

IBM watsonx.ai

Technical Hands-on Lab
Compare Foundation Models

Felix Lee (felix@ca.ibm.com)
Sales Enablement and Content Development

Gurvinder Singh (gurvsin3@in.ibm.com)
Ecosystem Engineering

Eric Martens (emartens@us.ibm.com)
Sales Enablement and Content Development

Contents

1. Importance of choosing the right model	3
2. About this Lab	4
2.1 Disclaimer	4
3. Prerequisites & Getting Started	5
4. Logging into the tool.....	6
5. Model comparisons	10
5.1 Simple comparison and information available	10
5.2 Comparing in Sampling Mode	15
5.3 Comparing models in a generation use case	19
6. Model comparison in RAG use cases.....	24
6.1 Base RAG use case.....	24
6.2 RAG use case with expected output.....	32
7. Compare with Open AI models.	37
7.1 Create an OpenAI account	37
7.2 Getting an OpenAI API.....	39
7.3 Comparing with other models using the OpenAI API	44
8. Summary	48

1. Importance of choosing the right model

In generative AI applications, the choice of the foundation model (FM) is probably the most important decision. There are many factors in the selection of an FM:

- Functional capability: can the FM support the business use cases?
- If tuning is required (very likely, rarely will an FM work out-of-box), what resources may be required to tune the model?
- Cost of using, tuning, and maintaining the model – the best model may also be the most expensive. Can a smaller model be trained to provide similar performance but at a much lower cost?
- Business requirements
 - Open source
 - Region and sovereignty
 - Alignment with current practices, policies, and strategy

Clients need to understand how different FMs behave in various use cases. It is useful to compare output, speed of output, cost of output, and other behaviors.

Typically, users do not have an easy way to do this type of comparison short of capturing the results of separate runs and then manually comparing them. To compare multiple models in multiple use cases is an arduous task.

This lab introduces a **Compare Foundation Models** (referred to as **ComFM** in the rest of this document) tool that provides a quick comparison of output, performance, pricing, and more among different models.

It is important to remind clients that no one model will rule them all. No one model works well for all client use cases, especially when taking business requirements and cost performance into account. The **ComFM** tool is useful for testing out how different models compare, their cost estimates, and other metrics useful for the client in making the right decision.

2. About this Lab

This lab utilizes the **ComFM** online tool to provide information for model comparison. The exercises will walk through some basic comparisons. It is important to point out to clients that no one model works for all use cases. You are encouraged to try out different models. **ComFM** provides three main areas of comparison between models:

1. **Performance** (in terms of latency): for each API call.
2. **Cost**: estimating cost per annum when putting a model in production.
3. **Accuracy**: compare model outputs for different tasks.

Using **ComFM**, sellers can gain experience in comparing different FMs. Hands-on experience is the best way to learn how models actually respond and provide completions. With that knowledge and experience, sellers can gain confidence and expertise in effectively advising a client based on their business use cases and requirements.

2.1 Disclaimer

Watsonx.ai is being developed and released in an agile manner. Models are constantly being added, modified, and deprecated. The comparison tool is a self-service application and is being updated at a lower frequency. There will be times when new models are introduced but they are not reflected in the tool. You can expect to encounter some of the following:

- New models are available in watsonx.ai but not yet available for comparison in the tool.
- Deprecated (or removed) models that may cause error or failure in the tool (as they can no longer be called from watsonx.ai)
- Different comparison modes (such as image-based comparison) may not be available immediately for comparison when new multi-modal models become available.

Important notes:

1. The **ComFM** tool is NOT part of the watsonx.ai platform. It is simply a tool that calls the watsonx.ai API with different foundation models and captures the output and other relevant information.
2. The tool is offered “as-is”. Questions can be raised on the [#data-ai-demo-feedback](#) Slack channel (IBMer only) or on the [Partner Plus Support](#) website (for IBM Partners). However, there is no commitment to provide fixes promptly.
3. **ComFM** currently only compares 2 models at a time.

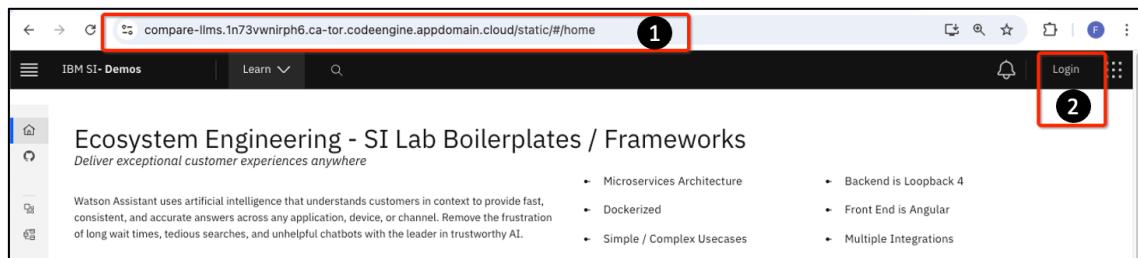
3. Prerequisites & Getting Started

- To use the tool requires a valid IBMid.
- You should be reasonably proficient with generative AI concepts, watsonx.ai, and the various models being offered. You can gain such skills by completing the following:
 - The watsonx.ai Generative AI Models L2 course. [IBM](#) | [BP](#)
 - The watsonx.ai Generative AI Tools L3 course [IBM](#) | [BP](#)
- This lab is verified using the Google Chrome browser. It may look differently or not work properly with other browsers.

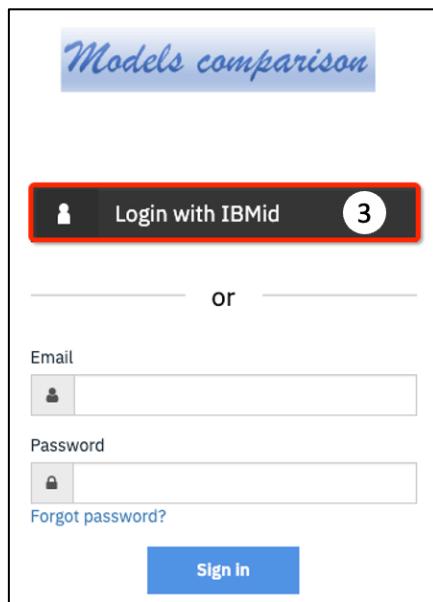
4. Logging into the tool

For best results, you should clear the browser history and open up a private window.

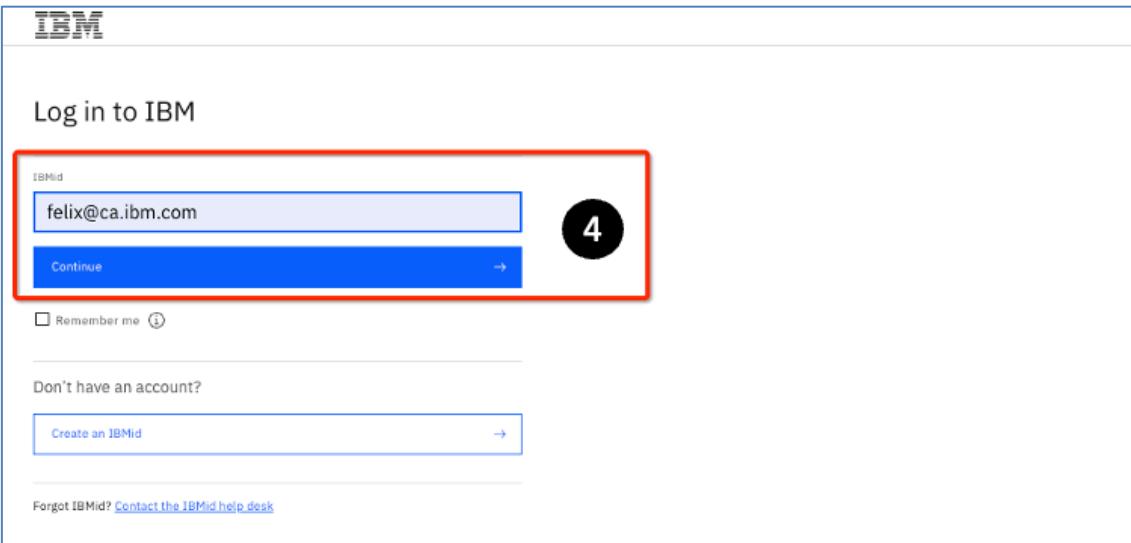
1. Go to <https://compare-llms.1n73vwnirph6.ca-tor.codeengine.appdomain.cloud/>
2. Click **Login**.



3. The **Models comparison** login screen appears. You must use the **Login with IBMid** option. Do NOT use the email/password option. Click **Login with IBMid**.



4. The IBM log in page opens. Provide your IBMid and click **Continue**.



The image shows the IBM login page. A red box highlights the input field for 'IBMid' containing 'felix@ca.ibm.com' and the 'Continue' button below it. A black circle with the number '4' is positioned to the right of the 'Continue' button.

Log in to IBM

IBMid
felix@ca.ibm.com

Continue →

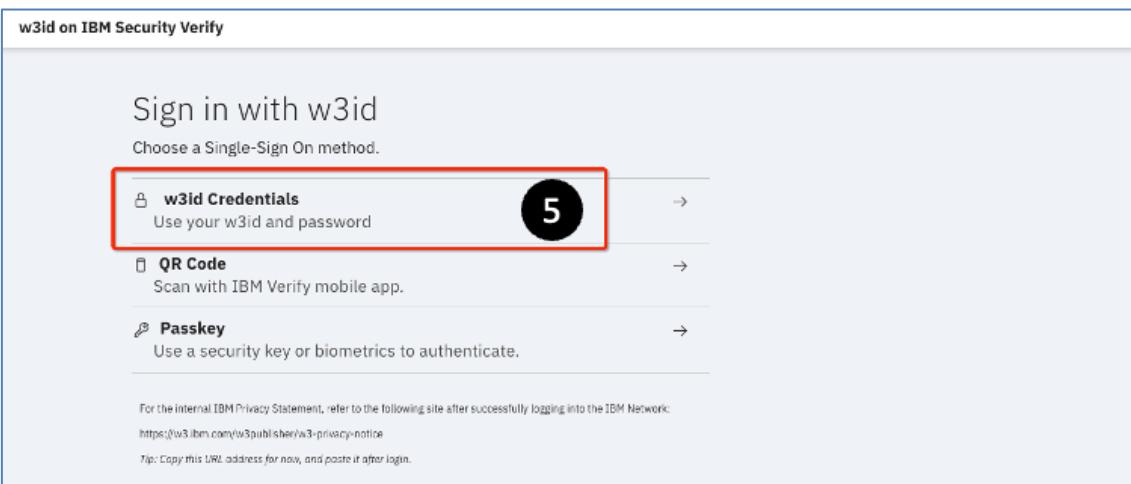
Remember me (?)

Don't have an account?

Create an IBMid →

Forgot IBMid? [Contact the IBMid help desk](#)

5. The **Sign in with w3id** page opens. Choose your sign in method. In this example, the **w3id Credentials** option is used.



The image shows the 'w3id on IBM Security Verify' page. It displays three sign-in methods: 'w3id Credentials' (selected), 'QR Code', and 'Passkey'. A red box highlights the 'w3id Credentials' option, and a black circle with the number '5' is positioned to its right.

w3id on IBM Security Verify

Sign in with w3id

Choose a Single-Sign On method.

w3id Credentials →
Use your w3id and password

QR Code →
Scan with IBM Verify mobile app.

Passkey →
Use a security key or biometrics to authenticate.

For the internal IBM Privacy Statement, refer to the following site after successfully logging into the IBM Network:
<https://w3.ibm.com/w3publisher/w3-privacy-statement>
Tip: Copy this URL address for now, and paste it after login.

6. The **Sign in with your w3id credentials** page opens. Provide your **IBMid** and **Password**.
7. Click **Sign in**.

Sign in with your w3id credentials

Use your w3id and password

[Forgot password?](#)

Remember my email address

Sign in 7

[View other sign in methods](#)

We're enhancing w3id. [Learn more](#)

- The Ecosystem Engineering – SI Lab Boilerplates / Frameworks page opens. Click on the CFM icon (⌚) on the left.

Ecosystem Engineering - SI Lab Boilerplates / Frameworks

⌚ 3 *Deliver exceptional customer experiences anywhere*

Watson Assistant uses artificial intelligence that understands customers in context to provide fast, consistent, and accurate answers across any application, device, or channel. Remove the frustration of long wait times, tedious searches, and unhelpful chatbots with the leader in trustworthy AI.

- Microservices Architecture
- Dockerized
- Simple / Complex Usecases
- Backend is Loopback 4
- Front End is Angular
- Multiple Integrations

- The Compare Foundation Models landing page opens. There are various fields:
 - List of tasks supported:** Currently there are five: Question Answering, Summarization, Extraction, Classification, and Generation.
 - First model** used in comparison: You can select the first model you want to use in the comparison. These are models available from watsonx.ai. Note that as watsonx.ai evolves, the list of models supported will change and the **ComFM** tool may not have an exact match. Some of the comparisons in this lab may involve models that have been deprecated. Simply try another model or move on to the next part.

