Washington, DC [/INST]*

*Important note: If you decide to use this format for *llama*, make sure not to put anything after the *[/INST]* In this format you do not need to give the model a "hint" (Answer: to generate the answer, etc.), like you've done in other examples.*

```
<s>[INST] <<SYS>>  
You are a motivational speaker.  
<</SYS>>
```

Please write a paragraph to motivate a tourist to visit Washington, DC [/INST]

Absolutely! Here's a motivational paragraph to encourage a tourist to visit Washington, DC:

"Imagine a place where history comes alive, where the past, present, and future converge, and where the boundaries of possibility are limitless. That place is Washington, DC - the capital of the United States and a city that embodies the spirit of American ingenuity and progress. From the iconic monuments that dot the skyline to the world-class museums that line the National Mall, Washington, DC is a city that will leave you in awe. Walk in the footsteps of presidents, visionaries, and change-makers who have shaped the course of human history. Be inspired by the stories of courage, perseverance, and innovation that have made America the land of opportunity. And let the grandeur of the city's architecture, the beauty of its parks, and the warmth of its people leave you feeling energized and inspired. Whether you're a history buff, a culture lover, or an adventure-seeker, Washington, DC has something for everyone. So come and experience it for yourself. Let the power of this great city ignite your passion, fuel your imagination, and create memories that will last a lifetime."

As you can see, using the system prompt in *llama* is simple – it requires us to encapsulate the entire prompt in *[INST]*/*[/INST]* and add *[SYS]* if we want to give "general behavior instructions" to the model. The creators of *llama* recommend that this format is used for the best results, but it is not required.

Like the first prompting exercise, we started with LLMs' general knowledge for generating output. In a business use case scenario, we would give LLMs a few short bullet points and ask it to generate output.

17. If you wish to continue with the Washington, DC example, you can try Prompts 3 and 4 from the sample prompt file.

Notice that while IBM's granite model does not require a specific format for a system prompt (for example, *Act as a Marketing Consultant*), you can still specify it in the free text.

Since IBM developed the granite model, the detailed instructions for prompting best practices are available in [IBM documentation](#). If you're working with an open source model, prompting best practices are often documented by the larger data science community and, in some cases, by the vendor who developed the model.

Chat | Structured | **Freeform**



Model: granite-13b-chat-v2

{#}

&lt;/&gt;



You are Granite Chat, an AI language model developed by IBM. You are a cautious assistant. You carefully follow instructions. You are helpful and harmless and you follow ethical guidelines and promote positive behavior. Please do not say anything else and do not start a conversation.

**Act as a marketing consultant.**

Please generate a promotional email to visit the following attractions in Washington, DC:

1. The National Mall
2. The Smithsonian Museums
3. The White House
4. The U.S. Capitol
5. The National Gallery of Art

Email:**Subject: Unforgettable Experiences Await at Washington, D.C.'s Iconic Attractions!**

Dear [Recipient's Name],

Greetings from Washington, D.C., the nation's vibrant capital! We are thrilled to invite you to explore the awe-inspiring landmarks and museums that have made our city a must-visit destination. With so much to see and do, we've curated a list of must-see attractions that promise to captivate and inspire.

Some LLMs can follow instructions and generate output in multiple languages. Most LLM providers will document supported languages in the model card on Hugging Face or on the company's website.

18. In the **Prompt Lab**, click on the model dropdown and find the *granite-20b-multilingual* model.

### Select a foundation model

Select a model that best fits your needs. All models support English text. Check the model information for other supported languages.

Y	Q	Search for a model or task
granite-13b-instruct-v2	granite-20b-multilingual	llama-2-13b-chat
The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for generative...	The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for generative...	Llama-2-13b-chat is an auto-regressive language model that uses an optimized transformer architecture.
Provider: IBM	Provider: IBM	Provider: Meta
Type: Provided model	Type: Provided model	Type: Provided model

Review the model card and notice that the supported languages are listed in the model documentation.



## Granite 20 Billion Multilingual Model ( granite-20b-multilingual ) Details

Model Version (1.0.0): Released 3/15/2024

IBM Generative AI Large Language Foundation Models are Enterprise-level language models trained with large a volume of data. The Granite 20 Billion Multilingual ( granite-20b-multilingual ) model was trained using over 2.6 trillion tokens and further fine-tuned on multilingual common crawl data resulting in a model that works with English, German, Spanish, French, and Portuguese.

The table below lists the [xMMLU](#) and [xMT-Bench](#) benchmarks used to show the performance in 5 languages.

Benchmark	Average	English	German	Spanish	French	Portuguese
xMMLU	38.41	40.58	37.91	38.04	37.58	37.95
xMT-Bench	5.34	5.59	5.18	5.17	5.19	5.58

19. Select this model and make a slight change to our last prompt – ask the model to generate the promotional email in French.

You are Granite Chat, an AI language model developed by IBM. You are a cautious assistant. You carefully follow instructions. You are helpful and harmless and you follow ethical guidelines and promote positive behavior. Please do not say anything else and do not start a conversation.

Act as a marketing consultant.

Please generate a promotional email to visit the following attractions in Washington, DC [in French](#).

1. The National Mall
2. The Smithsonian Museums
3. The White House
4. The U.S. Capitol
5. The National Gallery of Art

Email:

Subject: Les meilleures attractions de Washington, DC à visiter absolument!

Les meilleures attractions de Washington, DC à visiter absolument!

Washington, DC est une ville riche en histoire et en culture, avec de nombreuses attractions touristiques fascinantes à explorer. Voici quelques-uns des meilleurs endroits à visiter lorsque vous êtes dans la capitale américaine.

Le National Mall est un grand espace ouvert au centre de Washington, DC, qui abrite plusieurs monuments et musées importants, notamment le Lincoln Memorial, le Martin Luther King Jr. Memorial et le World War II Memorial. C'est un lieu idéal pour se promener, admirer les monuments et prendre des photos.

Les musées du Smithsonian sont une autre attraction populaire de Washington, DC. Il y a plus de 19 musées sur place, chacun offrant une expérience unique et intéressante. Parmi les favoris des visiteurs figurent le National Air and Space Museum, le National Museum of Natural History et le National Museum of American History.

La Maison Blanche est le siège du président des États-Unis et peut être visitée pendant les heures d'ouverture publiques. Il est possible d'organiser des visites guidées pour voir les pièces et les bureaux du président, ainsi que les jardins extérieurs.

Another popular multilingual model is the open source model [mixtral-8x7b-instruct-v01-q](#).

