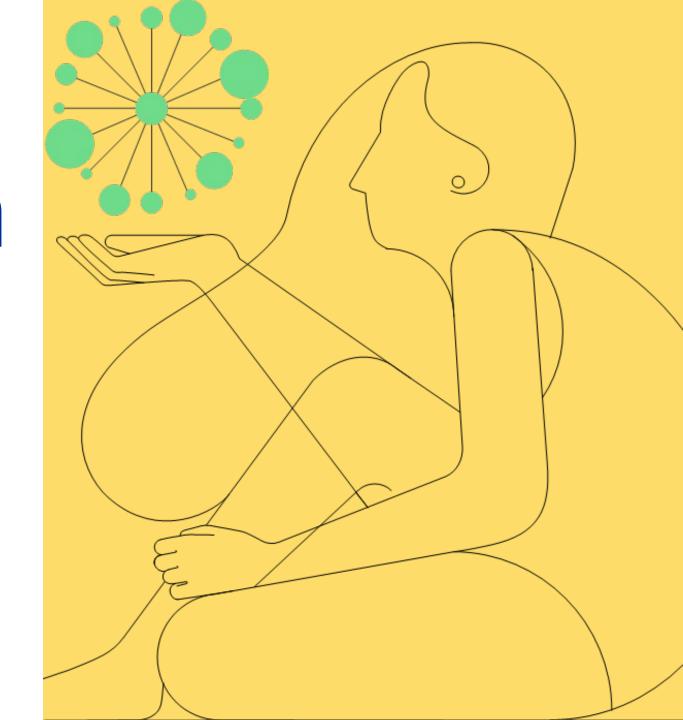
GenAI Challenge – Retail Industry

IBM Consulting



Introduction to the Customer



The Customer

Our customer is a leading retailer in the European market, with a strong presence in Spain and Portugal. With over 90 years of history, the company has grown to become a household name in the retail industry. The main activities of our customer include offering a diverse range of products and services across various categories such as clothing, electronics, home goods, and more. By continuously adapting to market trends and prioritizing customer satisfaction, our customer has successfully maintained its position as a prominent retailer, attracting millions of shoppers annually. As a pioneer in the retail sector, our customer is always seeking innovative solutions to enhance the shopping experience and optimize their operations.



The Market's challenges

In the last years, new players have appeared in the Spanish market, and the conditions have changed. Customers demand new services and products from retailers, and they need to adapt to continue leading the sector.

These are the main challenges for the sector:

- Adapting to changing consumer behaviour: Responding to shifts in customer preferences, shopping habits, and demands for personalized experiences.
- Embracing digital transformation: Implementing and integrating new technologies to enhance customer experience and streamline operations.
- Maintaining profitability amid competition: Balancing competitive pricing, product offerings, and quality to retain and attract customers.
- **Supply chain management:** Ensuring efficient inventory management, product availability, and timely delivery to meet customer expectations.
- Sustainable practices: Implementing eco-friendly and socially responsible initiatives to address growing environmental and ethical concerns.

How well prepared is our Customer?

Our customer, as a prominent retailer, is continuously adapting to meet these challenges. They have invested in their digital presence, providing customers with seamless online shopping experiences. Additionally, they work on improving supply chain management and embracing sustainable practices. They strive to maintain their competitiveness by offering a diverse range of products and focusing on customer satisfaction.

The Competition

New players in the market are leveraging their international presence, volume and experience to navigate the market's challenges.

