

## Use Case - Configure built-in conversational search using the Granite LLM in watsonx.ai and Elasticsearch🔗

In this lab you will learn how to set up the out-of-the-box conversational search using the RAG pattern by connecting the assistant to an existing knowledge base (Elasticsearch) to search for relevant FAQs and generate the answer to the user's query using the Granite LLM model in watsonx.ai.

The high-level steps to accomplish this are as follows:

1. Open your assistant builder instance created in the previous lab
2. Configure the conversational search extension in the assistant instance
3. Test the virtual assistant's **content-grounded answering** by asking questions contained in the FAQ document (stored in Elasticsearch knowledge base) as well as additional questions answered by LLM only (**general-purpose answering**).

### Environment Details🔗

1. **IBM watsonx Orchestrate URL:** <https://dl.watson-orchestrate.ibm.com/>
2. **Credentials:** your IBM-id

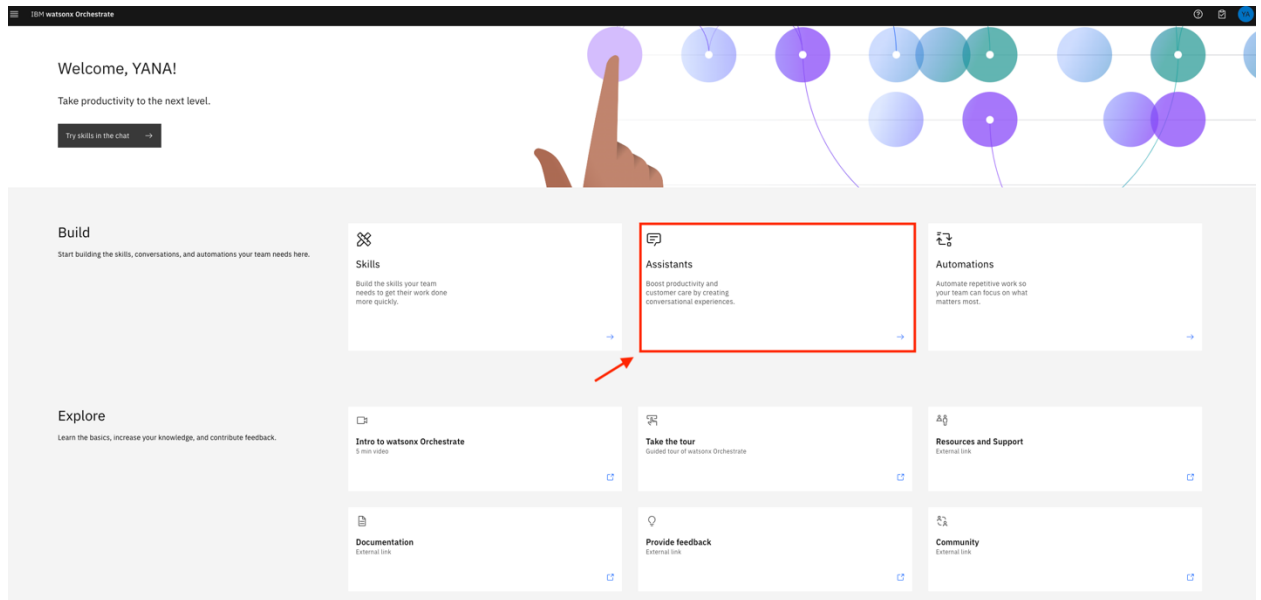
### Assumptions🔗

To accomplish this lab you need:

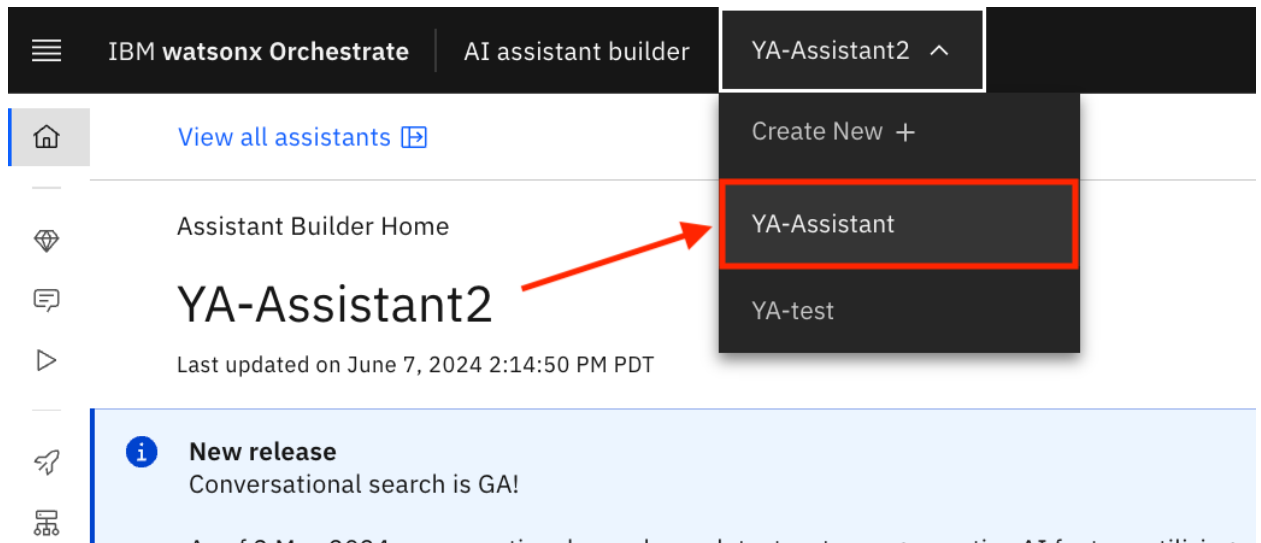
1. You have access to IBM watsonx Orchestrate URL.
2. You have access to Assistant Builder in the watsonx Orchestrate instance
3. Elasticsearch URL, username, and password (provided by email)

Step 1: Set up conversational search in your assistant builder instance🔗

1. Open the assistant builder:



2. Select your assistant instance if it isn't selected already:



3. Open the **Integrations** tab:



Home

Build



Generative AI



Actions



Preview

Deploy



Publish



Environments

Improve



Analyze



Integrations



Activity log



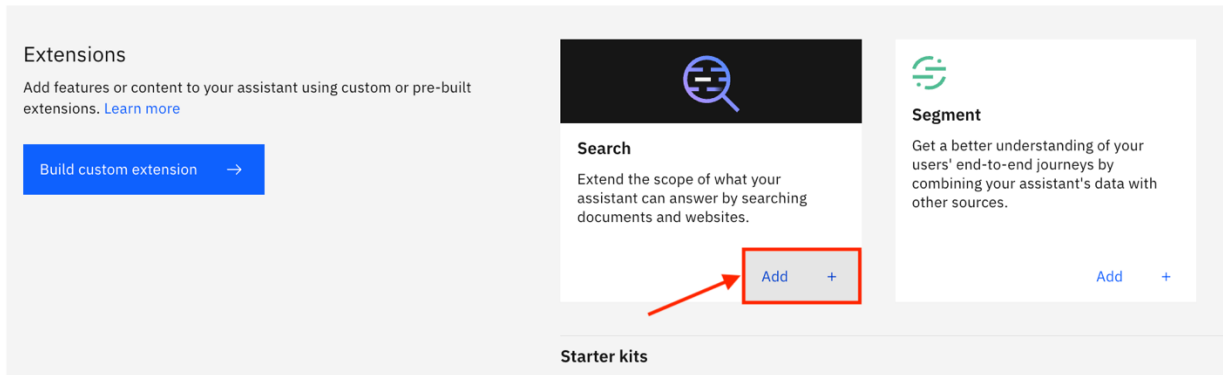
Assistant settings

cture



depicts the structure of

Integrations

4. Scroll down to **Extensions** and add **Search**:



5. Connect to Elasticsearch. Fill in the URL and port number (provided by your proctor). Leave authentication type as **Basic authentication** and fill in the username and password (provided by your proctor) and click **Next**:

 Connect Elasticsearch Select index Conversational search (optional)

## Connect to your Elasticsearch instance.

Fill in the information below to access your Elasticsearch instance. [Learn more](#)