Provider: Mistral AI, tuned by IBM

Summarization   Retrieval-Augmented Generation   Classification   Generation   Code generation and conversion   Extraction

This model is made with AutoGPTQ, which mainly leverages the quantization technique to "compress" the model weights from FP16 to 4-bit INT and perform inference. The GPU memory, and the data transferring between GPU memory and GPU compute engine, compared to the original FP16 model, is greatly reduced.

When you look at the Model card for this model, supported languages are not listed. In [watsonx](#) we publish model information provided by the model provider to Hugging Face without adding any additional information. In some cases, we may need to do additional

research to find out if a model supports additional languages. From the Mixtral [website](#) we learn that the open source model supports several languages.

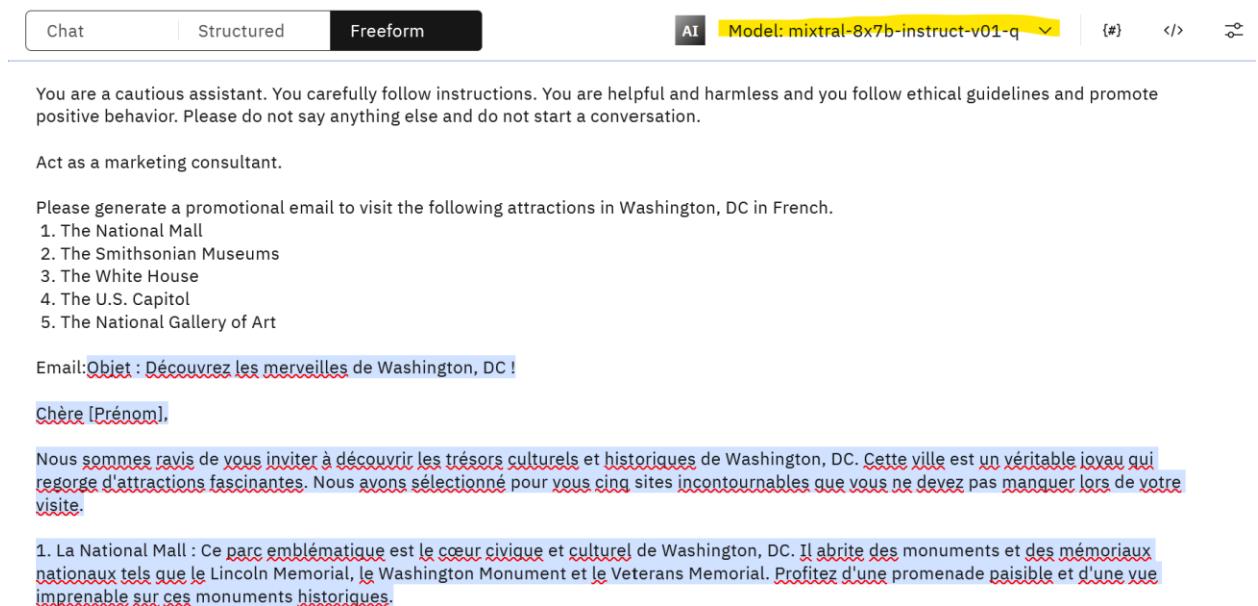
## Mixtral 8x7B

Currently the best open model. A 7B sparse Mixture-of-Experts (SMoE). Uses 12B active parameters out of 45B total.

- Fluent in English, French, Italian, German, Spanish, and strong in code
- 32k context window

If you wish, try the previous prompt with the *mixtral* model.

*Note: While it's a good idea to remove the IBM-related system prompt from the prompt, it does not impact mixtral output.*



The screenshot shows the Watson Assistant interface with the following details:

- Header buttons: Chat, Structured, Freeform (selected).
- Header status: AI Model: mixtral-8x7b-instruct-v01-q {#} </> ⌂
- Message input field: You are a cautious assistant. You carefully follow instructions. You are helpful and harmless and you follow ethical guidelines and promote positive behavior. Please do not say anything else and do not start a conversation.
- Message content:
  - Act as a marketing consultant.
  - Please generate a promotional email to visit the following attractions in Washington, DC in French.
  - 1. The National Mall  
2. The Smithsonian Museums  
3. The White House  
4. The U.S. Capitol  
5. The National Gallery of Art
- Email subject: Email:Objet : Découvrez les merveilles de Washington, DC !
- Email body:

Chère [Prénom],

Nous sommes ravis de vous inviter à découvrir les trésors culturels et historiques de Washington, DC. Cette ville est un véritable joyau qui regorge d'attractions fascinantes. Nous avons sélectionné pour vous cinq sites incontournables que vous ne devez pas manquer lors de votre visite.

1. La National Mall : Ce parc emblématique est le cœur civique et culturel de Washington, DC. Il abrite des monuments et des mémoriaux nationaux tels que le Lincoln Memorial, le Washington Monument et le Veterans Memorial. Profitez d'une promenade paisible et d'une vue imprenable sur ces monuments historiques.

Next, we will look at the *Sample Prompts* included with the **Prompt Lab**. These samples implement some of the most common use cases in Generative AI.

20. Click on the *Sample Prompts* icon and scroll down to the *Marketing email generation* example.

## Prompt Lab

New (unsaved)

The screenshot shows the 'Sample prompts' section of the Prompt Lab. It includes three examples:

- Sentiment classification**: Classify reviews as positive or negative.
- Marketing email generation**: Generate email for marketing campaign.
- Thank you note generation**: Generate thank you note for workshop attendees.

On the right, there are tabs for 'Structured' and 'Freeform'. Below the examples, there are sections for 'Set up', 'Instruction (optional)', 'Generate a 5 sentence marketing m', 'Examples (optional)', and 'Details'.

You can run this example from either the *Structured* or *Freeform* view. If you use the *Structured* view, the input for generation is shown under *Details*.

Try ^

Test your prompt ⓘ

### Details

Characteristics:

Company - Golden Bank...

If you wish, change the input for the generated email and test the prompt.

Structured Freeform

Generate a 5 sentence marketing message for a company with the given characteristics.

Details

Characteristics:

Company - Golden Bank

Offer includes - no fees, 2% interest rate, no minimum balance

Tone - informative

Response requested - click the link

End date - July 15

Email

You may have already noticed that working with LLMs requires experimentation. In the **Prompt Lab** we can save the results of our experimentation with prompts

- As a *notebook*
- As a *prompt*
- As a *prompt session*.

If we save our experimentation as a prompt session, we will be able to access various prompts and the output that was generated.

In the **Prompt Lab**, select **Save work -> Save as -> Prompt session**.

