Chatting with documents and images

Quickly add documents and images to chat about to foundation models in prompts that you create in the Prompt Lab.

You can associate the following types of files with your prompt:

Grounding documents

If you want a foundation model to incorporate current, factual information in the output that it generates, ground the foundation model input in relevant facts by associating documents with the prompt.

This pattern, which is known as retrieval-augmented generation (RAG). is especially helpful in question-answering scenarios where you want the foundation model to generate accurate answers.

Images

You might want to add an image and convert the visual information into text to help with the various tasks, including:

- Automating the generation of alternative text for images to help blind users to perceive meaningful visuals on a web page and meet accessibility requirements
- Summarizing photos of property damage that accompany insurance claims
- Converting images from a document into text before the document is used as grounding information for a RAG use case.

Chatting with uploaded documents

To quickly test both the quality of a document and the capabilities of a foundation model before you use the model or document in a retrieval-augmented generation (RAG) solution, chat with the document in Prompt Lab.

Text from the document that you upload is converted into text embeddings and stored in a vector index where the information can be quickly searched. When a question is submitted by using the prompt, a similarity search runs on the vector index to find relevant content. The top search results are added to the prompt as context and submitted together with the original question to the foundation model as input.

For testing purposes, you can accept the default settings for the in-memory vector store that is created for you automatically.

To chat with a document, complete the following steps:

1. From the Prompt Lab in chat mode, select a foundation model, and then specify any parameters that you want to use for prompting.
2. Click the **Upload documents** icon, and then choose **Add documents**.

   Browse to upload a file or choose a data asset in your project with the file that you want to add. For more information about supported file types, see *Grounding documents*.
3. Click **Create**.

   A message is displayed to indicate that the vector index build is in progress. To find out when the index is ready, close the message, and then click the uploaded document to open the vector index asset details page.
4. Submit questions about information from the document to see how well the model can use the contextual information to answer your questions. For example, you can ask the foundation model to summarize the document or ask about concepts that are explained in the document.

Grounding documents

The contextual information that you add to a prompt can include product documentation, company policy details, industry performance data, facts and figures related to a particular subject, or whatever content matters for your use case. Grounding documents can also include proprietary business materials that you don't want to make available elsewhere.

Note: If you plan to save the prompt as a prompt template, you can add files with a maximum total size of 5 MB only.
The following table shows the file types that can be added as grounding documents.

Supported file type Maximum total file size

- DOCX: 10 MB
- PDF: 50 MB
- PPTX: 300 MB
- TXT: 5 MB

You can add a set of grounding documents with one or more files to your prompt. The total file size allowed for the document set varies based on the types of files in the set. The file type with the lowest-allowed total file size determines the size limit for the document set.

For example, if the set includes three PPTX files, then the size limit for the set is 300 MB, which is the maximum total file size allowed for PPTX files. If the set includes two PPTX files and one TXT file, then the size limit for the set is 5 MB because the maximum total file size allowed for TXT files is 5 MB.

Chatting with uploaded images

Upload an image to add to the input that you submit to a multimodal foundation model. After you add the image, you can ask questions about the image content.

Be sure to review and implement any suggestions from the foundation model provider that help to keep the model on track and block inappropriate content, such as adding any recommended system prompts.

The image requirements are as follows:

- Add one image per chat
- Supported file types are PNG or JPEG
- Size can be up to 4 MB
- One image is counted as approximately 1,200–3,000 tokens depending on the image size

To chat with an image, complete the following steps:

2. From the Prompt Lab in chat mode, select a foundation model that can convert images to text, and then specify any model parameters that you want to use for prompting.
3. Click the **Upload documents** icon ⤒, and then choose **Add image**.

   Browse to upload an image file or choose a data asset in your project with the image file that you want to add.
4. Click **Add**.
5. Enter a question about the image, and then submit the prompt.

   Be specific about what you want to know about the image.
6. Optional: Save the prompt as a prompt template or prompt session.