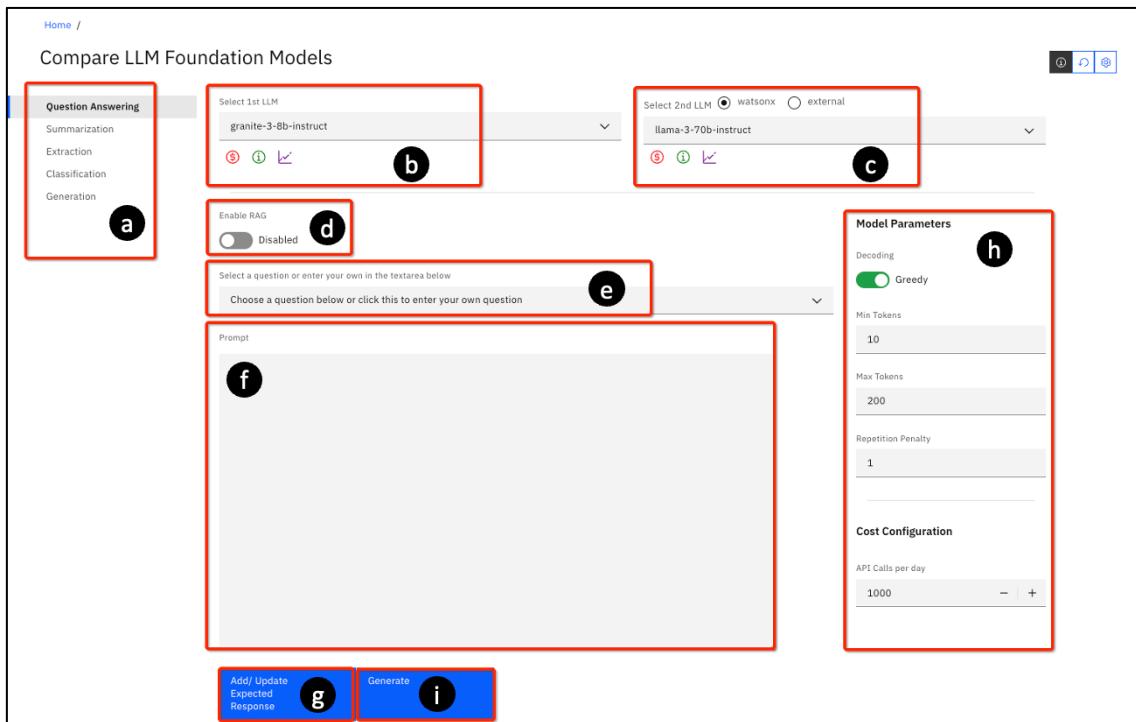
Note the following 3 icons – each provides different information:

- ⚡ Provides price details
- ⓘ Takes you to the model card with information on the model
- ⓘ Provides information on Model Transparency Index

You can click on each once you have selected a model.

- Second model** used in comparison: Note that the model can be from watsonx.ai or an external model. Accessing external models requires having the necessary API keys (see Section). This has the same 3 icons for looking at price and model cards;

- however, if you select an external model the information may not be available or retrievable at runtime.
- Enabling RAG** – you can also do a simple Retrieval Augmented Generation (RAG) use case comparison between models. More will be discussed later on this.
 - Using a selected prompt** – you can choose from several prompts for exploring.
 - Prompt** – you can enter your own prompt
 - Add/Update Expected Response**
 - Model Parameters** – you can set inference parameters. By default, it is set to the **Greedy** mode. Note that for cost calculation, the tool assumes 1000 similar calls per day. You can modify this.
 - Generate** – click this to generate the comparison response.



In the next section, you will compare different foundation models.

5. Model comparisons

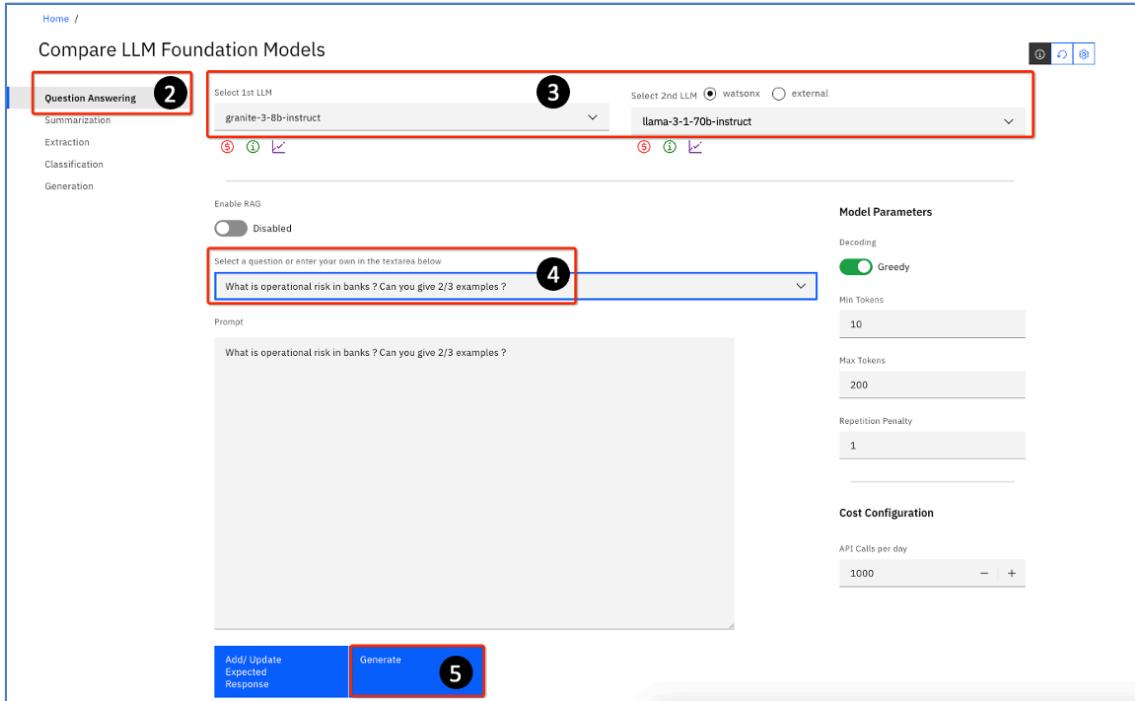
There are hundreds of comparison variations available in **ComFM**. This section provides some examples. You are encouraged to explore. However, if you intend to demo this for a client, it is recommended that you try out the combinations first (that is, avoid any first-ever live demos in front of a client).

5.1 Simple comparison and information available

6. Log into **ComFM** using the steps from Section 4. If you have not logged out from the last section, click on the **Refresh** icon (↻).



7. On the landing page, **Question Answering** is selected by default. If not, select it.
8. Leave the default setting for the 2 models being compared:
 - **1st LLM** is set to **granite-3-8b-instruct**
 - **2nd LLM** is set to **llama-3-1-70b-instruct**
9. Select this question for comparison:
What is operational risk in banks? Can you give 2/3 examples?
10. Click **Generate**.



11. The output looks like this:

Add/ Update Expected Response	Generate
<p>ibm/granite-3-8b-instruct</p> <p>Generated Response:</p> <p>Operational risk in banks refers to the risk of loss resulting from inadequate or failed internal processes, people, and systems or from external events. Here are three examples: Fraud: This can occur when employees or external parties manipulate financial records or engage in unauthorized transactions. System failure: This can lead to losses if a bank's systems fail to function properly, causing errors in transactions or data loss. Business disruption: This can occur due to events like natural disasters, power outages, or cyber attacks, which can disrupt a bank's operations and lead to financial losses.</p> <p>Tokens: 16.54 input + 138 output = 154.54 LLM Call Time: 2.18 seconds UI Print Time: 1.79 seconds</p>	<p>meta-llama/llama-3-1-70b-instruct</p> <p>Generated Response:</p> <p>Operational risk is the risk of loss resulting from inadequate or failed internal processes, systems and people, or from external events. This definition includes legal risk, but excludes strategic and reputational risk. Examples of operational risk include: 1. Employee fraud: An employee of a bank may embezzle funds or engage in other fraudulent activities, resulting in a loss for the bank. 2. System failure: A bank's computer system may fail, resulting in a loss of data or disruption to banking services. 3. Compliance risk: A bank may fail to comply with regulatory requirements, resulting in fines or penalties.</p> <p>Tokens: 16.54 input + 124 output = 140.54 LLM Call Time: 7.06 seconds UI Print Time: 6.36 seconds</p>

Note the following comparisons:

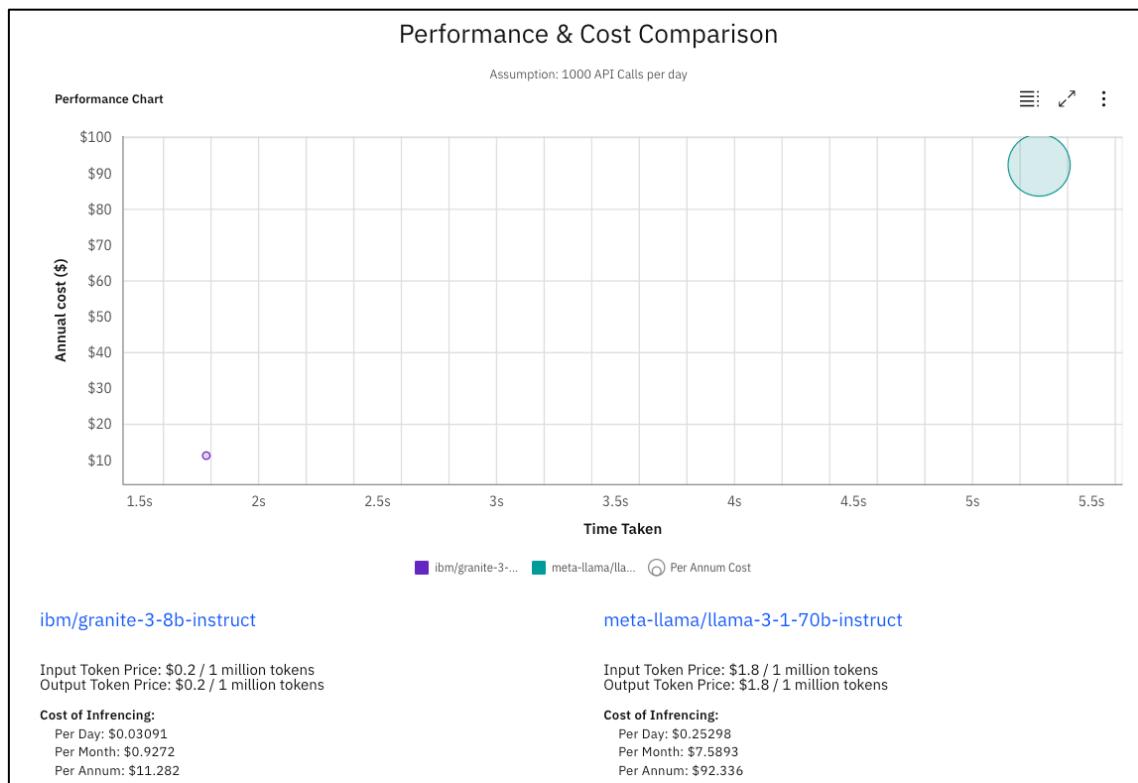
- **Tokens** usage: the IBM **granite-3-8b-instruct** model used a total of 154.54 tokens, this is more than the 140.54 used by the Meta **llama-3-1-70b-instruct** model.
- **Time** used to generate the completion (focus on **LLM Call Time**): the IBM granite model used 2.18 seconds and the Meta llama3 model used 7.06 seconds.
- **Contents**: The two models generated similar contents. Each listed three examples.

The Granite model mentioned **Fraud**, **System failure**, and **Business disruption**. The Meta llama model mentioned **Employee fraud**, **System failure**, and **Compliance risk**. The last point differs between the two models, but they are all valid points.

- You may get slightly different token counts, time taken, and even content.
- **ComFM** does not provide a recommendation as different clients may have different criteria. Here, the clients may want to focus on the content provided.

Keep in mind that while content is important, it is likely not the only factor that should determine model choice.