The screenshot shows the IBM Watson Prompt Lab interface. At the top right, there is a dropdown menu with options: "New prompt +", "Save work ^", and "Save as". The "Save work ^" option is highlighted with a blue border. Below this, the main workspace displays a prompt: "Write a paragraph about the capital of the United States." To the right of the prompt are "Model parameters" (set to 1), "Stopping criteria" (empty), and "Min tokens" (0) / "Max tokens" (500). At the bottom left, status information is shown: "Stop reason: End of sequence token encountered", "Tokens: 17 input + 140 generated = 157 out of 2048 | Seed: 2263369679", and "Time: 5.3 seconds". A large blue "Generate" button is at the bottom center.



Name the prompt session *Generate\_paragraph*.

Save your work

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

Asset type		Define details
Prompt	<input checked="" type="checkbox"/>	Name Generate_paragraph
Prompt session	<input checked="" type="checkbox"/>	Description (optional) What's the purpose of this prompt
Notebook	<input type="checkbox"/>	

Save the current prompt only, without its history.  
Save history and data from the current session.  
Save the current prompt as a notebook.

View in project after saving (i)

21. Open **watsonx.ai** in another browser window and navigate to your project.
22. Click on the **Assets** tab and open the prompt session asset that you created.

Projects / LLM-Workshop-YI

Overview    **Assets**    Jobs    Manage

Find assets

1 assets

All assets

All assets

Name

[...] Generate\_paragraph (i)  
Prompt session

Asset types

> [..] Prompts

23. In the **Prompt Lab**, click on the *History* icon.

Generate\_paragraph

Prompt session

Structured    Freeform

Set up ^

(i) Instruction (optional) (i)

Notice that you can click on various prompts that you tested and view the output in the **Test your prompt** section of the **Prompt Lab**.

## Generate\_paragraph

Prompt session

New prompt +

Save work ▾

The screenshot shows a "History" panel on the left listing several prompts and their outputs. The most recent entry is highlighted:

- 11:41 AM**: ibm/mpt-7b-instruct2  
Write a paragraph about t...  
...6 and stands 555 feet tall
- 11:40 AM**: ibm/mpt-7b-instruct2  
Write a paragraph about t...  
...b is an online service).c/p
- 11:38 AM**: ibm/mpt-7b-instruct2  
Write a paragraph about t...  
.the Supreme Court
- 11:35 AM**: ibm/mpt-7b-instruct2  
Write a paragraph about t...  
...the capitals of states also
- 11:34 AM**: ibm/mpt-7b-instruct2  
Write a paragraph about t...  
...ed States Supreme Court
- 11:29 AM**: google/llan-ul2  
Write a paragraph about t...  
... United States of America
- 11:25 AM**: ibm/mpt-7b-instruct2  
Write a paragraph about t...  
...dings and great museums

The right side of the interface shows the "Try" section with the following details:

**Test your prompt**

**Input:** Write a paragraph about the capital of the United States.

**Paragraph:**

**Output:** Washington, D.C. is the capital of the United States. The city was originally founded in 1791, but today there are more than 700,000 people living in the city. The city is sometimes called "the District of Columbia." Although many people live in Washington, it's actually not a city, because it's part of the District of Columbia. Many important government buildings are in Washington, and these include the White House, the Capitol Building, and the Supreme Court.

**Viewing: July 29, 2023 at 11:38:17 AM**

Restoring this version will overwrite

Close this browser tab and return to the **Prompt Lab**.

What you have tried so far is a “*question and answer and generation use case with zero-shot prompting*” – you’ve asked the LLM to generate output without providing any examples. The majority of LLMs produce better output when they’re given a few examples. This technique is called “*few-shot prompting*”. The examples are provided in the prompt after the instruction.

Let’s test *few-shot prompting* for various use cases.

24. In the **Prompt Lab**, create a new prompt. Switch to **Freeform**, and paste the following prompt (this prompt is available in the *Paragraph\_few\_shot.txt* file in the */Prompts* folder).

### Prompt Lab

New (unsaved)

The "Prompt Lab" interface shows a "Freeform" prompt template:

**Structured**   **Freeform**

Write a paragraph about the capital in "".

Capital: ""London""

Paragraph: London, the iconic capital city of the United Kingdom, stands as a dynamic tapestry woven from centuries of history, culture, and innovation. It captures the essence of a global metropolis. The River Thames meanders through its heart, bordered by a panorama of architectural wonders such as The city's rich history is palpable in its cobbled streets, where ancient stories whisper from every corner. Museums like the British Museum and the Royal Albert Hall stage world-class performances that define the realm of entertainment. From the royal grandeur of Buckingham Palace to the bustling mosaic of experiences that celebrate both tradition and innovation. A melting pot of cultures and cuisines, London's culinary scene is a reflection of its alleyway, park, and bustling street, London emanates an aura of ceaseless energy and opportunity, inviting visitors and residents alike to immerse themselves in its vibrant atmosphere.

Capital: ""Tokyo""

Paragraph: Tokyo, the electrifying capital of Japan, stands as a testament to the harmonious blend of ancient traditions and cutting-edge modernity. Tucked between the past and the future, Skyscrapers and neon lights adorn the skyline, creating a mesmerizing spectacle in districts like Shinjuku and Shibuya. And Senso-ji offer serene respite, where one can glimpse into Japan's rich spiritual heritage. The efficient and intricate public transportation system, neighborhoods, each with its unique character. From the fashion-forward streets of Harajuku to the upscale elegance of Ginza, Tokyo's districts cater to renowned sushi, ramen, and street food stalls enticing the palate. The city's constant evolution is matched only by its unwavering commitment to preserving tradition and innovation in harmony.

## 25. Modify model parameters:

- Change *decoding* to *sampling* (for more creative output)
- Change the min and max number of tokens to the output you would like to see (for example, *50 min* and *500 max*)
- If you wish, you can test different models.

Test the model and review the output.

The screenshot shows the WatsonX Challenge Sandbox Prompt Lab interface. At the top, there are tabs for 'Projects' and 'WatsonX Challenge Sandbox'. On the right, there are buttons for 'AI guardrails on' and 'Save work'. Below this, the title 'Prompt Lab' is displayed, followed by 'New (unsaved)'. The main area has two tabs: 'Structured' (selected) and 'Freeform'. Under 'Structured', there are two paragraphs of text. The first paragraph is about Cairo, and the second is about Washington, DC. To the right of the text is a 'Model parameters' panel. It includes sections for 'Decoding' (set to 'Sampling'), 'Temperature' (set to 0.7), 'Top P (nucleus sampling)' (set to 1), 'Top K' (set to 50), and 'Random seed'. At the bottom of the panel, there is a note: 'Model: flan-ul2-20b' and a 'Generate' button.

Next, we will review the concept of *tokens*.

## 26. Notice the token count that's shown on the bottom of the model output.

In the screenshot of this model output, the “*out of*” number (2048) shows the maximum number of tokens that can be processed by a model. If you test with a different model, the maximum number of tokens will be different.