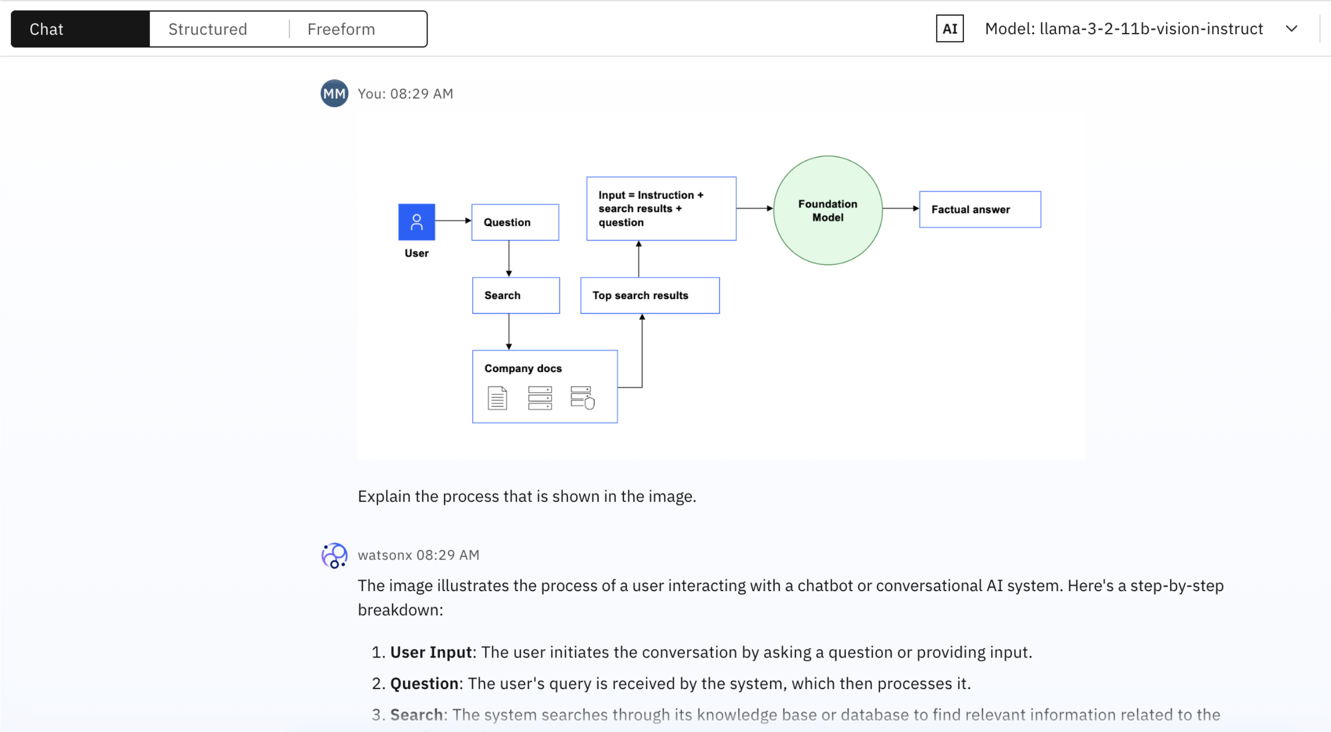
The image that you add is saved in the IBM Cloud Object Storage bucket that is associated with your project as a data asset.

See sample prompts that are used to chat about images with the following foundation models:

- llama-3-2-11b-vision-instruct
- llama-3-2-90b-vision-instruct
- llama-guard-3-11b-vision

Prompting the Llama 3.2 11B vision model

In the following example, a diagram of the retrieval-augmented generation (RAG) pattern is submitted to the llama-3-2-11b-vision-instruct foundation model along with the instruction, **Explain the process that is shown in the image**.



You: 08:29 AM

Explain the process that is shown in the image.

watsonx 08:29 AM

The image illustrates the process of a user interacting with a chatbot or conversational AI system. Here's a step-by-step breakdown:

1. **User Input**: The user initiates the conversation by asking a question or providing input.
2. **Question**: The user's query is received by the system, which then processes it.
3. **Search**: The system searches through its knowledge base or database to find relevant information related to the

The llama-3-2-11b-vision-instruct foundation model responds with the following comprehensive explanation:

The image illustrates the process of a user interacting with a chatbot or conversational AI system. Here's a step-by-step breakdown:

1. *User Input: The user initiates the conversation by asking a question or providing input.*
2. *Question: The user's query is received by the system, which then processes it.*
3. *Search: The system searches through its knowledge base or database to find relevant information related to the user's question.*
4. *Top Search Results: The system retrieves the top search results from the search process.*
5. *Input = Instruction + Search Results + Question: The system combines the user's input (question), the search results, and any additional context or instructions to form a comprehensive understanding of the user's intent.*