Here are some strategies from other players:

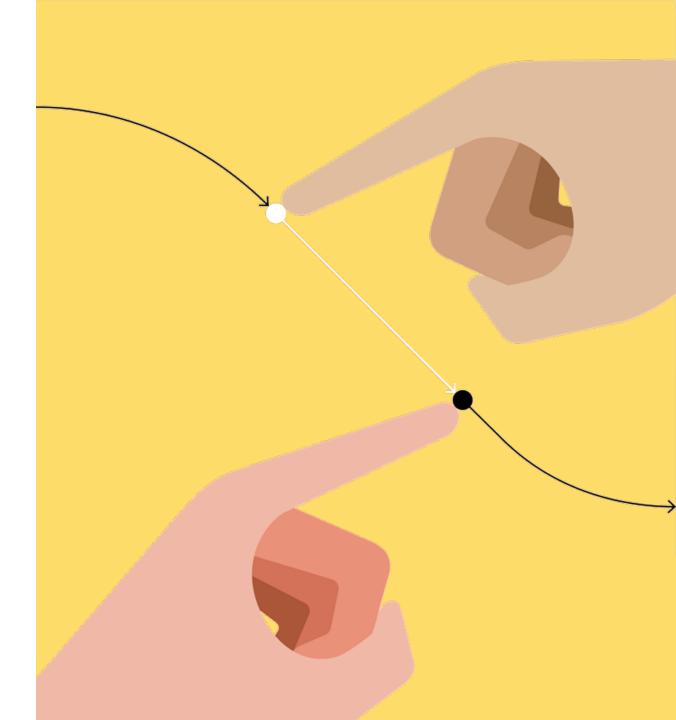
- Adapting to consumer behaviour: Use data analytics to provide personalized product recommendations, enhancing the customer experience.
- Embracing digital transformation: Develop mobile apps and instore technologies to streamline the shopping process and improve customer engagement.
- Maintaining profitability amid competition: Focus on offering exclusive products and collaborations to differentiate itself from competitors.
- **Supply chain management:** Use real-time data and an agile supply chain model to reduce lead times and respond quickly to market trends.
- **Sustainable practices:** Promote environmental responsibility and ethical supply chain practices, gaining customer loyalty and positive brand reputation.







Our challenge

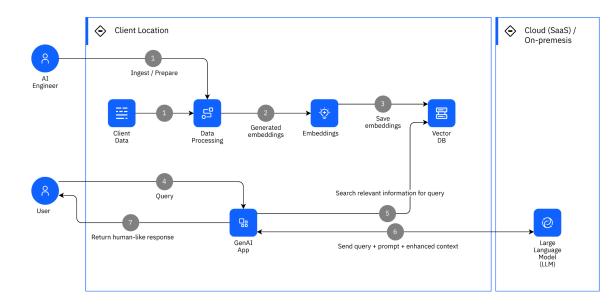


Our challenge

Develop an intelligent assistant for the retail industry using generative AI and RAG techniques. Students will work as a team to create a solution that vectors relevant information and uses this information to provide useful and accurate answers.

A few tips:

Organise yourselves, find your strengths and leverage them. Plan, then execute. Review, test, adjust.



Reference RAG architecture

Read more

Challenge's structure

We propose the following structure, which you can adjust:

1. Scope definition:

Determine what type of information will be used as the knowledge base for the virtual assistant (e.g., product descriptions, FAQs, customer reviews). Incluir/hacer referencia a los documentos

2. Development:

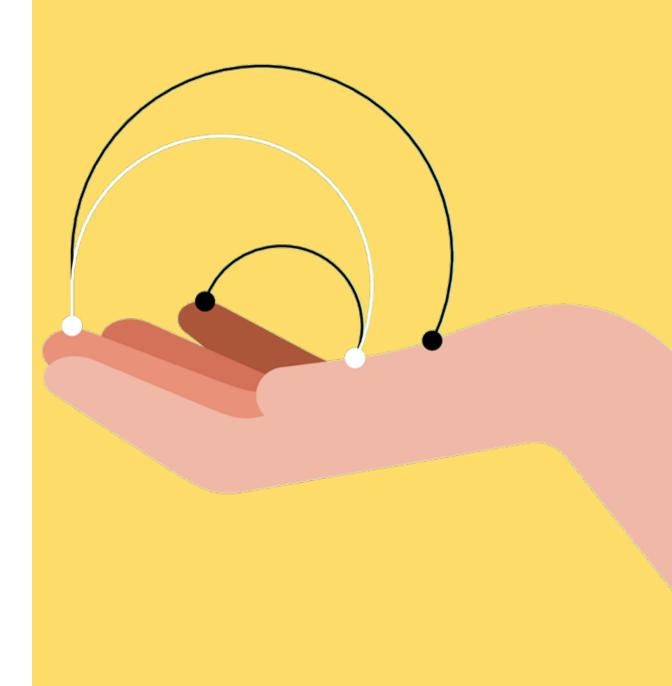
Review the chunking strategy for documentation storage