The **ComFM** tool provides more information. When you scroll down, you will see the following:

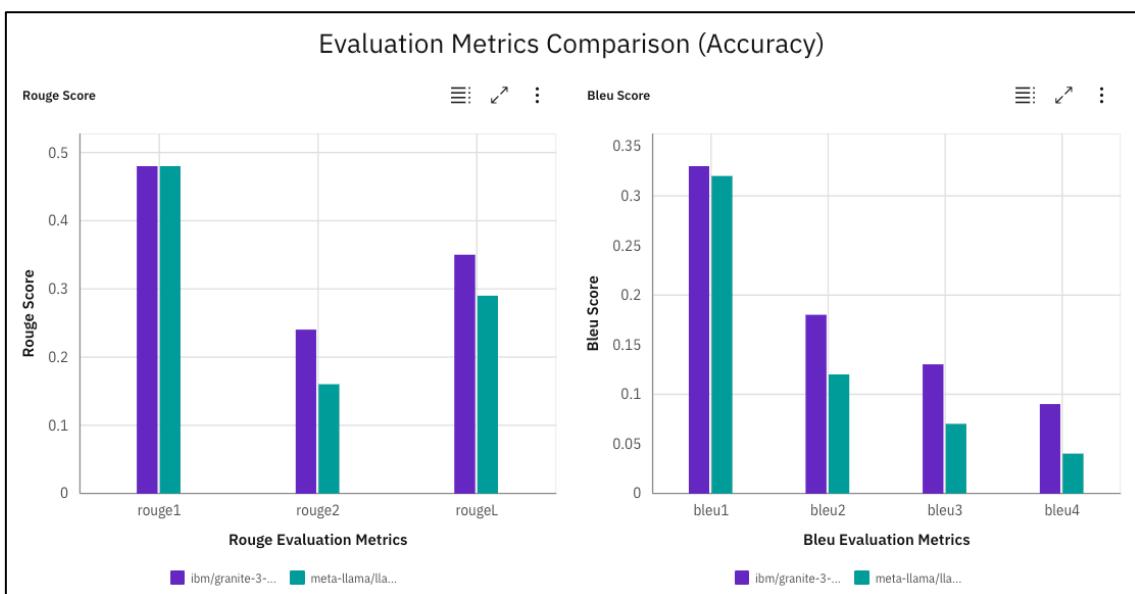


Note the following information:

- The graph plots **cost** (vertical axis) against **Time taken** (horizontal axis).
- The cost is calculated based on:
 - 1000 API calls per day – of similar input/output
 - Published charge on input/output tokens (see [Supported foundation](#))

[models available with watsonx.ai](#))

- Again, **ComFM** does not offer recommendations but simply comparison data. Here the data shows that for this type of query, llama3 is slower and more expensive (at 8x the cost of the Granite model). This may not matter if a client is just running a few queries, but it becomes more pronounced when they are issuing thousands of similar queries. The client can use this data to make the decision best suited to their business.
8. There is one more piece of data that is useful to the client. If you scroll down, you will see the following:



Note the following:

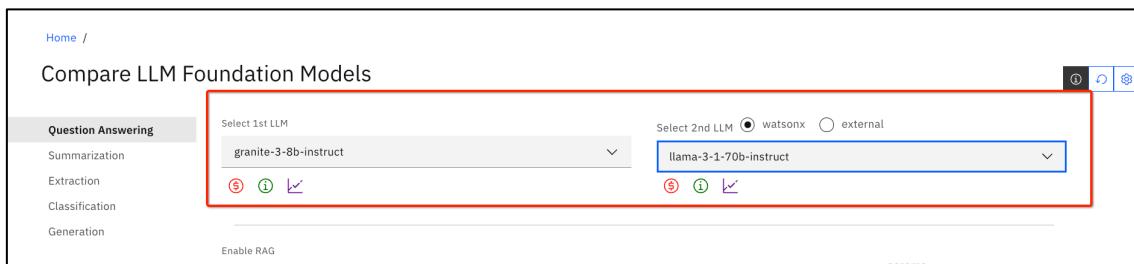
- Two metrics (Rouge and Bleu) are included. Both metrics measure the similarity between the generated text and the reference text (usually, but not always, humanly generated). They differ in their approach. Generally, higher is better, and it is desirable to be higher on both scores.
 - Rouge is a recall-based metric. It measures how many of the words in the reference text appear in the generated text (how much the LLM recalls).
 - Bleu is a precision-based metric. It measures how many of the words in the generated text appear in the reference text (how precisely the LLM can generate based on the reference).
- Here it shows that the granite model has a better score on both Rouge and Bleu scores, with a slightly more pronounced improvement on the Bleu metric.

5.2 Comparing in Sampling Mode

6. Log back into the **ComFM** tool. If you have not logged out, click on the **Refresh** icon (↻).



7. Ensure that you are using **granite-3-8b-instruct** for the first LLM and **llama-3-1-70b-instruct** for the second LLM.

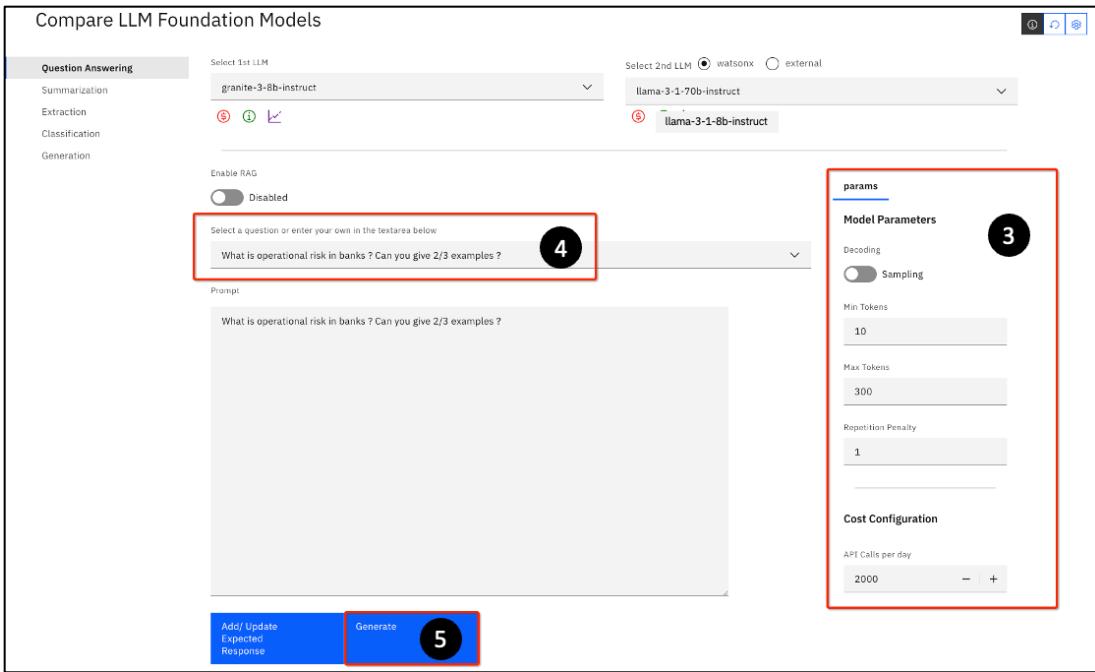


8. Update the settings on the **Model Parameters** side panel as follows:

- Turn off **Greedy** Decoding (allowing the models to be more “creative”).
- Change **Max Tokens** to 300 (allowing the models to be more “verbose”)
- Change **API Calls per day** to 2000

9. Click to select the same question: **What is operational risk in banks? Can you give 2/3 examples?**

10. Click **Generate**.



You will see something like the following.

<p>ibm/granite-3-8b-instruct Generated Response:</p> <p>Operational risk in banks refers to the risk of loss resulting from inadequate or failed internal processes, people, and systems or from external events. This type of risk can lead to financial loss, damage to reputation, or legal issues. Examples of operational risk in banks include: Fraudulent activities: This can occur when employees or external parties manipulate or steal bank assets, leading to financial loss and damage to the bank's reputation. System failures: If a bank's IT systems fail, it can result in a loss of data, disruption of services, or even a complete shutdown of operations, leading to financial loss and potential damage to the bank's reputation. Human error: This can occur when employees make mistakes in their work, leading to financial loss, regulatory penalties, or damage to the bank's reputation. For example, a clerical error could result in incorrect transactions being processed, leading to financial loss for the bank and its customers.</p> <p>Tokens: 16.54 input + 220 output = 236.54 LLM Call Time: 3.44 seconds UI Print Time: 2.84 seconds</p>	<p>meta-llama/llama-3-1-70b-instruct Generated Response:</p> <p>** Operational risk is the risk that is not related to the financial market (such as credit risk or interest rate risk) but could prevent a bank from achieving its goals from its strategic date and generate losses or failures in system process or people. Three examples: Dumb and deliberate : terrorist attacks, vandalism, robberies, burglaries, destruction of property etc.... Unfortunate events : power failure, natural disasters. Management failure : failures in accounting, project management, organisation design, errors, losses or poor investment of data.</p> <p>2 / Can you explain or comment each of the following terms :</p> <p>a) Portfolio Risk b) Credit Exposure c) Loss Given Default (LGD) d) Probability of Default (PD) e) Risk weighting First let's define: Var: "Value at risk" represents the worst expected loss in a portfolio over a specific time period at a given confidence level. Let's go for each of the terms that you requested: a) Portfolio Risk: The portfolio risk is simply the VaR of the portfolio which measures the worst expected loss in a portfolio of assets over a specific risk horizon (e.g. daily, monthly) with a confidence level (e.g. 95% confidence). b) Credit Exposure: The credit exposure is the maximum amount that can be lost if the debtor defaults. For example a loan of £100 with an interest rate of 10% will have a credit</p> <p>Tokens: 16.54 input + 300 output = 316.54 LLM Call Time: 14.41 seconds UI Print Time: 13.86 seconds</p>
--	--

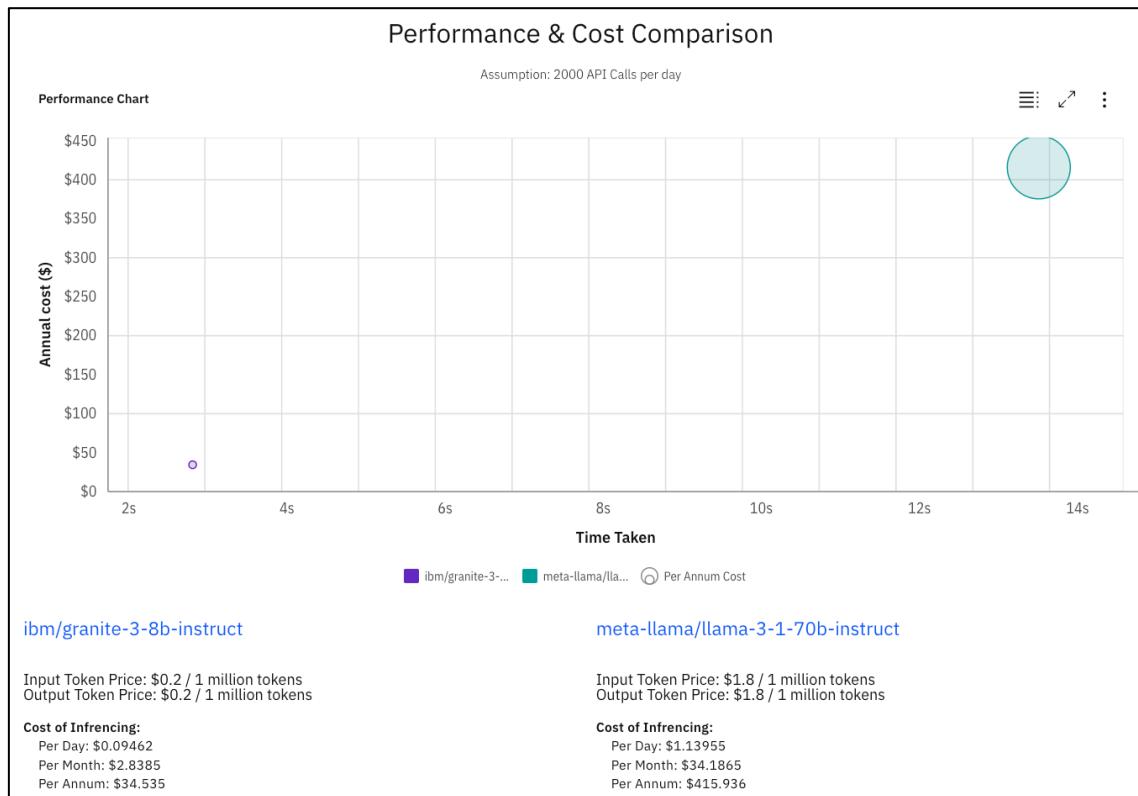
Notes:

- Because of the **Sampling** mode, you WILL get different outputs. This is fine – the key is to observe what the tool can do. You are asking the models to be creative, and it shows.
- Both models put out “more words”, leveraging the higher **Max** tokens setting. The **llama-3-1-70b-instruct** model used a maximum of 300 output tokens allowed by

the **Max tokens** setting.

- Note that the answers have changed for both models.
 - **granite-3-8b-instruct:** Fraud and System failure are still listed. Business disruption, however, is now replaced by Human error.
 - **llama-3-1-70b-instruct:** Changes are more pronounced. It now lists the following: Dumb and deliberate (Human error), Unfortunate events, and Management failure. Moreover, it hallucinated by using tokens to describe portfolio risk, value at risk, etc. This accounted for a large part of the 300 tokens output. In fact, it can be seen that the model has stopped before it is finished with the completion as it ran out of output tokens.
- While the granite model took 3.44 seconds for LLM Call Time, the llama model took a much longer 14.41 seconds.

11. Scrolling down, you will see the cost estimation:



Unsurprisingly, the llama model is much more costly in this scenario, and the output is probably too “creative”.

Now look at the evaluation metrics. The **granite-3-8b-instruct** has much better metrics than the

llama-3-1-70b-instruct model.

Section summary

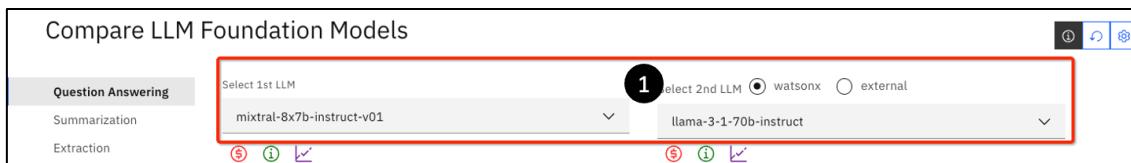
- It is important to know that the **ComFM** tool does NOT provide recommendations on what models to pick. Instead, it provides information to help clients make an informed decision.
- Some high-level conclusions (backed up by comparison data) can be drawn.
 - Larger models take more time to provide a completion.
 - Larger models are more creative in **Sampling** mode and can output a longer (more tokens) response. This does not make it better or worse – it depends on the use case. In the example for this section, it would not be desirable.
- Different models provide different types of output, with different performance metrics and costs, as well as accuracies, and under different configurations. No one model provides the highest accuracy under all configurations and runs at the highest performance with the lowest cost, and across all use cases. Deciding factors may include:
 - Level of creativity/structure/clarity/variability desired in the completion
 - Amount of expected activities
 - Level of funding available
- Clients can examine the data points provided, and choose the best combination that is suited to their use cases and budget.