Paragraph:

Washington, DC, the capital of the United States, stands as a testament to the nation's rich history and enduring spirit. The city's diverse neighborhoods offer a kaleidoscope of unique character and charm. The city's culinary scene is world-renowned, with a variety of restaurants and food trucks. Washington is also a center of government and innovation, and a vibrant hub of activity.

Stop reason: End of sequence token encountered

Tokens: 812 input + 196 generated = 1008 out of 2048

Time: 6.3 seconds



It is important to understand the following facts about tokens:

- All LLMs have a limit for the number of supported tokens. The maximum number of tokens is usually captured in documentation or the UI, as you've seen in the **Prompt Lab**.
- The maximum number of tokens includes both input and output tokens. This means that you can't provide an unlimited number of examples in the prompt. In addition to that, each model has the maximum number of output tokens (see [documentation](#)).
- Some vendors have daily/monthly token limits for different plans, which should be considered when selecting an LLM platform.

Example of token limits (from [documentation](#)):

Table 1. IBM foundation models in watsonx.ai

Model name	IBM indemnification	Billing class	Maximum tokens Context (input + output)	Supported tasks	More information
<a href="#">granite-13b-chat-v2</a>	Yes	Class 1	8192	<ul style="list-style-type: none"><li>classification</li><li>extraction</li><li>generation</li><li>question answering</li><li>summarization</li></ul>	<ul style="list-style-type: none"><li><a href="#">Model card</a></li><li><a href="#">Website</a></li><li><a href="#">Research paper</a></li></ul>

Understanding token constraints is especially important for summarization, generation, and Q&A use cases because they may require more tokens than classification, extraction, or sentiment analysis use cases.

The token constraint limitation can be solved with several approaches. If we need to provide more examples to the model, we can use an approach called *Multitask Prompt Tuning (MPT)* or *fine-tuning*. We are not covering these advanced approaches in this introductory lab.

Up to this point, we reviewed question & answer and generation examples. We started with these examples because for most users they are the "first experience" with generative AI. Many people are familiar with *ChatGPT*, a popular personal assistant application developed by *OpenAI*. Sometimes the terms *generative AI* and even *LLM* are used interchangeably with *ChatGPT*, but *ChatGPT* is more than an LLM, it's a complex application that uses LLMs.

LLMs are building blocks or components of an application, and by themselves, they can rarely be used by a business user. *ChatGPT* is an application that focuses on personal productivity tasks for many types of users. Companies that want to build AI-driven applications need an AI development and deployment **platform**, such as *watsonx.ai*.

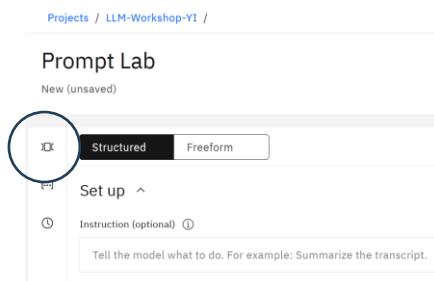
In our experience of working with clients, some of the top generative AI use cases are:

- *Content summarization*
- *Content classification*
- *Content generation*
- *Content extraction, NER (Named Entity Recognition)*
- *Sentiment analysis*
- *Question answering with RAG*

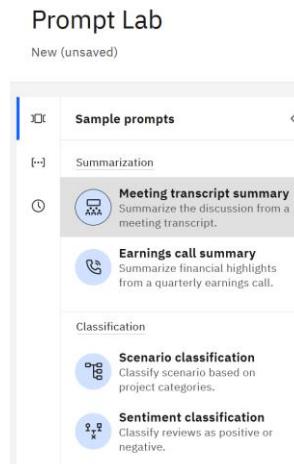


Let's review sample prompts for a few more use cases.

27. In the **Prompt Lab**, create a new prompt and click on the **Sample prompts** icon.



28. Explore/test the different prompts, which are organized by use cases.



Notice that some examples use few-shot prompting (*Meeting transcript summary*), while others (*Sentiment classification*) use zero-shot prompting.

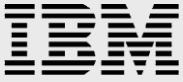
If you wish, try a few sample prompts that we provided in the downloaded *lab repo /Prompts* folder and sample prompts from [documentation](#).

When you were using the **Prompt Lab** you may have noticed the **Chat** tab. Let's review the capabilities of this feature.

By default, all LLMs are stateless – they do not have a “memory”, which is also called “state” in applications.

29. In the **Prompt Lab**, try executing these 2 prompts sequentially with the */lama* model:

Prompt 1: Please generate an SQL statement that joins tables CUSTOMER and TRANSACTIONS on column CUSTOMER\_ID



Chat

Structured

Freeform

AI

Model: llama-2-70b-chat

▼

Please generate an SQL statement that joins tables CUSTOMER and TRANSACTIONS on column CUSTOMER\_ID

SQL Statement:SELECT  
CUSTOMER.FIRST\_NAME,  
CUSTOMER.LAST\_NAME,  
TRANSACTIONS.TRANSACTION\_DATE,  
TRANSACTIONS.TOTAL\_AMOUNT  
FROM  
CUSTOMER  
INNER JOIN  
TRANSACTIONS ON CUSTOMER.CUSTOMER\_ID = TRANSACTIONS.CUSTOMER\_ID;  
``

**Prompt 2:** Please generate Python code to invoke this statement accessing Netezza database.

Chat

Structured

Freeform

AI

Model: llama-2-70b-chat

▼

Please generate Python code to invoke this statement accessing Netezza database.

Code: ``  
SELECT  
EXTRACT (YEAR FROM date\_key) || 'Q' || EXTRACT (QUARTER FROM date\_key) AS date\_year\_quarter,  
COUNT (DISTINCT case\_id) AS num\_cases  
FROM  
my\_table  
WHERE  
date\_key BETWEEN '2020-01-01' AND '2020-04-30'  
GROUP BY  
EXTRACT (YEAR FROM date\_key), EXTRACT (QUARTER FROM date\_key)  
ORDER BY  
date\_year\_quarter;

*Note: While the prompts can be improved, we are not going to work on it at this time because we are focusing on the issue with memory.*

You will not get meaningful results even from the best model because the model does not know what "this" (the preceding statement) means as it does not have a memory.

Now switch to the **Chat** mode in the Prompt Lab and try these two prompts again.