Prompt selection and adjustment according to desired responses

Implement the virtual assistant with the whole solution built.

3. Test and improve:

Test the virtual assistant with different types of questions and scenarios. Refine the solution as needed. Client's Dataset



Dataset overview

Here are some examples of the information you will receive:

Product department & description

Products will be categorized according to description, size, color and price.

Here are some examples:

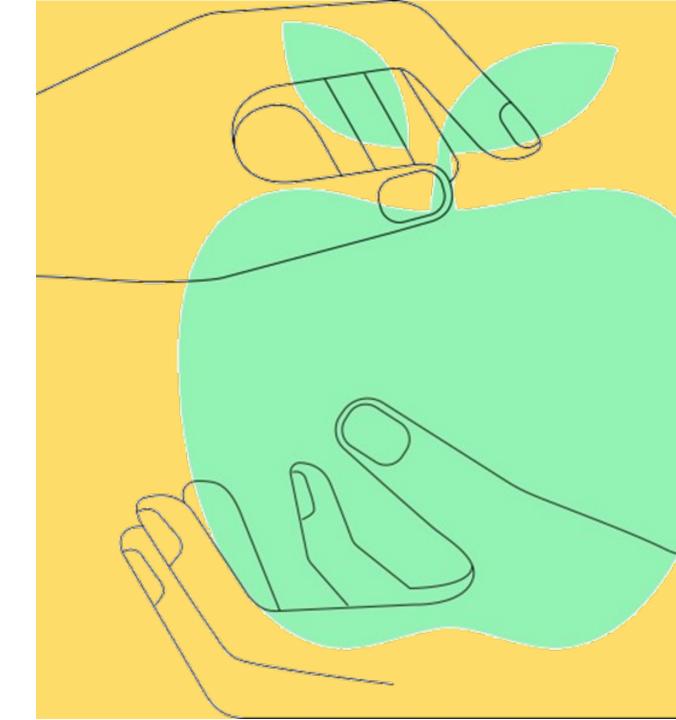
Cotton T-shirts

VintageCottonT-ShirtDescription: This t-shirt has a nostalgic air and a
retro touch. Its soft and slightly worn cotton will
transport you to another era. Imagine wearing it on
a sunny day, pairing it with high-waisted jeans.
Available
sizes: XS, S, M, L
Colors:
Light
gray, Beige
Price: €22.99

FAQs

- What are your customer service hours? Our customer service hours are Monday to Friday, from 9:00 AM to 6:00 PM. On Saturdays, we are open from 10:00 AM to 2:00 PM. We are closed on Sundays and holidays.
- What payment methods do you accept? We accept the following payment methods: Credit and debit cards (Visa, MasterCard, American Express), PayPal, Bank transfer, Cash on delivery (only available for domestic shipments)

Share your project



Solution proposal

You will be asked to

- 1. Present your proposal
- 2. Share the solution you've created with the format you are comfortable with: real time demo, prepared video demo....
- 3. Explain the process that you've followed to reach it, the decisions that took you there, and how you came up with the final architecture design.
- 4. The results you've obtained after testing and validating your solution.

We will assess based on the solution's approach, the level of innovation and your ability to present and defend your approach.