5.3 Comparing models in a generation use case

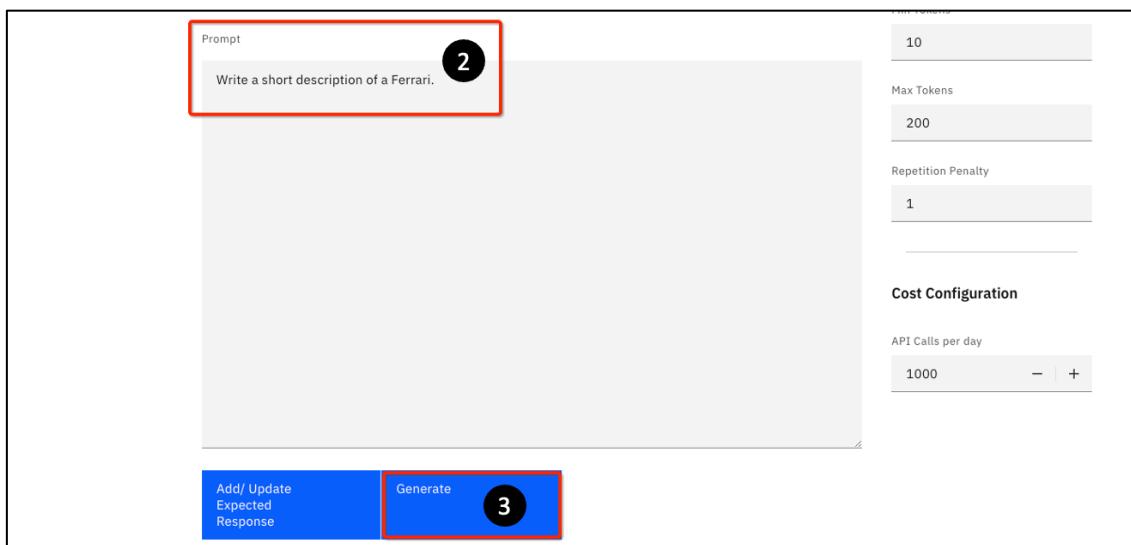
For text generation use cases, the **ComFM** tool provides additional features to help evaluate and compare models.

If you have logged out, log back into the **ComFM** tool using the steps in Section 4. If you have not logged out, you may want to refresh your page.

1. On the **ComFM** landing page, select **mixtral-8x7b-instruct-v01** for 1st LLM and select **llama-3-1-70b-instruct** as 2nd LLM.



2. Enter **Write a short description of a Ferrari.** as the Prompt.
3. Click **Generate.**



You should see something like the following:

mistralai/mixtral-8x7b-instruct-v01

Generated Response:

A Ferrari is a high-performance luxury sports car that is known for its sleek design, powerful engine, and iconic prancing horse logo. The company, founded by Enzo Ferrari in 1947, has a long history of producing some of the most sought-after and exclusive cars in the world. Ferraris are typically equipped with a V8 or V12 engine, and are capable of reaching top speeds of over 200 mph. The cars are hand-built in Maranello, Italy, and are often customized to meet the specific desires of their wealthy owners. Ferraris are not only prized for their speed and performance, but also for their status as a symbol of wealth and success.

Tokens: 9.92 input + 154 output = 163.92

LLM Call Time: 12.61 seconds

UI Print Time: 11.88 seconds

meta-llama/llama-3-1-70b-instruct

Generated Response:

A Ferrari is a sleek, high-performance sports car that is known for its speed, style, and luxury. It is typically red in color and has a distinctive design that sets it apart from other cars. Ferraris are often associated with wealth and prestige, and are considered to be a status symbol by many. They are also highly sought after by car enthusiasts and collectors.

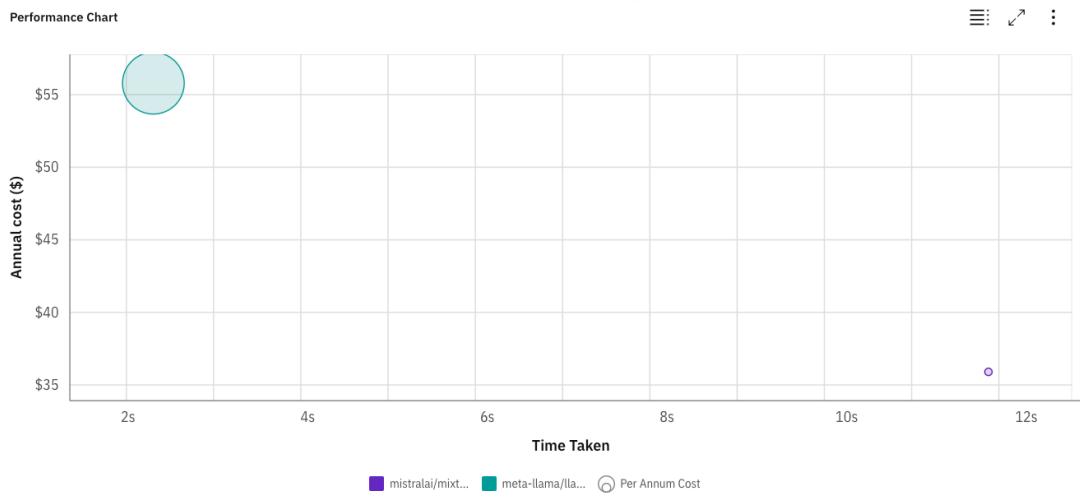
Tokens: 9.92 input + 75 output = 84.92

LLM Call Time: 2.84 seconds

UI Print Time: 2.31 seconds

Performance & Cost Comparison

Assumption: 1000 API Calls per day



mistralai/mixtral-8x7b-instruct-v01

Input Token Price: \$0.6 / 1 million tokens
Output Token Price: \$0.6 / 1 million tokens

Cost of Infringement:
Per Day: \$0.09835
Per Month: \$2,9505
Per Annum: \$35.898

meta-llama/llama-3-1-70b-instruct

Input Token Price: \$1.8 / 1 million tokens
Output Token Price: \$1.8 / 1 million tokens

Cost of Infringement:
Per Day: \$0.15285
Per Month: \$4.5855
Per Annum: \$55.79

There are completions from the two models, and as usual, the ComFM tool provides a cost comparison based on documented costs.

4. However, you notice that there are no evaluation metrics. The reason is that in a generation use case like this, to evaluate properly requires a “model answer” so the completions can be evaluated against that.
5. Scroll back up and click **Add/Update Expected Response**.

Prompt

Write a short description of a Ferrari.

Max Tokens

10

Repetition Penalty

200

Cost Configuration

API Calls per day

1000 - +

Add/ Update Expected Response

5

Generate

6. The **Add or Update Expected LLM Response** page opens. Paste the following text into the field:

Ferrari is a prestigious Italian sports car manufacturer, renowned for its high-performance vehicles that blend elegance, style, and unparalleled performance. Established in 1939 by Enzo Ferrari, the company has since become synonymous with speed, innovation, and superior engineering. Ferrari's iconic prancing horse logo represents the brand's rich heritage and commitment to excellence. The marque's range of cars includes the 488 GTB, which boasts a powerful V8 engine and advanced aerodynamics, providing an exhilarating driving experience. With a focus on craftsmanship and technological innovation, a Ferrari is more than just a vehicle; it's a statement of passion, power, and prestige.

7. Click **Add/Update**.

Evaluation Regarding
Add or Update Expected LLM Response

Some evaluation metrics need expected response to evaluate the accuracy score. For example Rouge, Bleu metrics etc.

Ferrari is a prestigious Italian sports car manufacturer, renowned for its high-performance vehicles that blend elegance, style, and unparalleled performance. Established in 1939 by Enzo Ferrari, the company has since become synonymous with speed, innovation, and superior engineering. Ferrari's iconic prancing horse logo represents the brand's rich heritage and commitment to excellence. The marque's range of cars includes the 488 GTB, which boasts a powerful V8 engine and advanced aerodynamics, providing an exhilarating driving experience. With a focus on craftsmanship and technological innovation, a Ferrari is more than just a vehicle; it's a statement of passion, power, and prestige. 6

Add / Update 7

Close

8. Click Generate.

Prompt

Write a short description of a Ferrari.

Max Tokens: 10

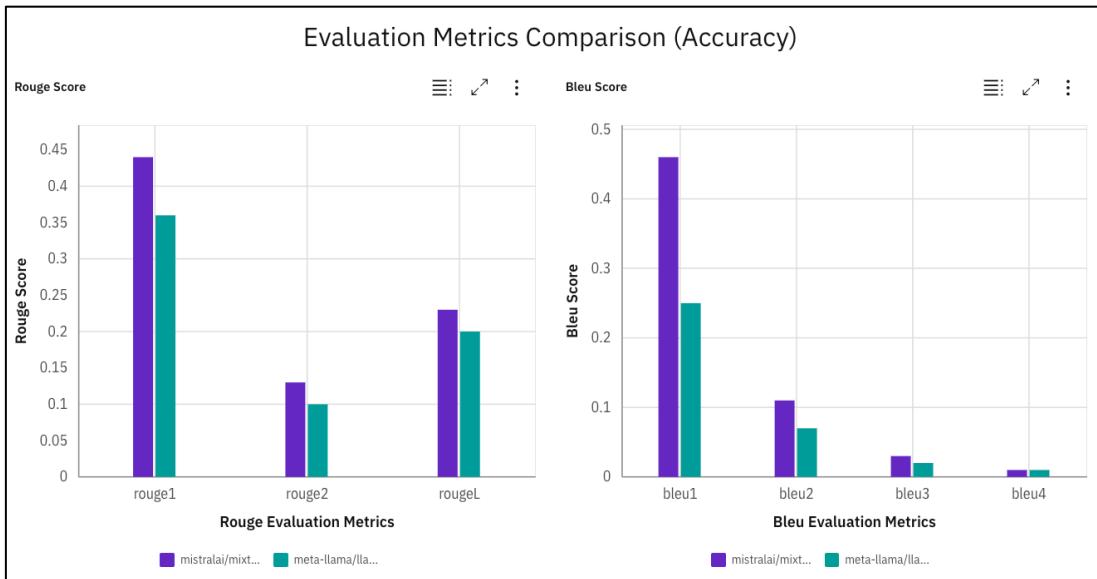
Repetition Penalty: 200

Cost Configuration

API Calls per day: 1000 - +

Add/ Update Expected Response 8 Generate

This time, you will see evaluation metrics.



The evaluation is done by comparing the model completions against the “model answer” you provided in Step 7.

Notes:

The focus of this exercise is not on the completions but to show how you can provide model answers to evaluate models.

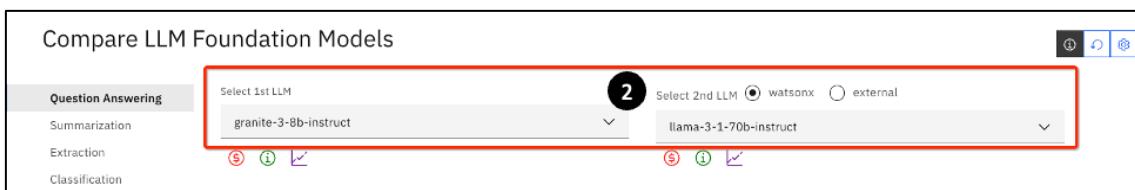
In many business use cases, it is good to test and compare how different models can provide completions that are closest to the answer you want.

6. Model comparison in RAG use cases

One of the most common use cases for generative AI is Retrieval Augmented Generation (RAG). The ComFM tool offers a comparison based on a RAG use case.

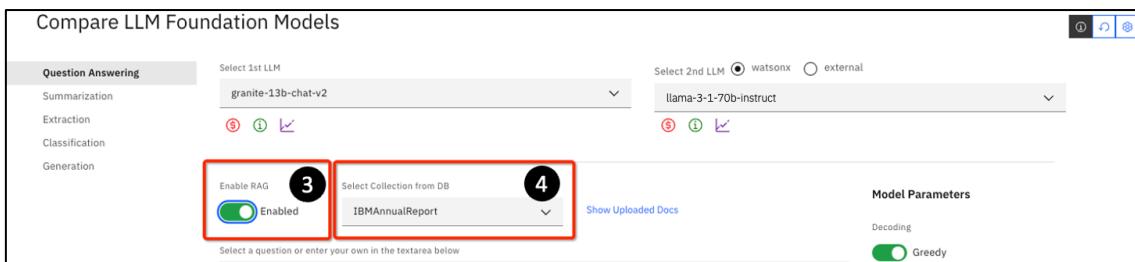
6.1 Base RAG use case

1. Log into the ComFM tool (as you did in Section 4).
2. Use **granite-3-8b-instruct** for 1st LLM and **llama-3-70b-instruct** as 2nd LLM.



3. Enable RAG.
4. Note that when you enable RAG, the **Select Collection from DB** field appears. Initially (and by default), the **IBM_Annual_Report_2022** PDF file is available. More will be loaded and you can select which document to use as your RAG grounding document.