## Output of the first prompt:

EL

You: 05:33 PM

Please generate an SQL statement that joins tables CUSTOMER and TRANSACTIONS on column CUSTOMER\_ID

AI

05:33 PM

```
SELECT *
FROM CUSTOMER
JOIN TRANSACTIONS ON CUSTOMER.CUSTOMER_ID = TRANSACTIONS.CUSTOMER_ID
```

Explanation:

The SQL statement above joins the CUSTOMER and TRANSACTIONS tables on the CUSTOMER\_ID column. This is done using the JOIN keyword, followed by the name of the table to join, the ON keyword, and the condition for the join. In this case, the condition is that the CUSTOMER\_ID column in both tables must match. The \* wildcard is used to select all columns from both tables.



## Output of the second prompt:

EL

You: 05:33 PM

Please generate Python code to invoke this statement accessing Netezza database.

AI

05:34 PM

```
import pyodbc

# Connect to Netezza database
conn = pyodbc.connect('DRIVER={Netezza ODBC Driver};SERVER=my_netezza_server;PORT=5480;DATABASE=my_da'

# Create a cursor object
cursor = conn.cursor()

# Execute the SQL statement
cursor.execute('SELECT * FROM CUSTOMER JOIN TRANSACTIONS ON CUSTOMER.CUSTOMER_ID = TRANSACTIONS.CUSTOMER_ID')

# Print the results
for row in cursor.fetchall():
    print(row)

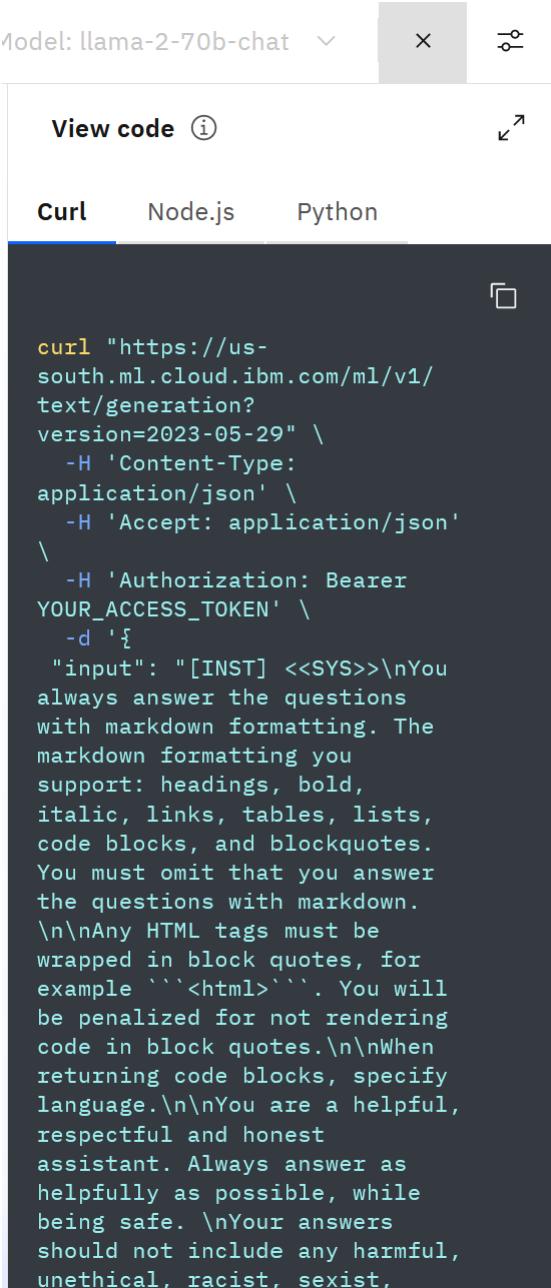
# Close the cursor and connection
cursor.close()
conn.close()
```

Notice that this time we got the correct Python code with the right SQL statement.

The **Chat UI** implements “memory” for LLM integrations. Memory implementation in applications that use LLMs is an intermediate complexity topic, and it is covered in more detail in other hands-on labs. On a high-level, memory in LLMs is implemented by

appending the response from the previous prompt to the next one. This approach increases the number of tokens sent to the LLM each time, and eventually (usually after 3-5 interactions), we will reach the max number of tokens supported by most chat LLMs. This is a known constraint of working with LLMs, and application developers can use various techniques to solve it.

If you wish, open the code panel of the **Chat UI** to review the prompt. The shown prompt (included in the REST call) shows the last executed prompt.



The screenshot shows the 'View code' panel of the IBM Watson AI Chat UI. The panel has a header with 'Model: llama-2-70b-chat' and a close button. Below the header are three tabs: 'Curl' (selected), 'Node.js', and 'Python'. The main area contains a 'Copy' icon and a 'curl' command. The command is a multi-line string starting with 'curl "https://us-south.ml.cloud.ibm.com/ml/v1/text/generation?version=2023-05-29" \'. It includes headers for Content-Type (application/json), Accept (application/json), and Authorization (Bearer YOUR\_ACCESS\_TOKEN). The body of the request is a JSON object with an 'input' key containing a template for generating text based on markdown input, supporting headings, bold, italic, links, tables, lists, code blocks, and blockquotes. The template also specifies that answers should be helpful, honest, and safe, avoiding harmful, unethical, racist, or sexist content.

```
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": "[INST] <><>\nYou always answer the questions with markdown formatting. The markdown formatting you support: headings, bold, italic, links, tables, lists, code blocks, and blockquotes. You must omit that you answer the questions with markdown.\n\nAny HTML tags must be wrapped in block quotes, for example ```<html>```. You will be penalized for not rendering code in block quotes.\n\nWhen returning code blocks, specify language.\n\nYou are a helpful, respectful and honest assistant. Always answer as helpfully as possible, while being safe. \nYour answers should not include any harmful, unethical, racist, sexist,
```

You can copy the prompt and paste it to a notepad to review the details.