Resources & recommendations

Large Language Models (LLM) deployments

The use case is designed to use one of the available LLMs deployed in <u>IBM's watsonx.ai platform</u> - such as LLaMA3, and Mixtral 8x7B Instruct, amongst others – but feel free to experiment and use any other model available to you.

Refer to <u>Annex I</u> to see how to generate a watsonx.ai API key to be used in Python.

Embedding models

In order to calculate text similarity, it must first be converted into a high-dimensional array called *embedding*.

You can find a list of the most popular models – such as gte-base-en-v1.5 or e5-mistral-7b-instruct – in the *retrieval* tab of <u>HuggingFace's MTEB Leaderboard</u>.

Vector Databases

In a RAG pattern, documents – or rather, documents' embeddings – are stored in what is known as vector databases. While no vector DB is provided for this exercise, you can easily deploy your own locally. The most popular choices are <u>Chroma</u> and <u>FAISS</u>.

Don't forget to select a good chunking strategy before indexing your documents!

Resources & recommendations

LLM framework

While you can create all your code from scratch using the original libraries, using an LLM-specific framework such as <u>LangChain</u> is often faster. It is open-sourced, has an active community, and <u>is compatible with watsonx.ai</u>.

It has tons of tutorials and reference code snippets, so searching for a good sample code to use as a starting point usually pays off.

Prompt engineering & RAG techniques

Talking to an LLM is an art only mastered by a few. Educate yourself and read about <u>how to properly write a prompt</u>, as well as <u>how to implement a RAG-powered chatbot in LangChain</u>!

Resources & recommendations

	Tools	Link	Requerimientos	
Text generation	ChatGPT	<u>https://chat.openai.com/</u>	Gmail account	
	Bing Image Creator	bing.com/create	hotmail account	
Images	Lexica	<u>lexica.art</u>	gmail account	
generation	Stable Diffusion XL (imágenes realistas usando advanced settings > Photographic)	https://huggingface.co/spaces/google/sdxl	gmail account	
	Ideogram (ideal para imágenes con texto)	https://ideogram.ai/login	gmail account	
Internet intelligent search	Gemini	https://gemini.google.com/app	gmail account	
Search	Perplexity	<u>https://www.perplexity.ai/</u>	gmail account	
AI web design	Website designer	https://hf.co/chat/assistant/65b24c4e4914c9 938c4a1a34	gmail account	
Assistants	Hugging Face Assistants	https://huggingface.co/chat/assistants	gmail account	

Support Team Schedule

Week Day	Schedule	Meeting	Support Type
Tuesday 11th	13:00 - 14:00		Business
Tuesday 11th	15:00 - 16:00		Technical
Wednesday 12th	10:00 - 11:00		Business
Wednesday 12th	15:00 - 16:00	<u>Click to join</u>	Technical
Thursday 13th	15:00 - 16:00		Technical
Friday 14th	15:00 - 16:00		Technical & Business

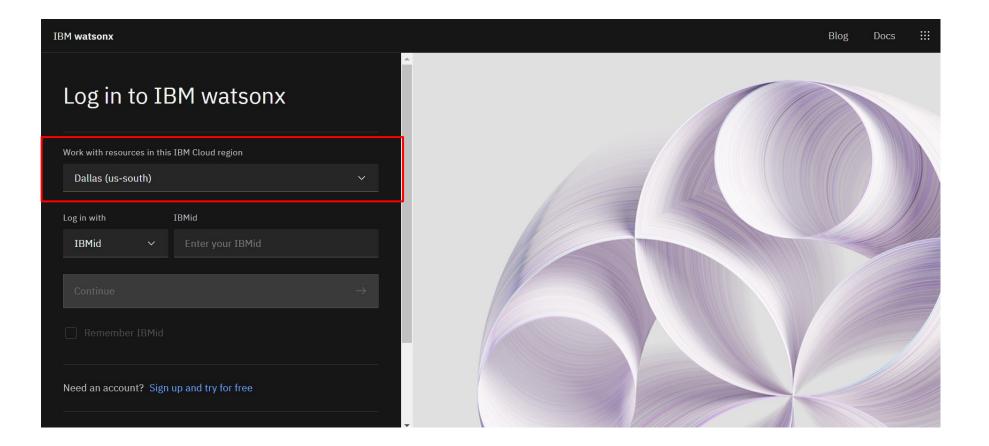
Thank you !