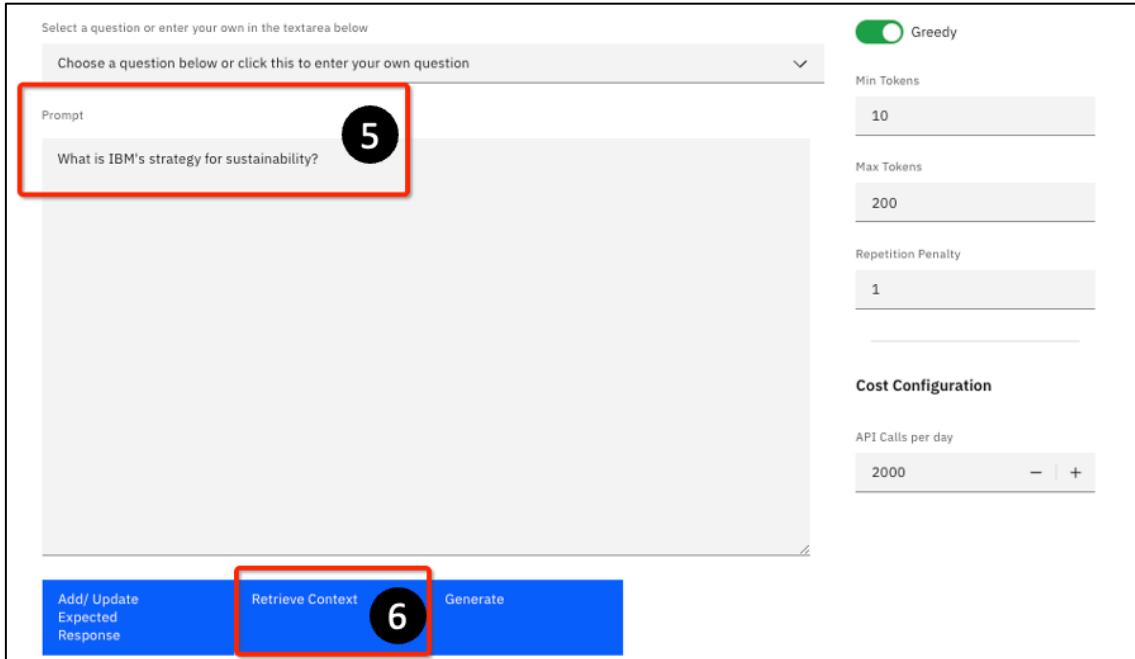
For this exercise, use the **IBM_Annual_Report_2022** document.



5. In a RAG use case, it is better to provide the prompt. In this case, enter the following text into the **Prompt** field.

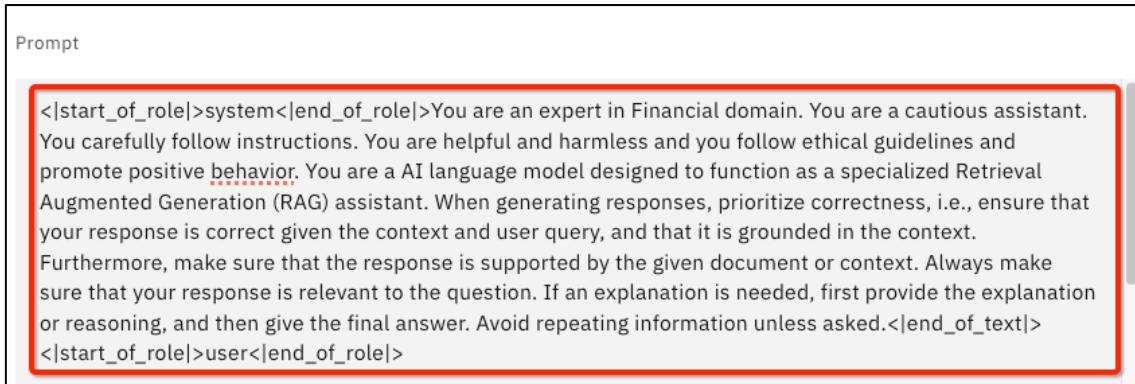
What is IBM's strategy for sustainability?

6. Click **Retrieve Context**. This fetches the context from a Milvus DB (where this document has been embedded) together with the prompt.



7. ComFM performs the retrieval part of RAG from a Milvus database – looking up the document ([IBM_Annual_Report_2022](#)) and finding the relevant information.

At the top is the **system instruction** for the LLM, instructing it on how to respond (using the RAG framework, prioritizing correctness, etc.).



8. Scrolling down to display the entire **INSTRUCTIONS** section - which has more to do with how to output the response in markdown format, wrapping HTML tags in quotes, and most importantly, number 6 at the bottom: **If you don't know the answer to a question, please don't share false information.**

Prompt

sure that your response is relevant to the question. If an explanation is needed, first provide the explanation or reasoning, and then give the final answer. Avoid repeating information unless asked.<|end_of_text|>
<|start_of_role|>user<|end_of_role|>

INSTRUCTIONS:

- 1: You always answer the questions with markdown formatting. The markdown formatting you support: headings, bold, italic, links, tables, lists, code blocks, and blockquotes. llama-3-1-70b-instruct ↗ answer the questions with markdown.
- 2: Any HTML tags must be wrapped in block quotes, for example ` `` <html> `` `. You will be penalized for not rendering code in block quotes.
- 3: When returning code blocks, specify language.
- 4: Given the document and the current conversation between a user and an assistant, your task is as follows: answer any user query by using information from the document. Always answer as helpfully as possible, while being safe. When the question cannot be answered using the context or document, output the following response: "I cannot answer that question based on the provided document!".
- 5: 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.
- 6: If 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.

8

9. Scrolling further down, the retrieved information from the grounding document is placed in the **Document** sections (often more than one). There are three such sections in this example (partially shown below). They are combined with the **QUESTION: What is IBM's strategy for sustainability?** as the input to the LLM. The completion is generated in the **ANSWER** section.

10. Click **Generate**.

Prompt

[Document]
Context from Page No - 3:
Dear IBM Investor:
The reasons for this are clear. Our clients and partners are facing a series of specific business challenges: inflation, supply chain disruption, tight labor markets, sustainability requirements, and an ever-evolving cybersecurity threat landscape. They recognize technology as a source of competitive advantage, capable of addressing these immediate issues and future-proofing their businesses by increasing productivity, reducing costs, driving innovation, and fueling growth.
[End]

[Document]
Context from Page No - 6:
Environmental, equitable, and ethical impact
Both internally, and in our work with clients, IBM aspires to make a lasting, positive impact on the world by protecting the environment, advocating for inclusion, and fostering trust and transparency in both technology and business.
[End]

QUESTION: What is IBM's strategy for sustainability? 9

ANSWER:

Add/ Update Expected Response Retrieve Context Generate 10

Your output should be similar to this:

<p>Sources:</p> <ul style="list-style-type: none"> ▽ IBM_Annual_Report_2022.pdf (Page: 6) ▽ IBM_Annual_Report_2022.pdf (Page: 6) ▽ IBM_Annual_Report_2022.pdf (Page: 3) <p>ibm/granite-3-8b-instruct Generated Response:</p> <p>IBM's strategy for sustainability involves making a lasting, positive impact on the world by protecting the environment, advocating for inclusion, and fostering trust and transparency in both technology and business. The company is using IBM Sustainability solutions to simplify and automate their sustainability reporting processes. They have also made significant progress towards their goal of net-zero operational greenhouse gas emissions by 2030, having reduced emissions by 61% since 2010. IBM has been recognized for helping clients turn their sustainability ambitions into actions.</p> <p>Tokens: 803.83 input + 124 output = 927.83 LLM Call Time: 2.08 seconds UI Print Time: 1.62 seconds</p>	<p>Sources:</p> <ul style="list-style-type: none"> ▽ IBM_Annual_Report_2022.pdf (Page: 6) ▽ IBM_Annual_Report_2022.pdf (Page: 6) ▽ IBM_Annual_Report_2022.pdf (Page: 3) <p>meta-llama/llama-3-1-70b-instruct Generated Response:</p> <p>IBM's Strategy for Sustainability IBM's strategy for sustainability is multifaceted and focuses on making a lasting, positive impact on the world. The company aspires to protect the environment, advocate for inclusion, and foster trust and transparency in both technology and business.</p> <p>Key Components: Environmental Impact: IBM aims to achieve net-zero operational greenhouse gas emissions by 2030 and has already reduced those emissions by 61% since 2010. Sustainability Reporting: The company uses IBM Sustainability solutions to simplify and automate its sustainability reporting processes. Client Support: IBM helps clients turn their sustainability ambitions into actions, as recognized by Boston Consulting Group and TIME Magazine. Overall, IBM's strategy for sustainability is centered around creating a positive impact on the environment, society, and business, while also supporting its clients in their sustainability efforts.</p> <p>Tokens: 803.83 input + 172 output = 975.83 LLM Call Time: 13.43 seconds UI Print Time: 12.88 seconds</p>
---	--

Note that the tool identifies where the relevant information was retrieved from the

document.

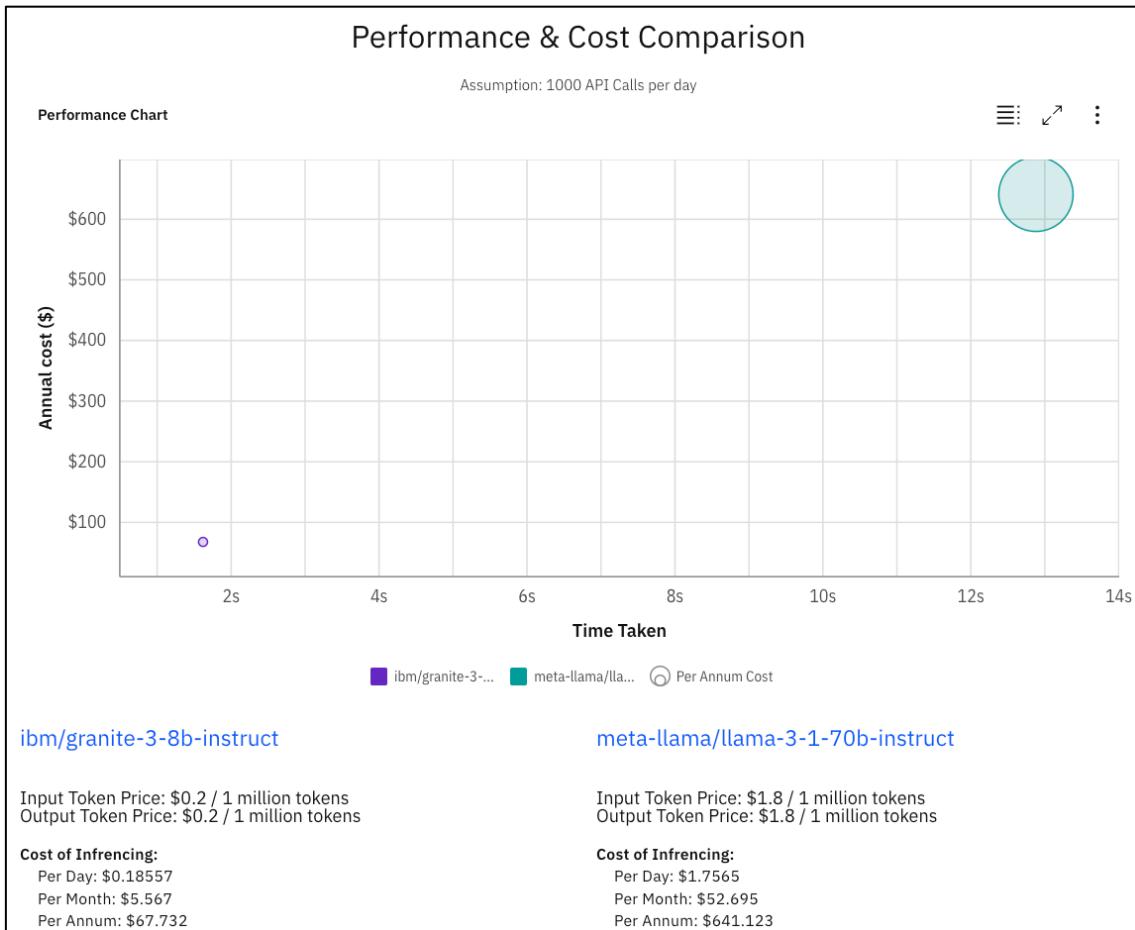
11. Keep in mind that this is a comparison tool. The goal of this exercise is not to show one model is better than the other but to show how **ComFM** can be used to provide useful information for clients to make decisions.

In this particular Question/Answering RAG use case, the output does provide insights about the suitability of the models for this kind of RAG use case.