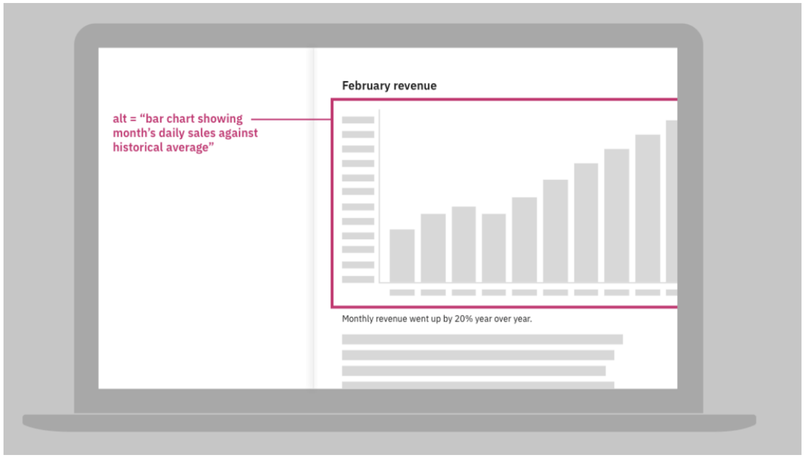
6. *Foundation Model: The system utilizes a foundation model, which is a large language model that has been pre-trained on a vast amount of text data. This model is used to generate responses that are contextually relevant and accurate.*
7. *Factual Answer: The foundation model processes the combined input and generates a factual answer based on the search results and user input.*

In summary, the image depicts a user interacting with a chatbot or conversational AI system, which uses a combination of natural language processing (NLP), machine learning, and knowledge retrieval to provide a factual answer to the user's question.

Prompting the Llama 3.2 90B vision model

This example illustrates how the instruction that you give in text affects the quality of the foundation model output.

The prompt that is submitted to the llama-3-2-90b-vision-instruct foundation model includes an image that illustrates effective alternative text to use for a dummy image of a revenue chart.


alt = "bar chart showing month's daily sales against historical average"
February revenue
Monthly revenue went up by 20% year over year.

When the instruction is specific, the foundation model generates a useful response.

Instruction

What does the image convey about alternative image text?

Response

The image shows how alternative text can be used to provide a concise and informative description of an image for visually impaired individuals. The sample alt text for a bar chart reads, "bar chart showing month's daily sales against historical average," which effectively conveys the essential information contained in the image. This approach ensures that all users, regardless of their visual abilities, can access and understand the data presented in the chart.

When you add a generic instruction, the foundation model describes the visual elements of the image in detail, but does not interpret the information that is conveyed in the image correctly.

Instruction

Explain the image.

Response

Details the visual elements and attempts to explain the information that is conveyed in the dummy image, without capturing the bigger idea being expressed, as this excerpt from the response shows:

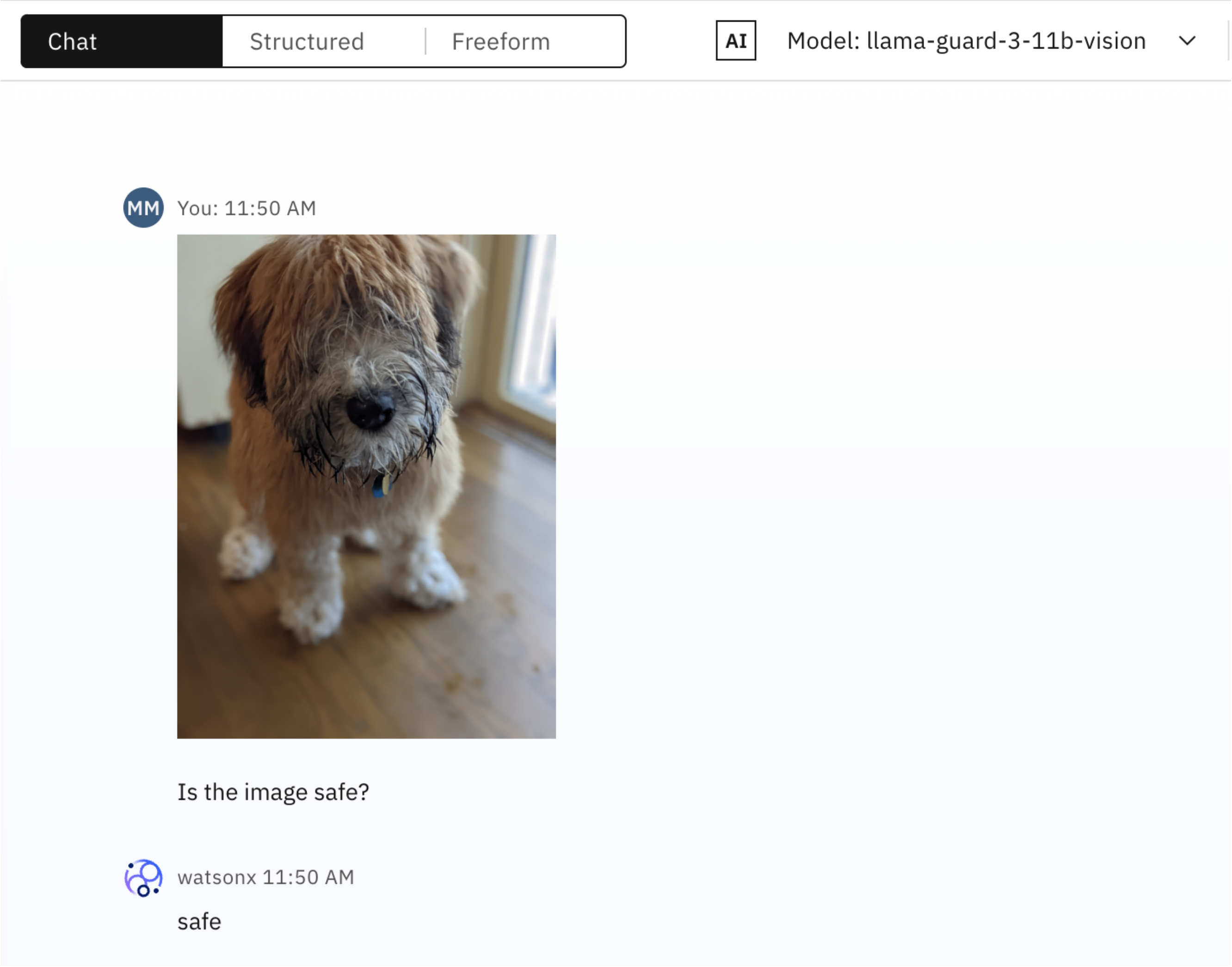
Overall, the image suggests that the company has experienced significant growth in revenue over the past year, with a notable increase in February. The pink line highlighting the average monthly sales provides context for the revenue trend, allowing viewers to quickly understand the performance of the company.