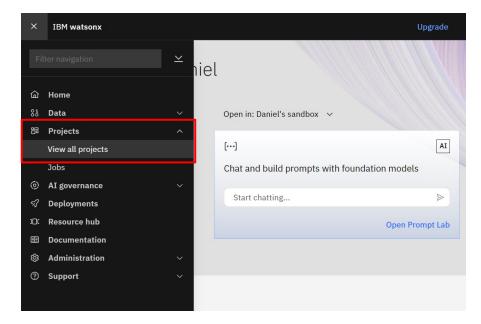
Annex I

Using watsonx.ai

Log in to the watsonx platform using your IBMid. For this workshop, set the region to Dallas.



Create a new project where you will deploy your Large Language Model

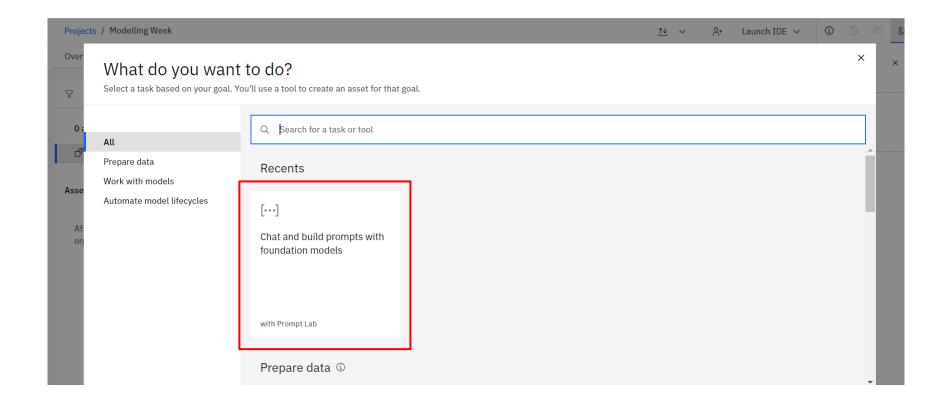


Projects							
√ Q Find a project						New project	+
Name	Date created	\downarrow	Your role	Collaborators	Tags		
ikc integration	3 months ago		Admin	DC			:

Let's add a new asset to the project

Projects / UCM Mode	elling Week	_			<u>+</u> +	°+	Launch IDE 🗸	í	J F	01 00
Overview	Assets	Jobs	Manage				Data in this proj	ect		×
√ Q Find asse	ts	1		Import assets 문 New as	set -		Drop data files here			
0 asset		All assets			G		browse for files to u			
□ [□] All assets										
Asset types After you create ass organized by asset				Start adding assets To get started with assets, click New asset create one in a tool, or Import assets to add existing ones.						

Select the asset called "Chat and build prompts with foundation models"



Come back to the "Assets" tab and add the prompt builder again. You can now open it and start using it. Let's switch from chat to structured form to better engineer our prompts.

Projec	cts / Modelling Week							
Over	What do you want Select a task based on your goal. Yo	to do? ou'll use a tool to create an asset for that goal.	Projects / Mod	elling Week / Prompt Lab	Freeform		New prompt + Model: granite-13b	AI guardrails off
0 ; Di Asse Af	All Prepare data Work with models Automate model lifecycles	Q [Search for a task or tool Recents [***] Chat and build prompts with foundation models	[··] ©		Quick start samples What are more efficient alternatives to a 'for loop' in Python? →	What is the Transformers architecture?	→	
		with Prompt Lab Prepare data ③			Create a chart of the top NLP use-cases for foundation models. →	Describe generative AI using emojis.	→	

Come back to the "Assets" tab and add the prompt builder again. You can now open it and start using it. Let's switch from chat to structured form to better engineer our prompts.