Elasticsearch url

exampleUrl.com

Elasticsearch port (optional)

123

Choose an authentication type

Basic authentication

Elasticsearch username

exampleUser

Elasticsearch password

\*\*\*\*\*



Here is the Mapping:

Field	Value
Elasticsearch url	watsonx Discovery URL
Elasticsearch port (optional)	Watsonx Discovery port
Elasticsearch username	Watsonx Discovery username
Elasticsearch password	Watsonx Discovery password

- On the next screen select **Upload documents to a new index in your Elasticsearch instance** to automatically create a new index in Elasticsearch and click **Next** to continue:

Elasticsearch Draft

CloseNext

Connect Elasticsearch

Select index

Conversational search (optional)

## Select an index

Select the index you want to access from your Elasticsearch instance, or create a new index by uploading documents. [Learn more](#)

☐ Use my index

Elasticsearch index

exampleIndex

☒ Upload documents to a new index in your Elasticsearch instance Beta

2

1

7. On the next screen make sure conversational search is **On** and click **Save**:

Elasticsearch Draft

CloseSave

Connect Elasticsearch

Select index

Conversational search (optional)

## Enable conversational search (optional)

By enabling conversational search, your assistant can generate responses from supplied documents.

Conversational search will use a watsonx generative AI model hosted in Dallas, TX US, to generate conversational responses. [Learn more](#)

☒ On

By using this feature you agree to the [Pricing](#) and [Terms](#)

×

### What's different?

When your assistant receives a message, the default search behavior returns short excerpts of the best search results.

When conversational search is enabled, your assistant will instead generate a conversational response grounded in those search results.

#### Default search behavior (conversational search off)

Welcome, how can I assist you?

Can I use my points on airfare?

I searched my knowledge base and found this information which might be useful

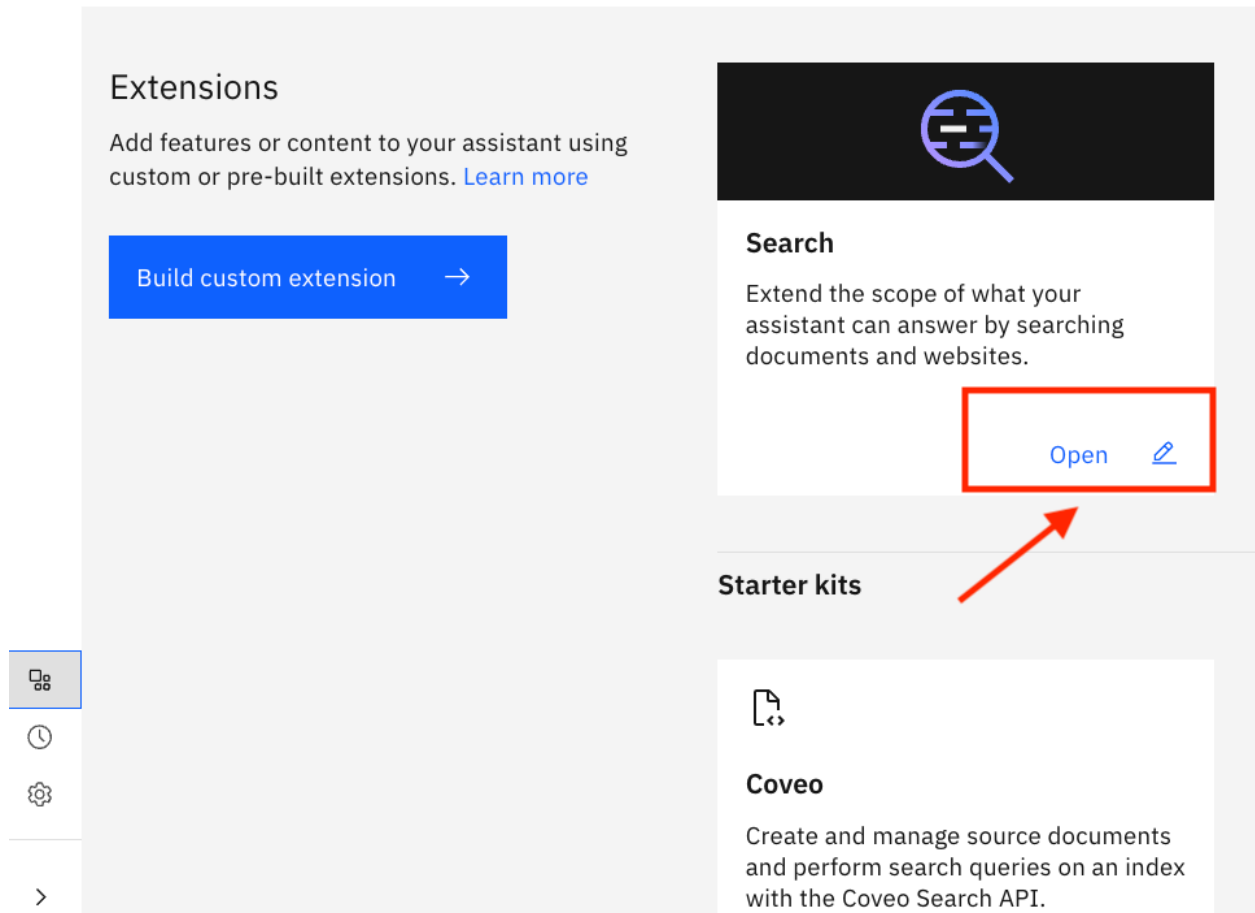
We offer credit cards offer zero liability to our clients, protecting you from having to pay for unauthorized purchases. Can I use my points on airfare? Using points and miles to book your flights is a great option. Not only do you save money, but you can use your points to book...