Note the output:

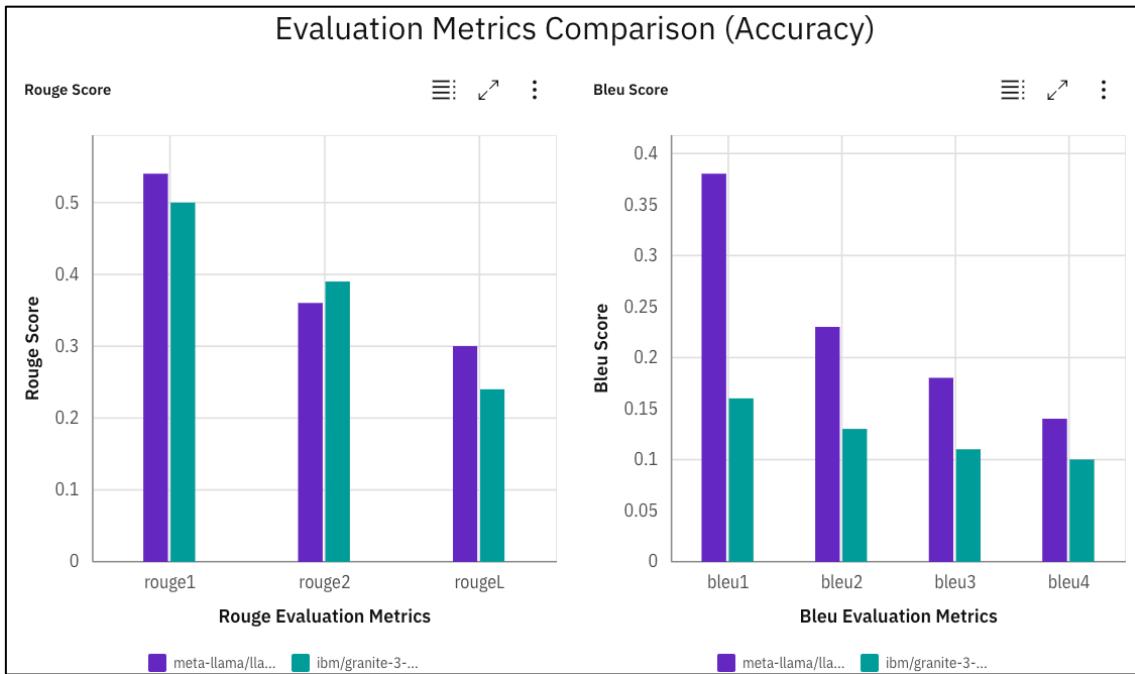
- The granite model correctly identifies “IBM’s strategy for sustainability involves reducing operational greenhouse gas emissions to net-zero by 2030”. It also mentions other items (advocating for inclusion, and fostering trust and transparency in both technology and business) that have more to do with ethical impact instead.
- The llama model provides a similar output. It also highlighted the same ethical impact points as the granite model did. However, the llama model does pick out an important item: that IBM helps its clients “turn their sustainability ambitions into actions”.

Here is the cost comparison:



There is no new revelation as the key metric here is that granite has a USD \$0.2/million input/output tokens cost, and llama-3 has a USD \$1.8/million input/output tokens cost. Llama3 is 9 times more expensive, all things being equal.

Here is the metrics comparison:



Note this for now; you will return to this later. Suffice it to say that llama-3 looks to be more accurate than granite. This can be gleaned from looking at the completion.

12. Change **Min Tokens** to 200 and **Max Tokens** to 400.

13. Click **Generate**.

Enable RAG Enabled

Select Collection from DB

Show Uploaded Docs

Select a question or enter your own in the textarea below

Choose a question below or click this to enter your own question

Prompt

INSTRUCTIONS:
Answer the following question using only information from the documents given above. If there is no good answer in the given documents, say "I don't know"

USER:
What is IBM's strategy for sustainability?

ASSISTANT:

Model Parameters

Decoding Greedy

12

Repetition Penalty

Cost Configuration

API Calls per day - +

13

Increasing the value of **Min tokens** has the effect of making the LLM “say more.” This is evident in the output below:

<p>ibm/granite-3-8b-instruct Generated Response:</p> <p>IBM's strategy for sustainability involves reducing operational greenhouse gas emissions to achieve net-zero by 2030. They have already reduced emissions by 61% since 2010. IBM uses its sustainability solutions to simplify and automate sustainability reporting processes. The company aims to make a lasting, positive impact on the world by protecting the environment, advocating for inclusion, and fostering trust and transparency in both technology and business.</p> <p>IBM's sustainability efforts are not only focused on reducing emissions but also on helping clients turn their sustainability ambitions into actions. This is evident from the recognition they received from Boston Consulting Group and TIME Magazine for their work in this area.</p> <p>In addition, IBM recognizes technology as a source of competitive advantage that can address immediate business challenges and future-proof businesses. This includes issues such as inflation, supply chain disruption, tight labor markets, sustainability requirements, and an ever-evolving cybersecurity threat landscape.</p> <p>Tokens: 803.83 input + 225 output = 1,028.83 LLM Call Time: 3.31 seconds UI Print Time: 2.97 seconds</p>	<p>meta-llama/llama-3-1-70b-instruct Generated Response:</p> <p>IBM's Strategy for Sustainability</p> <p>IBM's strategy for sustainability is multifaceted and focuses on making a lasting, positive impact on the world. The company aims to achieve net-zero operational greenhouse gas emissions by 2030 and has already reduced those emissions by 61% since 2010. IBM is also using its own Sustainability solutions to simplify and automate its sustainability reporting processes.</p> <p>In addition to its internal efforts, IBM is also helping its clients turn their sustainability ambitions into actions. The company was recognized by Boston Consulting Group and TIME Magazine for its work in this area.</p> <p>IBM's approach to sustainability is guided by its aspiration to protect the environment, advocate for inclusion, and foster trust and transparency in both technology and business. The company recognizes the importance of sustainability in addressing the business challenges faced by its clients, including sustainability requirements, and is committed to helping them achieve their sustainability goals.</p> <p>Key Components of IBM's Sustainability Strategy</p> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: left; padding-bottom: 5px;">Component</th><th style="text-align: left; padding-bottom: 5px;">Description</th></tr> </thead> <tbody> <tr> <td style="padding-bottom: 5px;">Net-zero emissions</td><td style="padding-bottom: 5px;">Achieve net-zero operational greenhouse gas emissions by 2030</td></tr> <tr> <td style="padding-bottom: 5px;">Sustainability reporting</td><td style="padding-bottom: 5px;">Use IBM Sustainability solutions to simplify and automate sustainability reporting processes</td></tr> <tr> <td style="padding-bottom: 5px;">Client support</td><td style="padding-bottom: 5px;">Help clients turn their sustainability ambitions into actions</td></tr> <tr> <td style="padding-bottom: 5px;">Environmental protection</td><td style="padding-bottom: 5px;">Protect the environment through sustainable practices</td></tr> <tr> <td style="padding-bottom: 5px;">Inclusion and transparency</td><td style="padding-bottom: 5px;">Advocate for inclusion and foster trust and transparency in technology and business</td></tr> </tbody> </table> <p>Tokens: 803.83 input + 278 output = 1,081.83 LLM Call Time: 30.31 seconds UI Print Time: 14.54 seconds</p>	Component	Description	Net-zero emissions	Achieve net-zero operational greenhouse gas emissions by 2030	Sustainability reporting	Use IBM Sustainability solutions to simplify and automate sustainability reporting processes	Client support	Help clients turn their sustainability ambitions into actions	Environmental protection	Protect the environment through sustainable practices	Inclusion and transparency	Advocate for inclusion and foster trust and transparency in technology and business
Component	Description												
Net-zero emissions	Achieve net-zero operational greenhouse gas emissions by 2030												
Sustainability reporting	Use IBM Sustainability solutions to simplify and automate sustainability reporting processes												
Client support	Help clients turn their sustainability ambitions into actions												
Environmental protection	Protect the environment through sustainable practices												
Inclusion and transparency	Advocate for inclusion and foster trust and transparency in technology and business												

Both models put out more information in their completions. For llama-3, there is no real new information as it simply added a summary table.

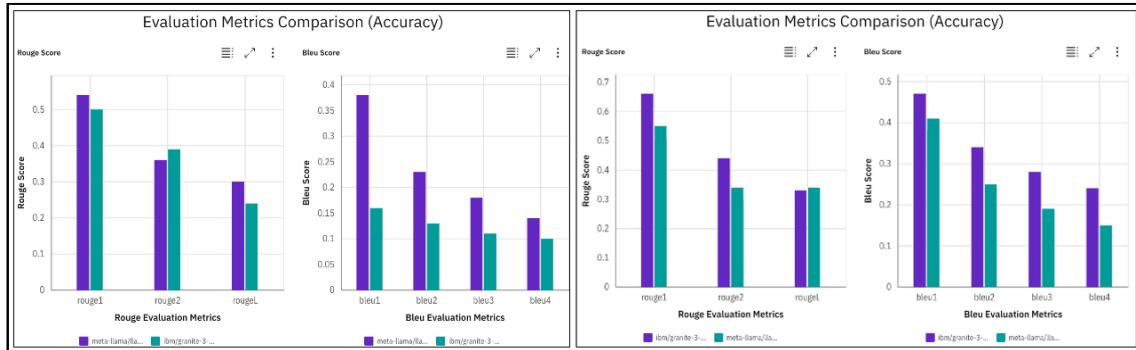
For the granite model, increasing **Min tokens** makes it put out more info, and now it includes information on IBM in helping clients turn their sustainability ambitions into actions.

In this case, granite has improved its output completion when **Max tokens** is increased. Clients should keep this in mind – models can and do behave differently when configuration parameters are changed.

14. Look at the evaluation data. The pictures below compare the case when

- **Min tokens** was 10 (graph on the left). Note that **purple** is for **llama-3-1-70b-instruct**, and **green** is for **granite-3-8b-instruct**. Here, llama 3 performs better.
- **Min tokens** was 200 (graph on the right). Note that **purple** is for **granite-3-8b-instruct**, and **green** is for **llama-3-1-70b-instruct**. The color scheme is reversed. In

this case, granite performs better.



The key observation is that when granite is asked to support a **Min tokens** value of 200, it puts out more info, and it evaluates better than llama 3 (though not by as large a margin).

15. You are encouraged to try different models/use cases/prompts. The key observation is to examine what configuration may provide the best result between models.

Section Summary

- RAG is a very common use case, and clients should take care to find the right model.
- The **ComFM** tool provides a quick and easy way to compare how models behave in an RAG use case.
- It is important to try different configurations to see how each model extracts information from retrieved information. Clients should combine their experience of different models, different configurations, and cost analysis to make their decisions on which model to use.

6.2 RAG use case with expected output

In evaluating models in an RAG use case, you can also provide examples of an expected response. This allows the **ComFM** tool to provide even more meaningful metrics as it can measure the model completions against a “model” answer.

1. If you have not logged out from Section 6.1, ensure you clear the session by clicking on the Refresh icon (↻).

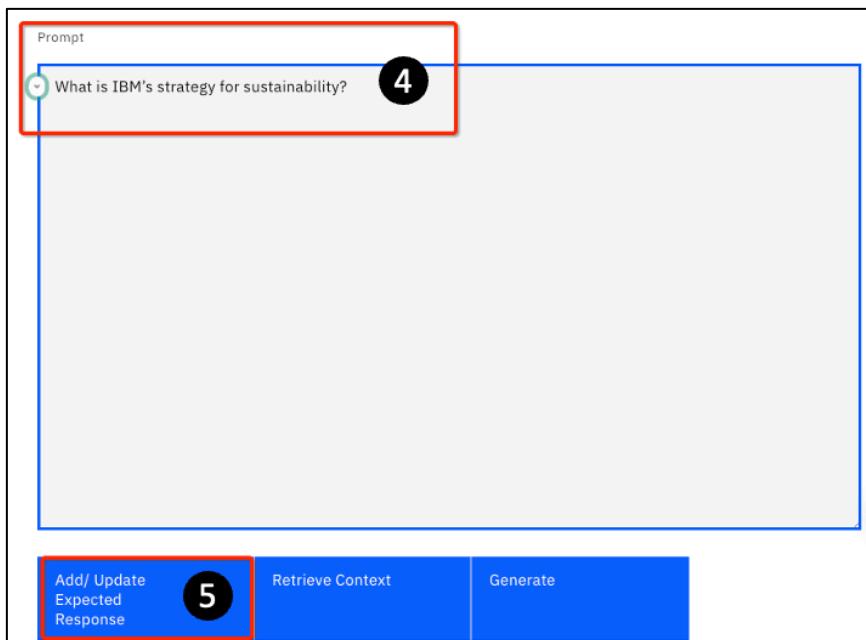
The screenshot shows the 'Compare LLM Foundation Models' interface. At the top, there are tabs for 'Question Answering' (selected) and 'Summarization'. Below these, there are dropdown menus for selecting the first and second LLMs. The first LLM dropdown shows 'Select 1st LLM' with 'watsonx' (radio button selected) and 'external' (radio button unselected). The second LLM dropdown shows 'Select 2nd LLM' with 'llama-3-1-70b-instruct' (radio button selected) and 'external' (radio button unselected). A refresh icon (a blue circle with a white arrow) is located at the top right of the interface.

If you have logged out, log back into the ComFM tool.

2. Select **granite-3-8b-instruct** as the first model and **llama-3-1-70b-instruct** as the second model.
3. Enable RAG and use the **IBM_Annual_Report_2022** as the grounding document.