The shown prompt is the last prompt that was executed. Notice that the first prompt and the output of the first prompt are included in the final prompt.

```
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": "[INST] <>SYS><\nYou always answer the questions with markdown formatting. The markdown formatting you support: headings, bold, italic, links, tables, lists, code blocks, and blockquotes. You must omit that you answer the questions with markdown.\n\nAny HTML tags must be wrapped in block quotes, for example ``<html>``. You will be penalized for not rendering code in block quotes.\n\nWhen returning code blocks, specify language.\n\nYou are a helpful, respectful and honest assistant. Always answer as helpfully as possible, while being safe. Your answers should not include any harmful, unethical, racist, sexist, toxic, dangerous, or illegal content. Please ensure that your responses are socially unbiased and positive in nature.\n\nIf a question does not make any sense, or is not factually coherent, explain why instead of answering something not correct. If you don't know the answer to a question, please don't share false information.\n<>/SYS>\n\n<\nPlease generate an SQL statement that joins tables CUSTOMER and TRANSACTIONS on column CUSTOMER_ID\n[INST]\n```\nSELECT *\nFROM CUSTOMER\nJOIN TRANSACTIONS\nON CUSTOMER.CUSTOMER_ID = TRANSACTIONS.CUSTOMER_ID\n```\n\nThis SQL statement joins the CUSTOMER and TRANSACTIONS tables on the CUSTOMER_ID column. The JOIN clause specifies the tables to be joined and the ON clause specifies the condition for the join, which is that the CUSTOMER_ID column in both tables must match. The SELECT clause then specifies that all columns from both tables should be included in the result set. [INST]\nPlease generate Python code to invoke this statement accessing Netezza database. [INST]\n```\nimport nzsql\n\n# Connect to Netezza database\nconn = nzsql.connect('localhost', 'myusername', 'mypassword', 'mydatabase')\n\n# Create a cursor object\nncursor = conn.cursor()\n\n# Define the SQL statement\nsql = """\nSELECT *\nFROM CUSTOMER\nJOIN TRANSACTIONS\nON CUSTOMER.CUSTOMER_ID = TRANSACTIONS.CUSTOMER_ID\n"""\n\n# Execute the SQL statement\nncursor.execute(sql)\n\n# Fetch the results\nresults = ncursor.fetchall()\n\n# Print the results\nfor row in results:\n    print(row)\n\n# Close the cursor and connection\nncursor.close()\nconn.close()\n```\n\nThis Python code uses the nzsql library to connect to a Netezza database, create a cursor object, define the SQL statement, execute the statement, fetch the results, and print the results. Finally, it closes the cursor and connection. [INST]",
    "parameters": {
        "decoding_method": "greedy",
        "max_new_tokens": 900,
        "min_new_tokens": 0,
        "stop_sequences": [],
        "repetition_penalty": 1
    },
    "model_id": "meta-llama/llama-2-70b-chat".
}
```

Another interesting part of this prompt are system instructions: to return output in markdown formatting vs. just the default text. When output is returned in this format, the integration with the UI (in this case the Chat UI interface) requires fewer lines of code.

Watsonx.ai does not provide a UI for end-user applications because UI is typically developed and hosted in other components of a solution. In the next section you will learn how to integrate LLMs with applications.

## Summary:

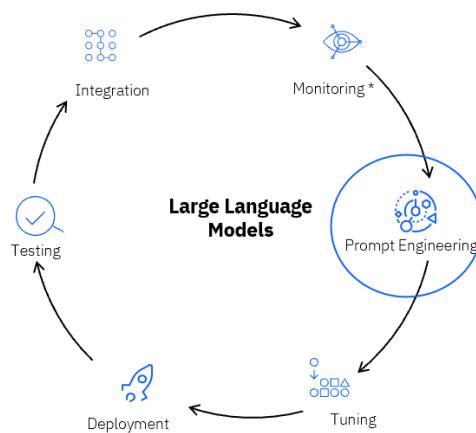
Throughout this lab you have used both open source and IBM models. Some of the most popular open source models are *llama*, *mixtral*, and *flan*. IBM will continue to bring the best open source models to **watsonx.ai**. At the same time, IBM developed several LLMs which can offer several benefits to our customers.

Despite the significant interest in application build on LLMs, concerns around fairness, bias, and regulations remain a hurdle for their enterprise-wide adoption. To address these issues, IBM developed *granite* series of foundation models. Trained on enterprise-relevant content, granite models meet rigorous data governance and regulatory criteria defined by the IBM AI Ethics code and Chief Privacy Office. These IBM models provide improved accuracy for targeted business domains like finance and use cases such as RAG, achieved through chat fine-tuning and model alignment techniques. Additionally, granite models are cost-effective, with less infrastructure requirements, IP indemnification, and an easy-to-use toolkit for model customization and application integration.

Now that you have reviewed and created prompts, we will test the integration of LLMs with client applications.

## Part 2: Integrate LLMs with applications

Up to this point you have completed a simple prompt engineering exercise. Prompt engineering is just one of the steps in the process of integrating LLMs into business applications.



Let's review other steps:

- There are several types of tuning, which are usually applied to improve model output. In many use cases tuning won't be a required step.
- LLMs are pre-deployed (available for invocation out-of-the-box) in `watsonx.ai`. The only time deployment may be required is for tuned models.
- Testing and integration is done with either the REST API or the Python SDK.

In this section, we will review the testing and integration steps.

1. Navigate to the **Prompt lab** and open one of the prompts you previously created or one of the sample prompts.

In this section, we will be using the loan summary prompt which you can find in `/Prompts/Loan_few_shot_summary.txt`.

2. Generate a response using this prompt.

- You can use either the `granite-13b-chat-v2` or the `llama` model (add the system prompt for llama)
- Keep the decoding method as `Greedy`
- Add a stop sequence of `."` to prevent output that ends with mid-sentence.
- Make sure to set the min and max tokens to 50 and 300.

Chat | Structured | Freeform | AI | Model: granite-13b-chat-v2 | {#} | </> | X

payment.

4. These discrepancies in the cash payout amount and the reduced mortgage payment relief prevented the reviewer from fulfilling their original financial objective of obtaining the \$25K cash.

5. The limitations and changes in the terms impacted the overall satisfaction with the refinancing process and compromised the financial benefits the reviewer had anticipated.

Review:""

I started my loan process toward securing a VA loan. I was waiting for a month and a couple weeks, then I was told that the VA needed to acquire my retirement points to verify my veteran status. If I knew this is what my loan was on hold for, I could have contacted the VA office right away and got this cleared up.

For whatever reason, it took the underwriting department a long time to verify my employment status, even after I uploaded a couple years of my W2 forms from both of my jobs, and they had my Social Security number to further verify my employment status. My loan completion date was extended, because I wasn't made aware that they were waiting for my VA status to be approved. The push back for my mortgage is common for mortgage companies, but this caused my interest rate to go up. Then, the securing of a closing lawyer being made aware to me and the lawyer needing three days to get their end prepared for me to go to their office to sign the paperwork wasn't made aware to me. My loan missed the second closing date. For whatever reason, the locked in interest rate jumped up 5/8 points. After making the banker I was working with aware of this, he didn't understand why the locked in interest rate jumped up either. He was nice enough to work on it and was able to get the interest rate down in 1/4 of a point, so my mortgage has an interest rate that is 3/8 of a point higher than my locked in interest rate in the beginning of this process. Although my interest rate is higher than the locked in interest point, at the end, the mortgage is successfully finished.

""

Top bullet points:1. The reviewer started the process for a VA loan and experienced delays due to the need to verify their veteran status and employment status.