[IBM watsonx Lendyr Demo Site](#)

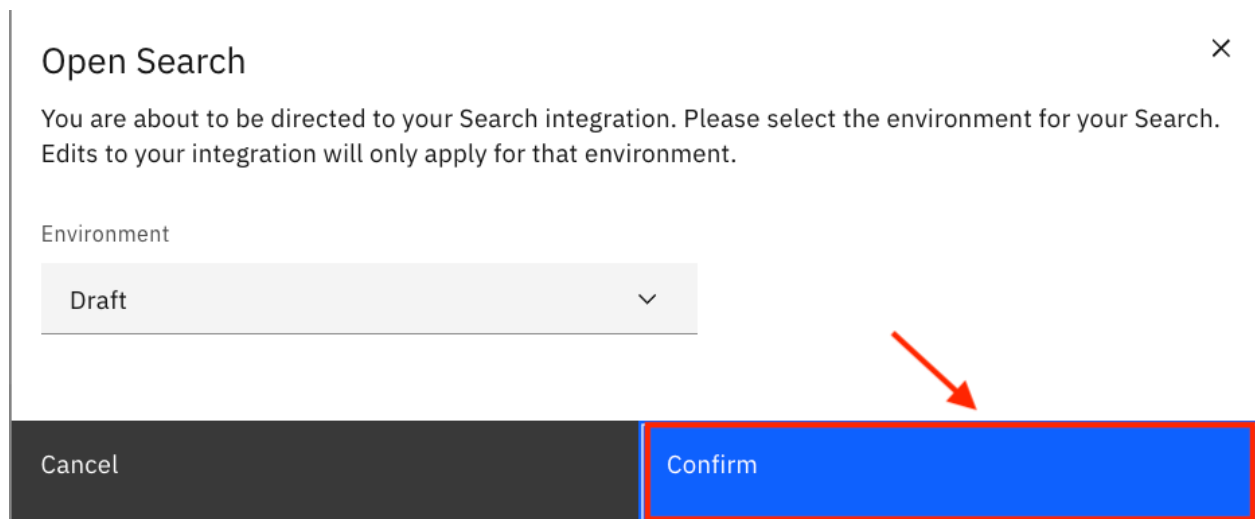
12

Step 2: Upload documents to the knowledge base<sup>1</sup>

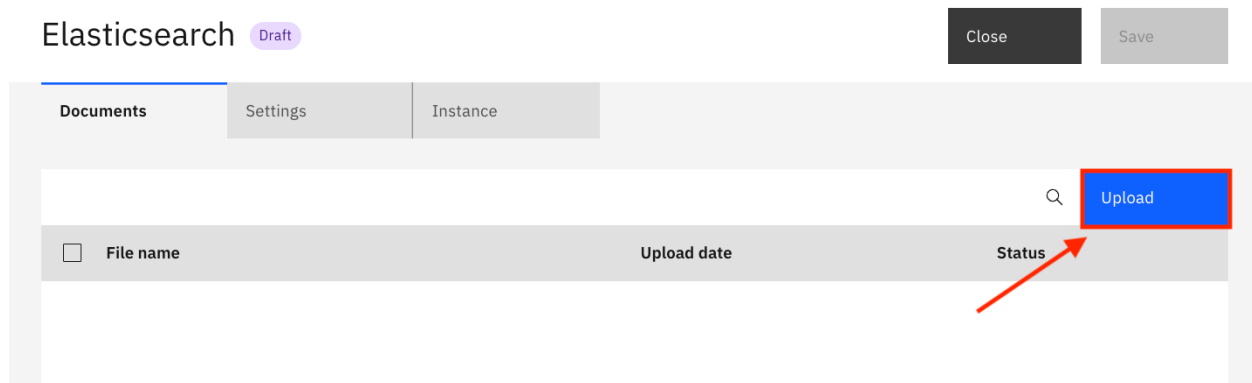
1. Now we need to upload our documents into the knowledge base. Go back into **Integration** tab and open the **Search** extension:



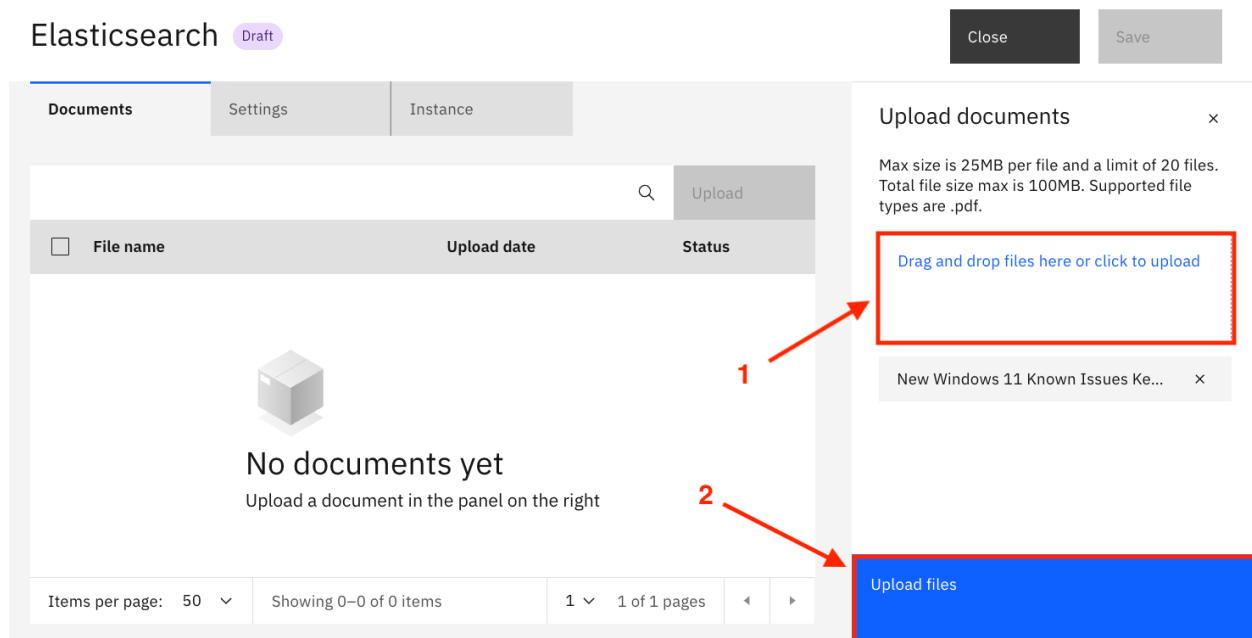
2. Click **Confirm** to open the extension:



3. In the **Documents** tab click **Upload**:



- Download [New Windows 11 Known Issues Keep Coming.pdf](#) (or anything you want) and drag and drop the .pdf file into the **drag and drop area**. Click on **Upload files**:



- Once the document is ready, the screen will look the following way and you can click on **Close**:



Elasticsearch Draft

Close Save

Documents Settings Instance

Upload

<input type="checkbox"/>	File name	Upload date	Status	
<input type="checkbox"/>	New Windows 11 Known Issues Keep Coming.pdf	6/7/2024, 5:23:03 PM	Ready	

Items per page: 50 Showing 1–1 of 1 items 1 1 of 1 pages

Upload documents ×

Max size is 25MB per file and a limit of 20 files. Total file size max is 100MB. Supported file types are .pdf.