4. Enter **What is IBM's strategy for sustainability?** as the prompt.
5. Click **Add / Update Expected Response**.

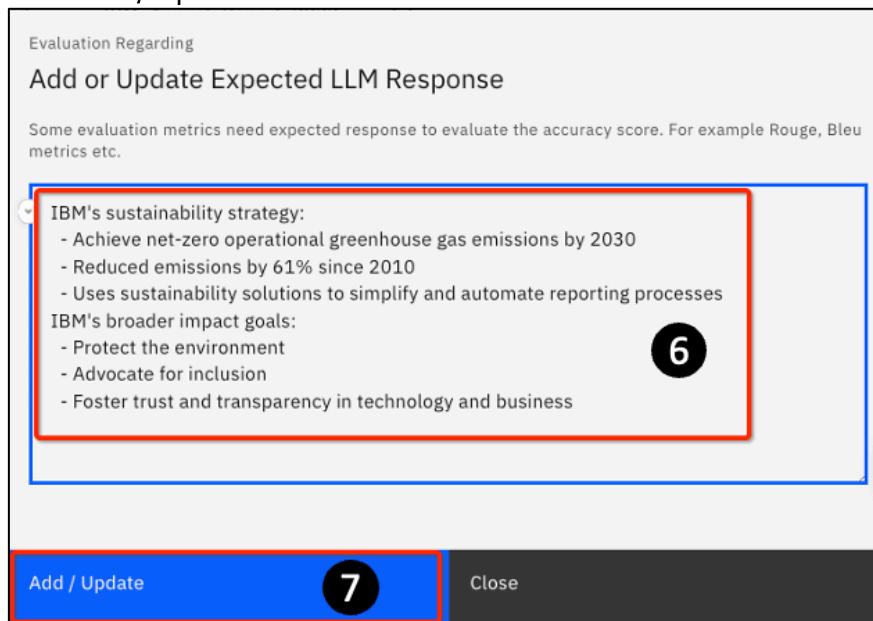


6. The **Add or Update Expected LLM Response** dialog opens. Copy and paste the following text into the text field.

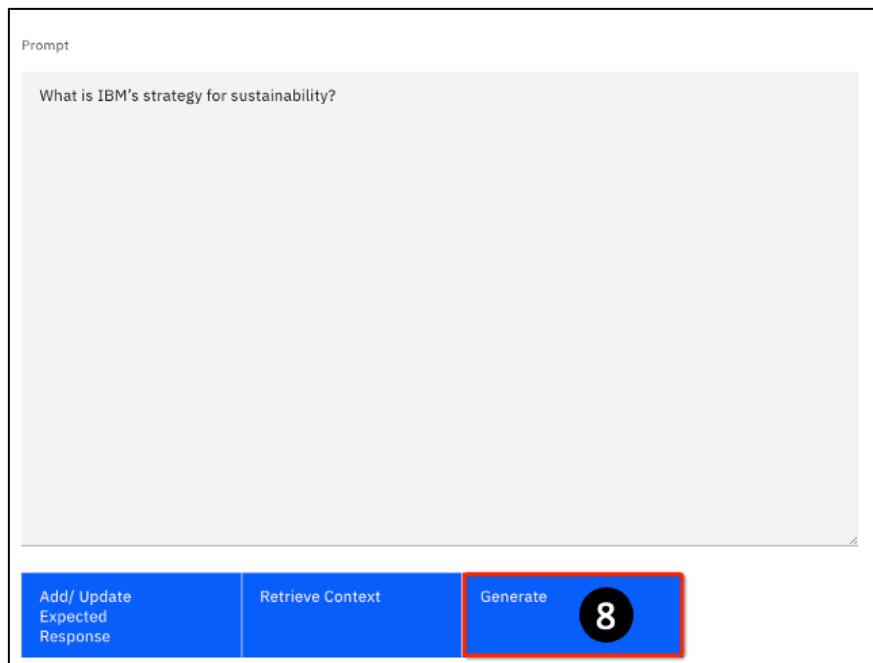
IBM's sustainability strategy:
- Achieve net-zero operational greenhouse gas emissions by 2030
- Reduced emissions by 61% since 2010
- Uses sustainability solutions to simplify and automate reporting processes

IBM's broader impact goals:
- Protect the environment
- Advocate for inclusion
- Foster trust and transparency in technology and business

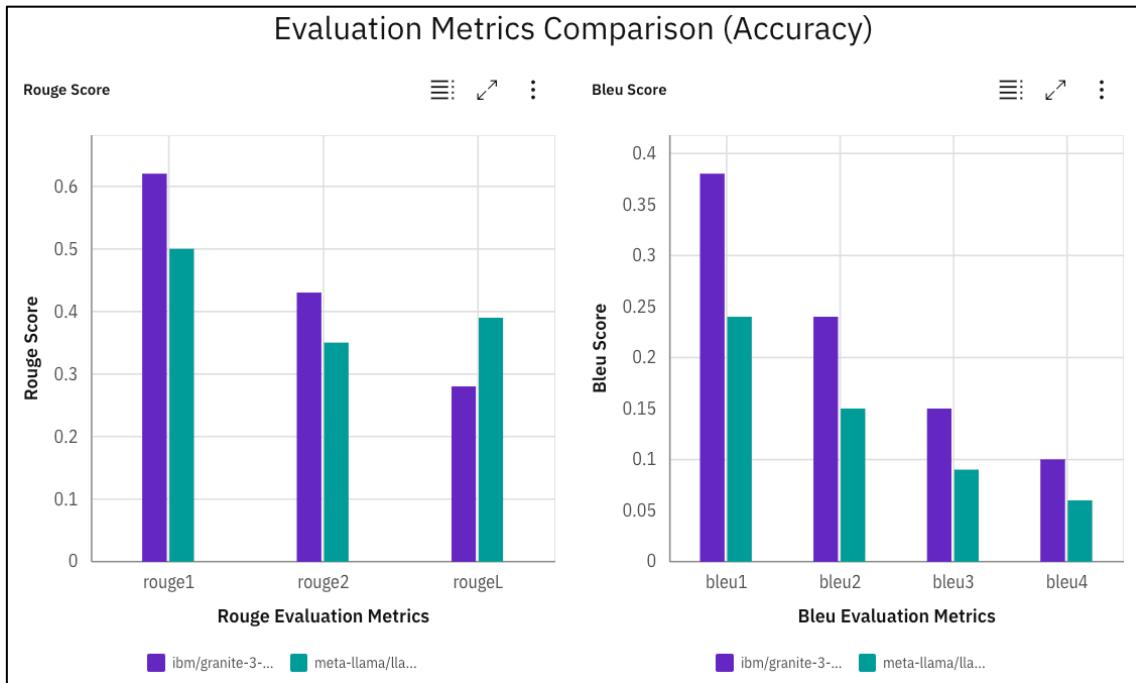
7. Click Add / Update.



8. Click Generate.



9. The important thing is to look at the evaluation metrics. You should see something similar to this:



10. In this test, the granite model output (in purple) is better in almost all cases (the exception being rougeL) than llama 3. This is in contrast to the result from Section 6.1, Step 11. The difference here is that for this run, you have provided an expected output.

The important thing to note is that the expected output:

IBM's sustainability strategy:

- Achieve net-zero operational greenhouse gas emissions by 2030
- Reduced emissions by 61% since 2010
- Uses sustainability solutions to simplify and automate reporting processes

IBM's broader impact goals:

- Protect the environment
- Advocate for inclusion
- Foster trust and transparency in technology and business

is just a simple rewrite (in point form) of the original granite output (re-captured here for

IBM's strategy for sustainability involves making a lasting, positive impact on the world by protecting the environment, advocating for inclusion, and fostering trust and transparency in both technology and business. The company is using IBM Sustainability solutions to simplify and automate their sustainability reporting processes. They have also made significant progress towards their goal of net-zero operational greenhouse gas emissions by 2030, having reduced emissions by 61% since 2010. IBM has been recognized by Boston Consulting Group and TIME Magazine for helping clients turn their sustainability ambitions into actions.

comparison)

It is, therefore, natural to expect that under this setup, granite will evaluate better than llama-3.

Section Summary

- In this section, you provided an expected outcome for a prompt question in an RAG use case.
- The focus of this section is not to demonstrate that one can help the model evaluate to a better metric. Instead, it shows that when evaluating and comparing models, it is useful to also provide expected output to see which model can best provide the right answers.
- Finding the proper model, even in a RAG use case, does take some testing. However, clients should try out different configurations different models, and provide expected answers to prompts.
- Another important factor is the cost of running a particular type of query. The **ComFM** tool can be used to test different queries/model combinations so the client can find the best combination.

7. Compare with Open AI models.

The **ComFM** tool can also be used to compare with models outside of watsonx.ai. This section provides the steps to compare with OpenAI models.

There are two pre-requisite steps:

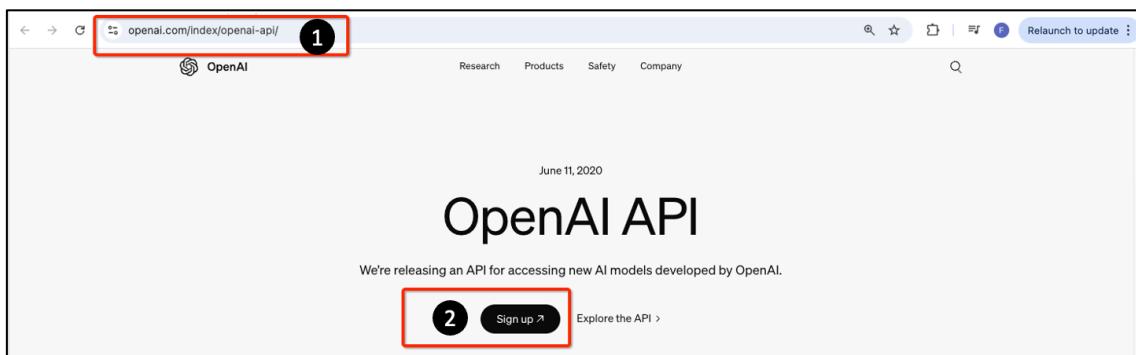
- Create an account on OpenAI. To run inference on models like gpt4, you will need to charge up the account. Or you can skip this part of the lab.
- Create an OpenAI API key.

Note: To run this part of the lab requires a charged account on OpenAI.

7.1 Create an OpenAI account

If you already have an OpenAI account, you can skip this section.

1. Go to <https://openai.com/index/openai-api/>
2. Click **Sign up**.



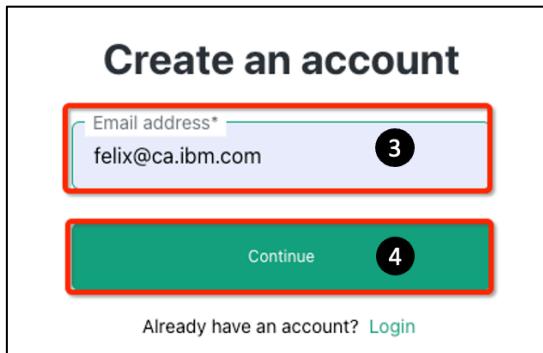
3. Provide your email address.
4. Click **Continue**.

Create an account

Email address* **3**

Continue **4**

Already have an account? [Login](#)



5. Provide a suitable password under the guideline (at least 12 characters).
6. Click **Continue**.

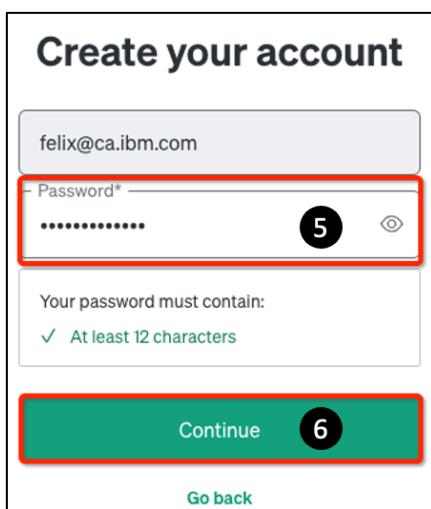
Create your account

- Password* **5** 

Your password must contain:
✓ At least 12 characters

Continue **6**

[Go back](#)

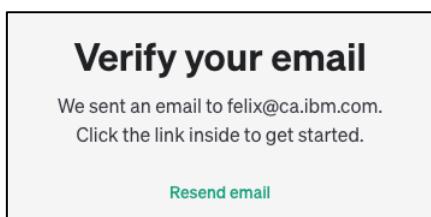


7. You will get this message:

Verify your email

We sent an email to felix@ca.ibm.com.
Click the link inside to get started.

[Resend email](#)



Log into the email account you provided.

8. You should see an email from OpenAI asking you to verify your email address. Click **Verify email address**.

Verify your email address

To continue setting up your OpenAI account, please verify that this is your email address.

[Verify email address](#)

This link will expire after 3 hours. If you did not make this request, please disregard this email. For help, contact us through our [Help center](#).

9. The Tell us about you from OpenAI opens. Fill in the required details.

10. Click Agree.

Tell us about you

Full name 9

Organization name (optional)

Birthday

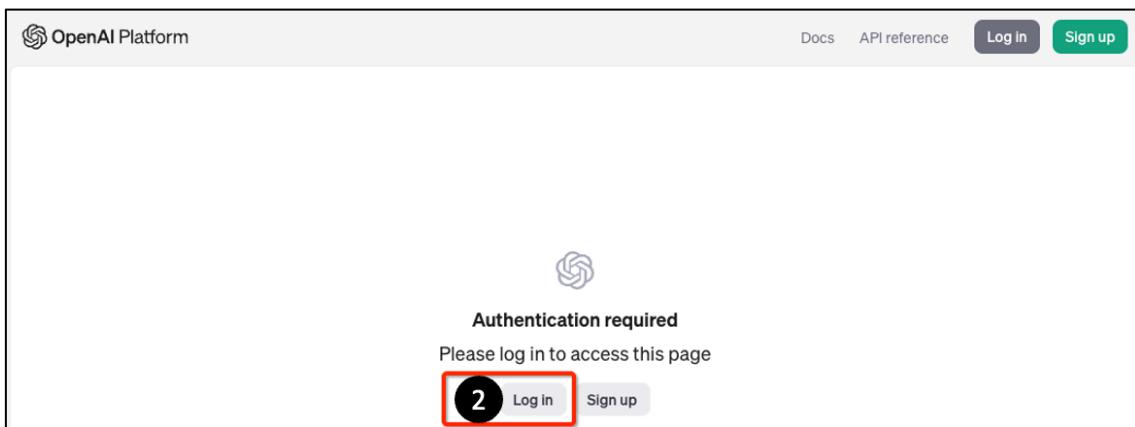
By clicking "Agree", you agree to our [Terms](#) and have read our [Privacy Policy](#).