Proje	cts / Modelling Week								
Over	What do you want		Projects / Modellin	g Week / Prompt Lab	Freeform		Unsaved 🖾 🗸	New prompt + Model: granite-13b	-chat-v2 V TXT
♥ 0; Asse Af	Select a task based on your goal. Yo All Prepare data Work with models Automate model lifecycles	Q Search for a task or tool Recents [+++] Chat and build prompts with foundation models with Prompt Lab	() ()		Quick start samples What are more efficient alternatives loop' in Python? Create a chart of the top NLP use-ca foundation models.	→	What is the Transformers architecture?	 → → → → → 	
	l	Prepare data ^①							

This will be our main playground for refining and iterating our LLM prompt before implementing the final version in Python, where we can test different models and settings.

Pr	rojects / Modelling Week / Prompt Lab		Unsave	d 🖺 🖌 New prom	pt + 🖂 💽 AI guardrails on 🧭	
101	Sample prompts $< \kappa^2$	Chat Structured Freeform		AI Model: llama-3-8b	o-instruct ∨ {#} ππ ×	
[]	Not sure how to improve × your prompt?	Hint: This model works better when you provide at least 1 example.		×	Model parameters	
0	Try these prompt tips 🖸	Set up ^	Decoding	<u>^</u>		
	Summarization	Instruction (optional)			Greedy 💽 Sampling 🗿	
	Meeting transcript summary Summarize the discussion from a meeting transcript.	Answer the following question using only information from the article. If there is no good answer in the article, sa	ay "I don't know"		Temperature	
	Earnings call summary	Examples (optional)			0 <u> </u>	
	Summarize financial highlights from a quarterly earnings call.	Question:	Answer:		Top P (nucleus sampling) 0	
	Classification	Is growing tomatoes easy?	Yes, if you select varieties that are resistant to disease and pests.		Тор К	
	Scenario classification Classify scenario based on project categories.				1 100 50	
	Feedback classification Classify feedback about insurance customer service.	What varieties of tomatoes are there?	There are endless heirloom and specialty varieties.		Random seed	
	Generation	Add example +		Ū		
	Marketing email generation Generate email for marketing campaign.	Try ^			Repetition penalty	
	A Definition Senerate thank you note for workshop attendees.	Test your prompt ①				
	View all (3)	Question:	Answer:	IA	Stopping criteria ③	
	Extraction	Why should you use mulch when growing tomatoes?	Generated output appears here.		Stop sequences	Note:
	Act extraction Extract information from SEC 10-K sentences.	New test +		Ŵ	++ ×	Setting a random
	Question answering			_	Min tokens Max tokens	seed, such as 42,
	Questions about an article Answer questions about a body of text.				0 100	is recommended for replicable
	Questions about insurance Answer questions about insurance and technology		Clear output 🕥 Gener	ate \rightarrow	Enter up to 6 sequences to stop output after the minimum number of tokens is reached.	results.

Check the box saying "View in project after saving" and save the prompt as a prompt session to avoid losing your progress.