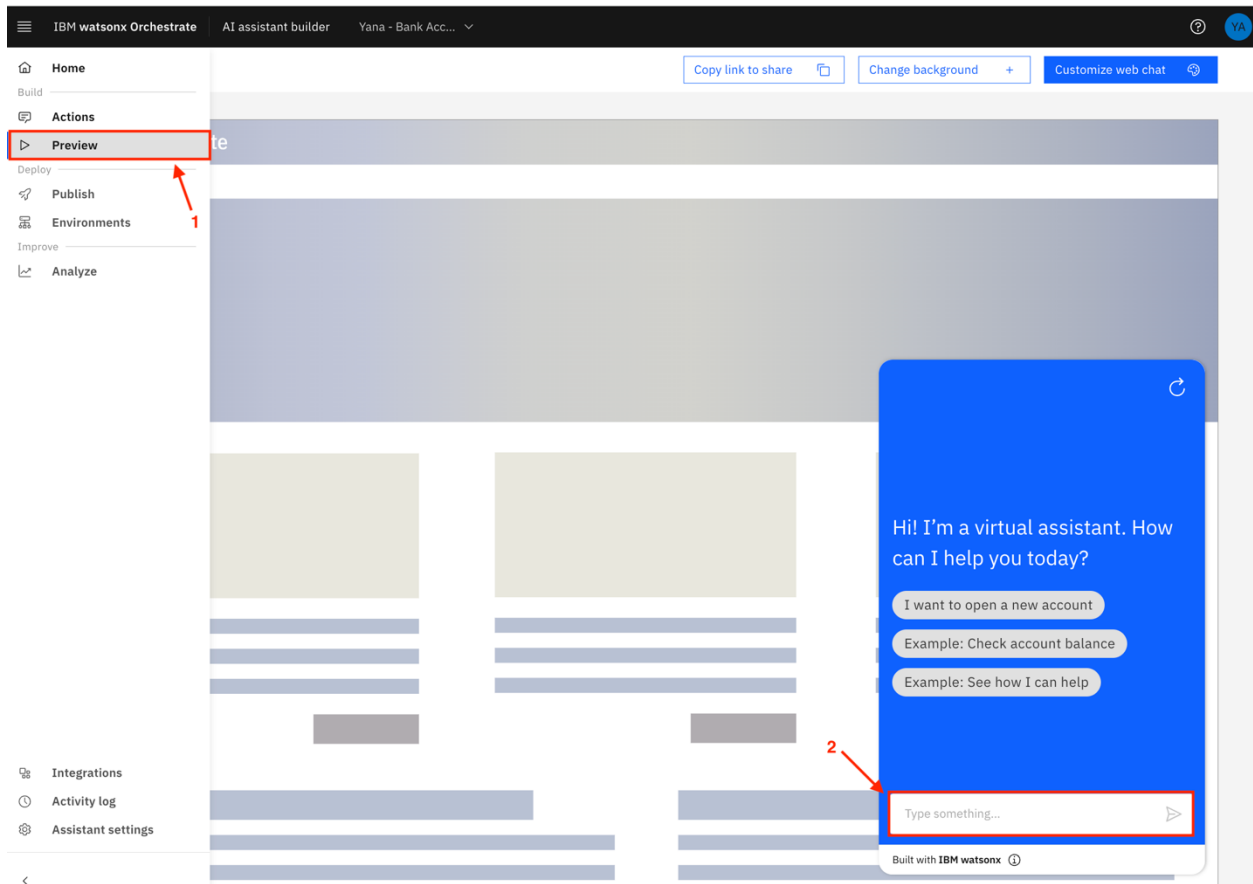
Drag and drop files here or click to upload

Upload files

### Step 3: Test the virtual assistant

Now that conversational search is set up and the document has been uploaded to the knowledge base, we can test the virtual assistant by asking some questions 1) based on the Q&As in the document 2) general questions that may be answered by a built in LLM.