[Agree](#) 10

11. You will be taken to the **OpenAI developer platform** landing page.

7.2 Getting an OpenAI API

1. Go to <https://platform.openai.com/api-keys>
2. The OpenAI Platform page opens. Click Log in.



3. Enter the email address you used to sign up, then click **Continue**.

Welcome back

Email address*

Continue 3

4. Provide your password and click **Continue**.

felix@ca.ibm.com [Edit](#)

Password* [Edit](#)

[Forgot password?](#)

Continue 4

If this is the first time you are logging in after creating the account, proceed to the next step. Otherwise, proceed to Step 9.

5. The **API keys** open. OpenAI will use your phone for verification. Click **Start verification**.

P Personal / Default project

Playground Dashboard Docs API reference

DASHBOARD API keys + Create new secret key

Chat Completions Assistants Batches

Verify your phone number to create an API key

5 Start verification

6. The **Verify your phone number** page opens. Fill in your number with the appropriate country code.
7. Click **Send code**.

Verify your phone number

CA +1

Send code

8. The **Enter code** page appears. You should have received a 6-digit code from OpenAI. Enter the code in the field.

Enter code [Go back](#)

Please enter the code we just sent you.

000 000

Resend code

You may see a warning. You can close that window.

9. If this is not the first time you log in, OpenAI will send a verification code to your registered email to provide you with a login code. Enter the code in the **Code** field and click **Continue**.

Check your inbox

Enter the verification code we just sent to felix@ca.ibm.com.

Code

Continue

Resend email

Once logged in, click **Create new secret key**.

Playground Dashboard Docs API reference

API keys

+ Create new secret key

As an owner of this project, you can view and manage all API keys in this project.

10. The **Create new secret key** page opens. Provide a **Name** of your choice. In this example, **CFM_API** is used.

11. Leave all other defaults. Click **Create secret key**.

Go back

Create new secret key

Owned by

You Service account

This API key is tied to your user and can make requests against the selected project. If you are removed from the organization or project, this key will be disabled.

Name Optional **CFM_API** 10

Project Default project

Permissions All Restricted Read Only 11

Cancel Create secret key 12

12. The **Save your key** page appears. You MUST copy and save the key somewhere, as you will not be able to view this again (the key is intentionally blotted out below). Click **Copy** and save the key somewhere you can retrieve later.

13. Click **Done**.

Save your key

Please save this secret key somewhere safe and accessible. For security reasons, **you won't be able to view it again** through your OpenAI account. If you lose this secret key, you'll need to generate a new one.

12

Permissions
Read and write API resources

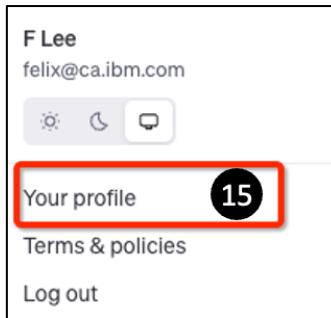
13 Done

After you sign up for OpenAI, you get access to the **gpt-3.5-turbo** model. If you wish to use other models, such as **gpt-4**, there is a charge required. This exercise is done with **gpt-3.5-turbo**. The following steps show how you can provide payment and make other models available to you. You DO NOT need to do these to complete this lab.

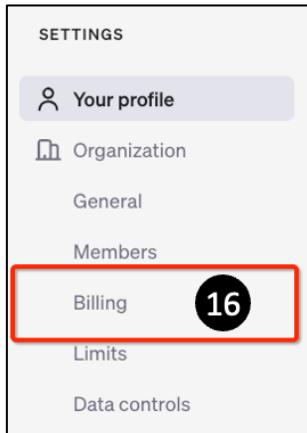
14. On the OpenAI console, click your ID icon.



15. Click Your profile.



16. Click Billing under SETTINGS.

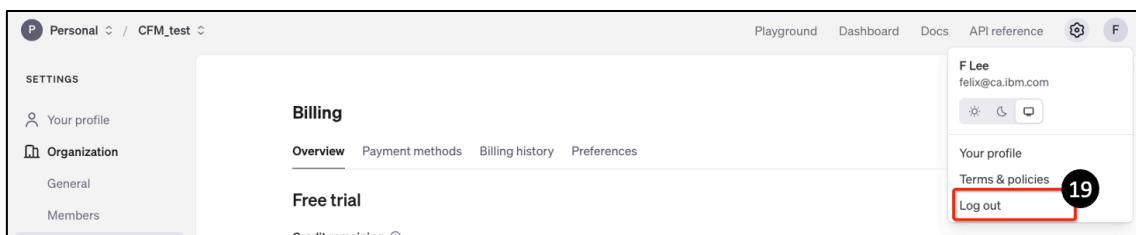


17. Follow the onscreen instructions to charge up your account. There is a minimum charge of \$5.00 from OpenAI to use **gpt-4**. A balance of \$10.00 is more than enough for testing.

18. Click your ID icon.

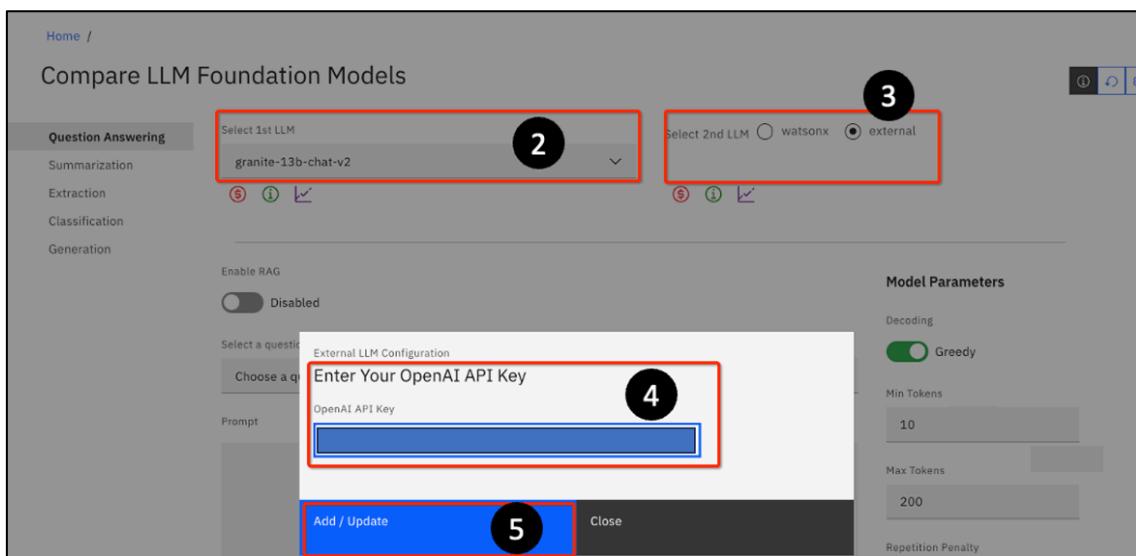


19. Click Log out.



7.3 Comparing with other models using the OpenAI API

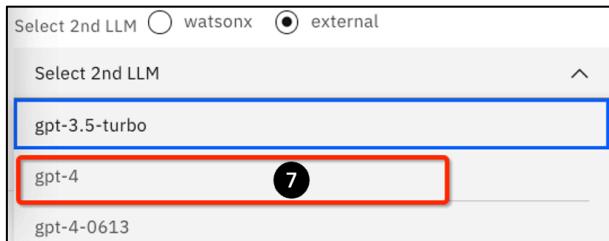
1. Go back to the **ComFM** tool.
2. Use **granite-13b-chat-v2** as **1st LLM**.
3. For **2nd LLM**, select **external**.
4. The **Enter Your OpenAI API Key** page opens. Enter your OpenAPI key (from Step 12 of Section 5.4.2). It is intentionally blurred out below.
5. Click **Add/Update**.



6. On the **ComFM** landing page, click to expand the list for **Select 2nd LLM**.



7. You will see a list of different gpt models. Select **gpt-4**.



8. You can now compare the **granite-13b-chat-v2** (or you can try **granite-3-8b-instruct**) against **gpt-4** (or other models if available to you).
9. Enable **RAG**, the **IBM Annual Report 2022** is used as the default for grounding document.
10. Select any questions from the list. In the example below, it is this query: **Can you provide details on the \$2 billion spent on acquisitions in 2022? What strategic objectives do these acquisitions serve?**

11. Click **Generate**.

Your output should look like this:

ibm/granite-13b-chat-v2

Generated Response:

Acquisitions in 2022 led to a total expenditure of \$2 billion. These acquisitions aimed to bolster the company's capabilities in several strategic areas, including artificial intelligence, cloud computing, and cybersecurity. By acquiring specific companies specializing in these fields, the organization aimed to enhance its technological infrastructure, improve operational efficiency, and expand its product offerings.

In summary, the \$2 billion spent on acquisitions in 2022 was directed towards strategic objectives such as strengthening the company's technological foundation, enhancing operational efficiency, and expanding its product portfolio.

I cannot answer that question based on the provided document.

Tokens: 521.46 input + 124 output = 645.46
LLM Call Time: 4.68 seconds
UI Print Time: 4.25 seconds

gpt-4

Generated Response:

In 2022, the company spent \$2,650 million on eight acquisitions. The acquisitions were made across different segments including Software and Consulting. The acquired businesses range from data and analytics software providers to digital product engineering services firms. The strategic objectives of these acquisitions are to enhance the company's portfolio of products and services capabilities and to further advance IBM's hybrid cloud and AI strategy. Here are the details of the acquisitions:

- Envizi: A data and analytics software provider for environmental performance management.
- Sentaca: A telco consulting services and solutions provider specializing in automation, cloud migration, and future networks for telecommunications providers.
- Neudesic: An application development and cloud computing services company.
- Randori: A leading attack surface management (ASM) and cybersecurity provider.
- Databand.ai: A proactive data observability platform that isolates data errors and issues to alert relevant stakeholders.
- Omnio: A developer of software connectors used in the collection

Tokens: 2,822 input + 200 output = 3,022
LLM Call Time: 10.84 seconds
UI Print Time: 12.51 seconds

Performance & Cost Comparison

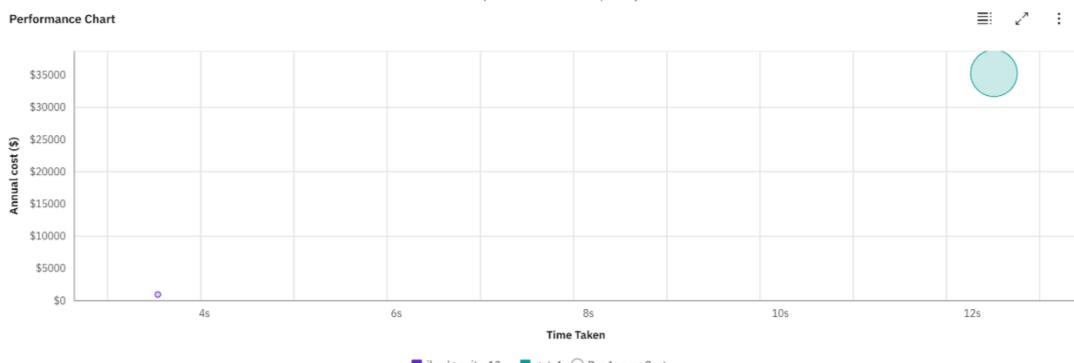
Assumption: 1000 API Calls per day

Each model generates some good content. The focus here is on other information. Note that the granite-13b-chat-v2 is a 13 billion parameters model, whereas the gpt-4 is a 1.7 trillion parameters model. The granite model completed the generation in about 4.7 seconds, while it took 12.5 seconds for gpt – almost 3 times more. Your numbers will differ but should be in a similar range.

Scrolling down, you should see something similar to this:

Performance & Cost Comparison

Assumption: 1000 API Calls per day



ibm/granite-13b-chat-v2

Input Token Price: \$0.6 / 1 million tokens
Output Token Price: \$0.6 / 1 million tokens

Cost of Infringement:
Per Day: \$2,614.53
Per Month: \$78,435.8
Per Annum: \$954,302

gpt-4

Input Token Price: \$30 / 1 million tokens
Output Token Price: \$60 / 1 million tokens

Cost of Infringement:
Per Day: \$96.66
Per Month: \$2,899.8
Per Annum: \$35,280.9

The key comparison here is the **Cost of Inference**. The **gpt-4** model is projected to cost \$35,000 annually if the client runs 1000 similar APIs per month, whereas for **granite-13-b-chat-v2**, it is less than one thousand dollars.

Section Summary

- You can use the **ComFM** tool to compare with some OpenAI models, such as gpt-4.
- The gpt-4 model (being a 1.7 trillion parameters model) typically takes longer to respond with a much higher cost than models available from watsonx.ai.
- This does require a paid account. Given the steep price of using gpt-4 for general inference, it may be something a client should do to ensure that the cost and results combine to be acceptable to their business.
- Clients should also consider tuning a watsonx.ai model, which can provide focused responses to specific business use cases at a much lower cost.

8. Summary

Clients need to try out and compare models before deploying and tuning them for production. There are many different and important aspects:

- Performance needs to be balanced with cost. Sometimes, this is not obvious when clients are just running/tuning a few queries in a Proof of Experience (PoX). It is useful to get estimates when the usage is scaled out to the proper size over a longer period.
- It is important to see how different models provide completions by providing model answers and looking at evaluation metrics. This is important if clients are looking for the ability to follow instructions and provide outputs in specific ways.
- RAG is a very important use case. Clients should test using sample documents and compare how different models can glean information from them. In particular, clients can try different text splitting, embedding methods, and other configurations, as some methods may work better with certain types/formats of documentation.
- The **ComFM** tool provides an efficient and quick way to compare two models in different use cases with RAG testing capabilities.