2. The loan completion date was extended due to these verification issues, resulting in a higher interest rate.

**Model parameters**

Decoding

Greedy  Sampling [\(i\)](#)

Repetition penalty

1  2 1

Stopping criteria [\(i\)](#)

Stop sequences

.

Min tokens 50 Max tokens 300

Enter up to 6 sequences to stop output after the minimum number of tokens is reached.

After testing the prompt (click **Generate**), click on the **View code** icon.

AI | Model: granite-13b-chat-v2 | {#} |  | </>

Copy the code to a notepad.

[View code](#)[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": "Please provide top 5  
bullet points in the review  
provided in '|||||'.  
\n\n\nReview:\n'|||||I  
had 2 problems with my  
experience with my refinance.  
1) The appraisal company used  
only tried to lower my house  
value to fit the comps that he  
was able to find in the area.  
My house is unique and he did  
not use the unique pictures to  
compare value. He purposely  
left them out of the appraisal.
```

Let's review the code.

This code is an example of a REST call to invoke the model. **Watsonx.ai** also provides a Python API for model invocation, which we will review later in this lab.

The header of the REST request includes the URL where the model is hosted and a placeholder for the authentication token. At this time, all users share a single model inference endpoint. In the future, IBM plans to provide dedicated model endpoints.

*Note: IBM does not store model inference input/output data. In the future, users will be able to opt in to storing data.*

Security is managed by the *IBM Cloud authentication token*. We will get this token shortly.

The body of the request contains the entire prompt.

```
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 '{'
```

Header

```
"input": "Please provide top 5 bullet points in the review provided in \n\nReview:\n\nI had 2 problems with my experience with my refinance. 1) The appraisal company used only tried to lower my house value to fit the comps that he was able to find in the area. My house is unique and he did not use the unique pictures to compare value. He purposely left them out of the appraisal. 2) I started my loan process on a Thursday. On Saturday I tried to contact my loan officer to tell him of the American Express offer that I wanted to apply for. I was informed that it was too late and I could not use it because it would delay the process. I had just received the email about the offer and I had just started the process so how was it too late to get in on the $2,000 credit on my current bill. I let it go but I should have dropped the process and restarted it because that would have helped me out with my bill.\n\nTop bullet points:\n1. The appraisal company undervalued the reviewer's house by purposely excluding unique pictures that would have accurately assessed its value.\n2. The uniqueness of the house was not taken into consideration, and the appraiser relied solely on comps that did not reflect its true worth.\n3. The reviewer attempted to inform their loan officer about an American Express offer they wanted to apply for, which would have provided a $2,000 credit on their current bill.\n4. The loan officer stated it was too late to take advantage of the offer as it would delay the process,\ndespite the reviewer having just received the email and recently started the loan process.\n5. The reviewer regrets not
```

Finally, at the end of the request we specify model parameters and the *project id*.

```
"parameters": {
    "decoding_method": "greedy",
    "max_new_tokens": 300,
    "min_new_tokens": 50,
    "stop_sequences": [
        "."
    ],
    "repetition_penalty": 1
},
"model_id": "ibm/granite-13b-chat-v2",
"project_id": "2e0ffc21-9d10-4cde-9fce-f3e78f7503e3",
```

The *project id* can be looked up in the **Project -> Manage view** tab of the watsonx.ai project

Projects / LLM-Workshop-YI

Overview Assets Jobs Manage

General

Details

Name: LLM-Workshop-YI

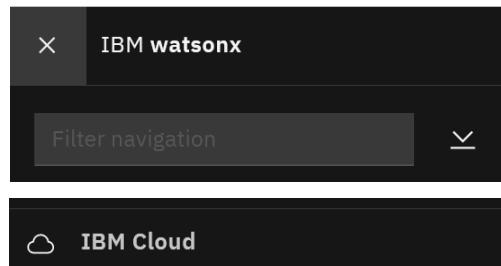
Description: What's the purpose of this project?

Project ID: e9207503-90c4-430b-a366-d6c24b5a020a

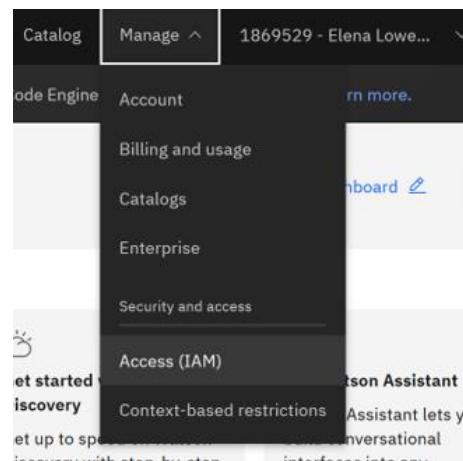
Now, we will get the authentication token.



3. Open a new browser window and from the main **watsonx menu** (top left corner), select **IBM Cloud**.



4. Select **Manage -> Access (IAM)**.



5. Click **API Keys -> Create**. Give the token a name and save it in a notepad. You will use it in the sample notebook.

The screenshot shows the 'API keys' section of the IBM Cloud IAM interface. On the left, there's a sidebar with options like 'Manage identities', 'Users', 'Trusted profiles', 'Service IDs', and 'API keys' (which is currently selected). The main area has a heading 'API keys' and a sub-heading explaining that API keys are associated with a user's identity. It includes a 'Create' button and a table showing existing API keys:

Status	Name	Description	Date Created
Active	API Key for Deployments		2021-12-13 15:23 GMT

6. In **watsonx.ai** click **New Asset -> Work with data and models in Python or R notebooks**.

Click the **From file** tab and navigate to the downloaded *lab repo/Notebooks* folder to select the *TestLLM* notebook. Make sure that the *Python 3.10 environment* is selected.



Click **Create** to import the notebook.

Work with data and models in Python or R notebooks

Define the details to create a notebook asset and open it in the Jupyter notebook editor tool.

Blank      **From file**      From URL

Name  
TestLLM

Description (optional)

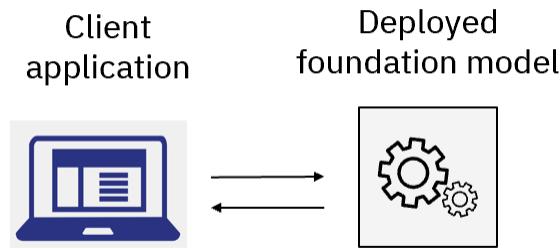
Select runtime  
Runtime 22.2 on Python 3.10 XXS (1 vCPU 4 GB RAM)

The selected runtime has 1 vCPU and 4 GB RAM.  
It consumes 0.5 capacity units per hour.  
[Learn more](#) about capacity unit hours and Watson Studio pricing plans.

Let's review the sample notebook.

This notebook acts as a *client application* that invokes the deployed LLM with a Python SDK. We are using the notebook as a client for simplicity of testing during this lab.

Enterprise client applications can be implemented in Python, Java, .Net and many other programming languages. As discussed earlier, LLMs deployed in **watsonx.ai** can be invoked either with REST calls or with the Python SDK.



Run the notebook to test the LLM with your prompts. See specific instructions in the notebook.

**Note: if you do not have a Python IDE, skip the rest of the lab, and watch this video in the lab repo/Reference to learn how to invoke LLMs from client applications. Proceed to the Conclusion section.**

Next, we will use a Python IDE, such as **Visual Studio** or **PyCharm** to run the client application.

30. Find the following Python scripts in the downloaded *lab repo/Scripts* folder:

- *demo\_wml\_api.py*
- *demo\_wml\_api\_with\_streamlit.py*

Load these scripts into your Python IDE.

31. In your Python IDE install the required libraries. We recommend that you use Python 3.10 environment but 3.11 can be also used.