1. Open **Preview** and go into the input window in the virtual assistant:



2. Type a question that is answered in the document, e.g. "how do I roll back from windows 11 to windows 10?" and check that the answer provided was retrieved from the document:

## Preview





You 5:35 PM

how do I roll back from windows 11 to windows 10?



5:35 PM

To roll back from Windows 11 to Windows 10, follow these steps: 1. Go to Settings > System and Recovery. 2. Click "Go back" next to the Previous version of Windows. 3. Check the reasons why you want to uninstall when prompted. 4. Click "No, thanks" when asked to check for updates instead of rolling back. 5. When a "What you need to know" window pops up, click "Next" to continue. 6. Click "Next" again when it warns you to remember your password from Windows 10. 7. Click "Go back to an earlier build." 8. The system will now restart and go through a restoration process that should take a few minutes. When that completes, you're back on Windows 10. 

 Accuracy of generated answers may vary.

 Content grounded answering called \_\_\_\_\_

Type something...



3. Type a question that is **NOT** answered in the document, e.g. "Are there lions in Yosemite?" and check if the LLM provides an answer:

## Preview



5:42 PM

I'm afraid I don't understand. Please rephrase your question.



General purpose answering called

You 5:42 PM

Are there lions in Yosemite?



5:42 PM

Yes, there are mountain lions, also known as cougars, in Yosemite National Park. Similar to Yellowstone, Yosemite is one of the few places in the United States where you can find free-roaming mountain lions. The park is home to a thriving population of these felines, which have adapted to the diverse habitats found within the park. There are typically between 20 and 30 mountain lions in Yosemite, making it an important habitat for this species. The presence of mountain lions in Yosemite is a reflection of the park's rich biodiversity and its ability to support a wide range of wildlife.



General purpose answering called

Use the up arrow for prior messages



4. You can test out different LLMs available under the **Generative AI** tab and observe how different the answers are to the same question:

The screenshot shows the IBM watsonx Orchestrate AI assistant builder interface. The top navigation bar includes 'IBM watsonx Orchestrate', 'AI assistant builder', and 'YA-Assistant'. The main section is titled 'Generative AI'. A red box highlights the 'Base large language model (LLM)' section, with a red arrow labeled '1' pointing to it. Below this, a dropdown menu is open, showing a list of LLMs: 'granite-13b-chat-v2 (recommended)', 'llama-3-70b-instruct', 'llama-3-8b-instruct', and 'mixtral-8x7b-instruct-v01'. A red arrow labeled '2' points to the dropdown menu. The 'Add instructions' button is visible below the dropdown. The 'Answer behavior' section is also visible, showing 'General-purpose answering' and 'Content-grounded answering' options, both of which are toggled on.

1. You can also experiment with the answer behavior by toggling **General-purpose answering** and **Content-grounded answering** options on and off. See how the answers change with different combinations of these settings:



## Base large language model (LLM) Beta

### Select a model

Select the large language model that your assistant uses for all base LLM functions.

granite-13b-chat-v2 (recommended) ▾



By using the model you agree to these terms.

[Read terms](#)

### Add prompt instructions

A default prompt is provided to produce the best general responses. Optionally, you may add instructions to refine the responses from the LLM. These instructions complement the default prompt, but are not integrated as part of it.

[Add instructions](#)

### Answer behavior

Specify how your assistant responds to customer requests.

#### General-purpose answering

Chat with an LLM that can converse on a wide array of general topics



#### Content-grounded answering ⓘ

Use content provided in the search integration as a basis for how the LLM will respond to inquiries

[View Search Integration](#) →



This final step concludes the lab. You configured the conversational search in your assistant with a knowledge base that stores FAQs (in Elasticsearch). When the user asks a question, the knowledge base is queried to retrieve any relevant FAQs which are then passed to a built-in watsonx.ai LLM (IBM Granite) to generate an answer for the customer. Great job!