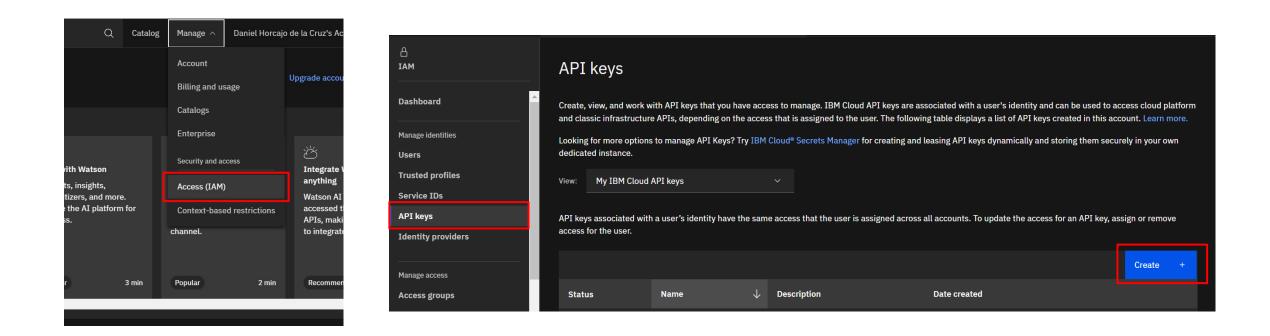
	Unsav	ed	8	^	New p	promp	t +	~
Freeform		S	ave as		llama-3	3-8b-i	nstruct	~
/hen you provide at least 1 example.					×	•	Mode	el para
							Deco	oding

Save your work Specify how to save your work by selecting an asset type and defining details.							
Asset type			Define details				
Prompt template Prompt session Notebook Name Save the current prompt only, without its history. Save history and data from the current session. Save the current prompt as a notebook. Description (optional) What's the purpose of this prompt asset?							
			✓ View in project after saving ①				
			Cancel Save				

Now, let's get the credentials needed to call the model from python. Head to the "Manage" tab and note down your project ID.

Overview Asets Jobs Maage Project © General © General © Access control © Access control © Access control © Access control Outling Week Modelling Week © Services & integration Obscription Used at as to make projects easist to find. Project ID Outling Weith Outling Week Project ID Outling Weith Outling Weith Project ID Outling Weith Outling Weith Outling Weith Outling Weith Outling Weith Outling Weith Project ID Outling Weith Outling Weith Outling Weith Outling Weith Project ID Outling Weith Outling Weith Project ID Outling Weith <th>Projects / Modelling We</th> <th>eek</th> <th></th> <th></th> <th></th> <th>11 ∨ A+ Launch IDE ∨</th> <th>()</th> <th>) 📮 🖁</th>	Projects / Modelling We	eek				11 ∨ A+ Launch IDE ∨	()) 📮 🖁
Image: Second of Constant Details Storage Storage	Overview	Assets	Jobs	Manage				
Petails Storage			General					
 I Resource usage Modelling Week Services & integrations Description Tags Adt ags to make projects easier to find. Friget ID 			Details		Ø	Storage		i
What's the purpose of this project? Bucket Tags 2 Add tags to make projects easier to find. Manage in IBM Cloud C ²	Resource usage	ons						
Add tags to make projects easier to find. Manage in IBM Cloud [] Project ID	-		What's the purpose of this proj	ect?	۵		6cs9ty	
				er to find.	<u>v</u>	Manage in IBM Cloud		
			-		-			

Now, let's generate our API key. Log in to IBM Cloud and click on "Manage" > "Access (IAM)". Next, click on "API keys" and generate a new one. Make sure to copy it before closing the dialog box – the API key is a secret and will not be shown again.



That's it! Now, we can call any model from watsonx.ai from python. Don't forget to set the URL to US south.

Here is a simple <u>example using LangChain</u>:

import os from langchain ibm import WatsonxLLM os.environ["WATSONX APIKEY"] = "YOUR API KEY" parameters = { "decoding_method": "sample", "max new tokens": 100, "min_new_tokens": 1, "temperature": 0.7, "top_k": 50, "top_p": 1, "stop sequences": ["["], 11m = WatsonxLLM(model_id="meta-llama/llama-3-70b_instruct", url="https://us-south.ml.cloud.ibm.com", project_id="YOUR_PROJECT_ID", params=parameters, response = llm.invoke("Tell me a joke about penguins") print(response)