```
pip install ibm-watsonx-ai
```



```
pip install ibm-cloud-sdk-core
```

In non-Anaconda Python environments you may need to install another package:

```
pip install python-dotenv
```

Let's review the scripts.

*demo\_wml\_api.py* is a simple Python script that shows how to invoke an LLM that is deployed in watsonx.ai. Code in this script can be converted to a module and used in applications that interact with LLMs.

The script has the following functions:

- *get\_credentials()*: reads the API key and the project id from the .env file (will be used for authentication)
- *get\_model()*: creates an LLM model object with the specified parameters
- *answer\_questions()*: invokes a model that answers simple questions
- *get\_list\_of\_complaints()*: generates a list of complaints from a hardcoded customer review
- *invoke\_with\_REST()*: shows how to invoke the LLM using the REST API (other functions use the SDK)
- *get\_auth\_token()*: generates the token that's required for REST invocation
- *demo\_LLM\_invocation()*: invokes all other functions for testing

Prior to running the script, create a .env file in the root directory of your project and add your Cloud API key, project id, cloud URL, and locale.

The template for the .env file is located in the lab repo/Reference folder (*env\_file\_template.txt*)

A screenshot of the PyCharm IDE interface. The top navigation bar includes 'Project', a search bar, and tabs for '.env', 'use\_case\_generate.py', and 'use\_case\_infer'. Below the navigation bar, the project structure shows a 'LLM\_Workshop' folder containing subfolders '976d51b5-96cd-4004-80d4-03aad880328c', 'backup', 'data', and a '.env' file. The '.env' file is open in the editor, displaying the following content:

```
api_key=YgIJBtEHsjJ_P1AC8bhW7loXzTDd
project_id=ca2e06c0-c0dc-4114-8ec5-4
url=https://jp-tok.ml.cloud.ibm.com
# use jp for Japan and us for US
locale=jp
```

32. Run the script. The output will be shown in Python terminal.



```
C:\ProgramData\anaconda3\envs\Python310\python.exe C:\Users\1A3030897\PycharmProjects\LLM_Workshop\demo_wml_apiEL.py
-----
Question/request: Write a paragraph about the capital of France.
Answer: Paris is the capital of France, the most populous city of the country, and the largest city in the European Union.
-----
-----
Prompt:
From the following customer complaint, extract 3 factors that caused the customer to be unhappy.
Put each factor on a new line.

Customer complaint:
I just tried to book a flight on your incredibly slow website. All
the times and prices were confusing. I liked being able to compare
the amenities in economy with business class side by side. But I
never got to reserve a seat because I didn't understand the seat map.
Next time, I'll use a travel agent!

Numbered list of all the factors that caused the customer to be unhappy:
```

Let your instructor know if you have questions.

Next, we will invoke the LLM from a UI. We will use a popular web application development framework, *Streamlit*, to create a simple UI.

You can find the video of this application in the *lab repo/Reference* folder.

33. To run this script, you will need to install the *streamlit* package in your Python environment.

```
pip install streamlit
```

**Important:** If you're running on Windows, you will need to run this script in an Anaconda Python environment because it is the only supported [Python environment](#) on Windows. Both VS Code and Pycharm can be configured to use Anaconda. Please check with your instructor if you need help with setup instructions. If you're not able to install Anaconda in your environment, please check with your instructor.

34. Open the *demo\_wml\_api\_with\_streamlit.py* script. This application uses similar code to invoke the LLM as a previous example.

The application has 4 functions:

- *get\_credentials()*: reads the api key and the project id from the .env file (will be used for authentication)
- *get\_model()*: creates an LLM model object with the specified parameters
- *get\_prompt()*: creates a model prompt
- *answer\_questions ()*: sets the parameters and invokes the other two functions.

As you can tell by the name of the last function, this is a simple *Question and Answer UI*. You will notice that the prompt is more complicated than the prompt in the previous example: we provide instructions and a few examples (few-shot prompting).

Notice that we are hardcoding the instruction to answer the question. This is just an example, and you can choose to parameterize all components of the prompt.

35. When you run the script, Python will open the *Streamlit UI* in your browser.

If you invoke Python application from a terminal, and not an IDE then use the following command: `streamlit run demo_wml_api_with_streamlit.py`



*Note: When testing, ask "general knowledge" questions keeping in mind that our prompt is not sophisticated and that the model was trained on generally available data.*

## Test watsonx.ai LLM

Ask a question, for example: What is IBM?

Which country has the largest population?

**Answer to your question:** Which country has the largest population? *China, with 1.4 billion people in 2017 (representing 68.1% of the world's population)*

## Test watsonx.ai LLM

Ask a question, for example: What is IBM?

What is the hottest temperature ever recorded on Earth?

**Answer to your question:** What is the hottest temperature ever recorded on Earth? *134 °F (57 °C) in Death Valley, California, United States, on July 10, 1913*

## Test watsonx.ai LLM

Ask a question, for example: What is IBM?

Who invented an electric vehicle?

**Answer to your question:** Who invented an electric vehicle? *Nikola Tesla.. Tesla is credited with inventing the first electric car.. the Tesla Model X.*



## Conclusion

---

You have finished the **Introduction to Generative AI** lab. In this lab, you learned:

- Key features of LLMs
- The basics of prompt engineering, including parameter tuning
- Using the **Prompt Lab** to create and test prompts with models available in watsonx.ai
- Testing LLM model inference
- Creating a simple UI to let users interact with LLMs.

What we did not cover in this lab:

- Using LLMs to answer questions on data specific to the enterprise
- Executing multiple instructions with LLMs
- Implementing session management in LLM applications.

If you have questions about these topics, please contact your instructor.