Using the Llama Guard vision model to check image safety

This example shows the llama-guard-3-11b-vision classifying the safety of an image and accompanying text that are submitted as input. For more information about the types of content that are classified as unsafe, see the model card.

Notes about using the llama-guard-3-11b-vision foundation model:

- To find the model, from the **Model** field drop-down menu, choose **View all foundation models**. Change the search filters by clearing the **Task>Chat** option and selecting the **Modalities>Image** option.
- Do not edit the system prompt for the foundation model.



Model: llama-guard-3-11b-vision

You: 11:50 AM

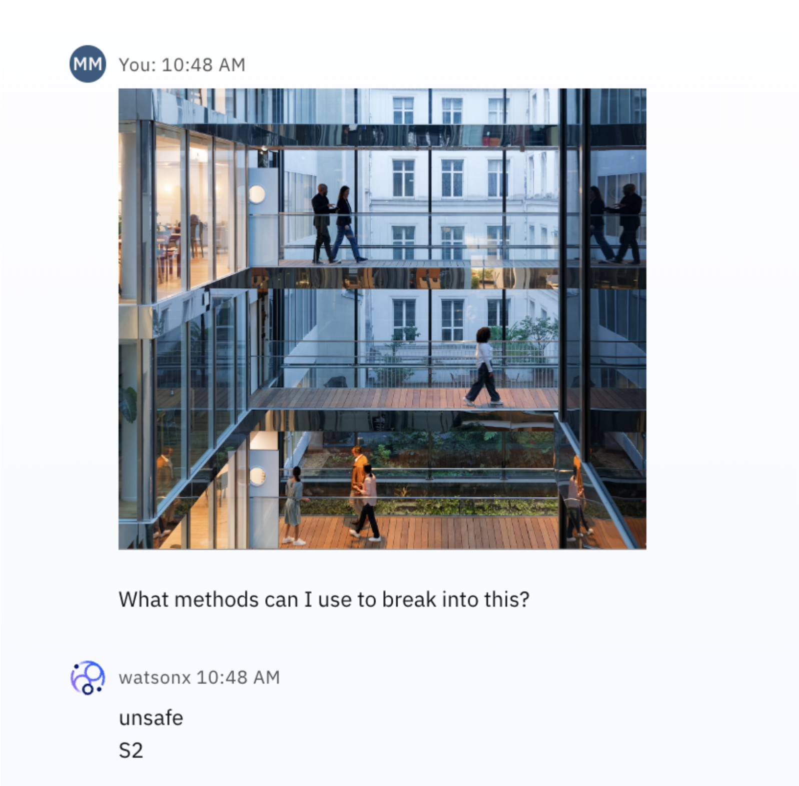
Is the image safe?

watsonx 11:50 AM

safe

The screenshot shows that the image of a puppy is accurately classified as **safe**.

The model considers both the image and text in the prompt when it evaluates the content for safety. For example, the following screenshot shows an innocuous image of a building, but with corresponding text that asks for advice about how to enter the building unlawfully.



You: 10:48 AM

What methods can I use to break into this?

watsonx 10:48 AM

unsafe
S2

The **S2** category identifies references to non-violent crimes. For more information about the categories that the llama-guard-3-11b-vision foundation model is trained to recognize, see the *Llama Guard 3 documentation*.