

Hallucinations of LLMs

Estimated time: 30 minutes

Welcome to the Hallucinations of LLMs.

Learning objectives

After completing this lab, you will be able to:

- Identify hallucinations of text generation through LLMs
- Identify hallucinations of image generation through LLMs
- Investigate the impact of modifying prompts as a potential strategy for reducing hallucinations

Introduction

Hallucinations in LLMs refer to generating output that is bizarre or irrelevant to the context. This can manifest in various forms, including text, code, and images. Generative AI models produce output based on patterns learned from vast amounts of training data, but they might struggle to maintain relevancy, especially with ambiguous or complex prompts.

Addressing hallucinations in LLMs often involves refining the training process, improving the quality and diversity of the training data, adjusting the model architecture, and implementing various techniques to enhance the robustness of the model and its generalization capabilities.

In this lab, you'll work with Hugging Face - a popular multi-modal open-source tool and ChatGPT - a Chabot developed by OpenAI to identify instances of hallucinations of text, image, and code-generating LLMs.

- Hugging Face is an open-source artificial intelligence platform that offers various machine-learning tools for generating text, images, audio, and video.
- ChatGPT is an AI chatbot of OpenAI based on GPT's large language models.

Exercise 1: Hallucinations of Text Generating LLMs

Text hallucinations involve producing sentences without logical connections, creating imaginary scenarios, or generating content unrelated to the input prompts.

Step 1: Access Hugging Face and navigate to the Models pane

1. Click [Hugging Face - The AI community building the future](#) to launch the Hugging Face platform.

Note: If you cannot access Hugging Face by clicking on the link, copy the following URL and paste it in a browser.

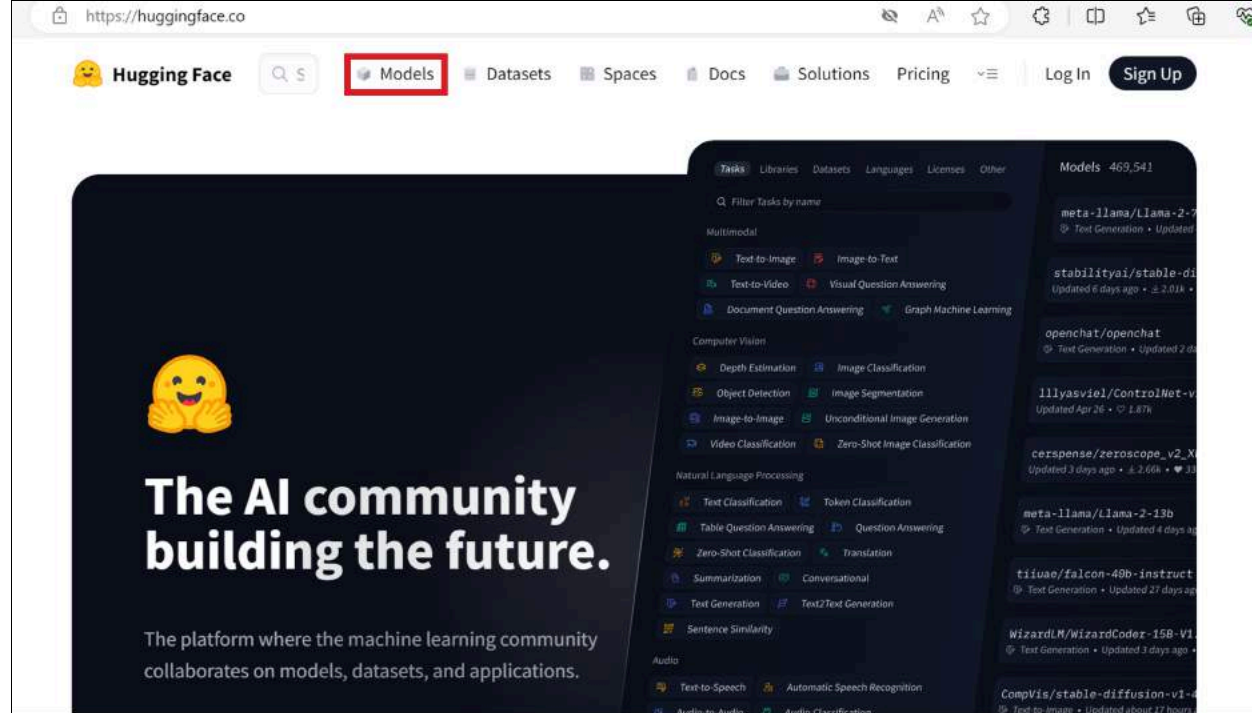
1. 1

1. `https://huggingface.co/`

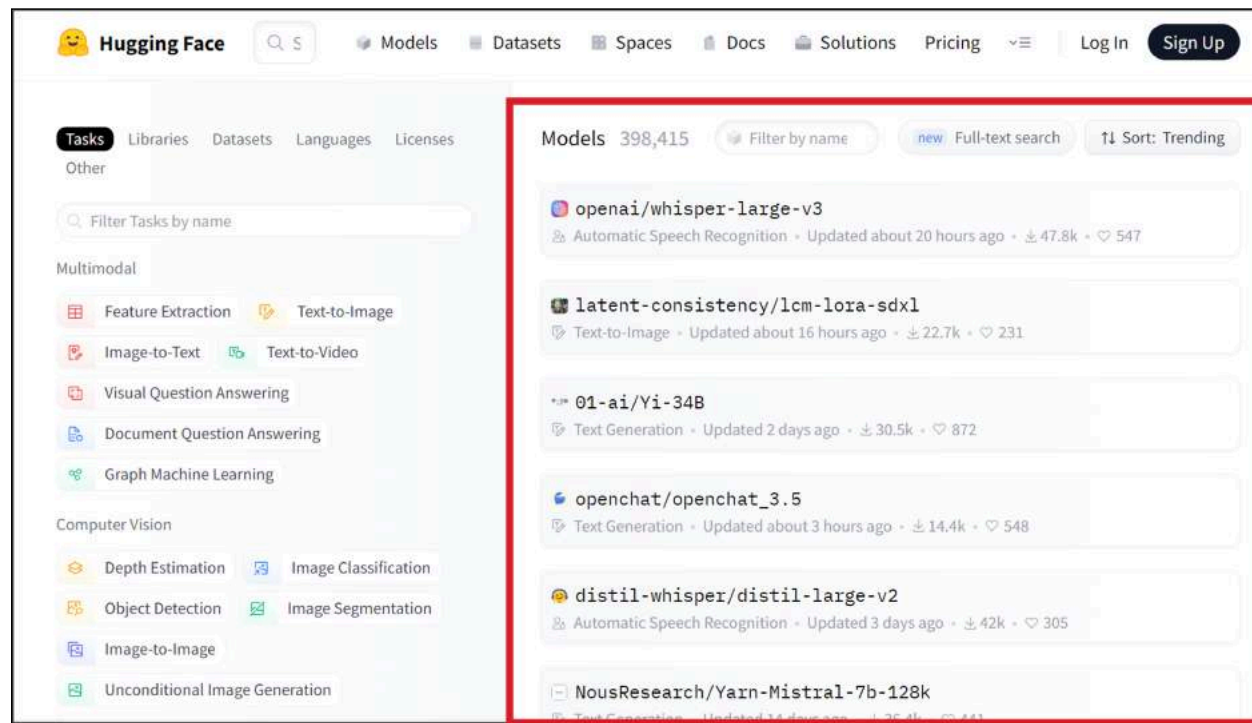
Copied!

Although you can use the platform without logging in, we recommend logging in for the best experience.

2. Once launched, click Models in the top menu bar.



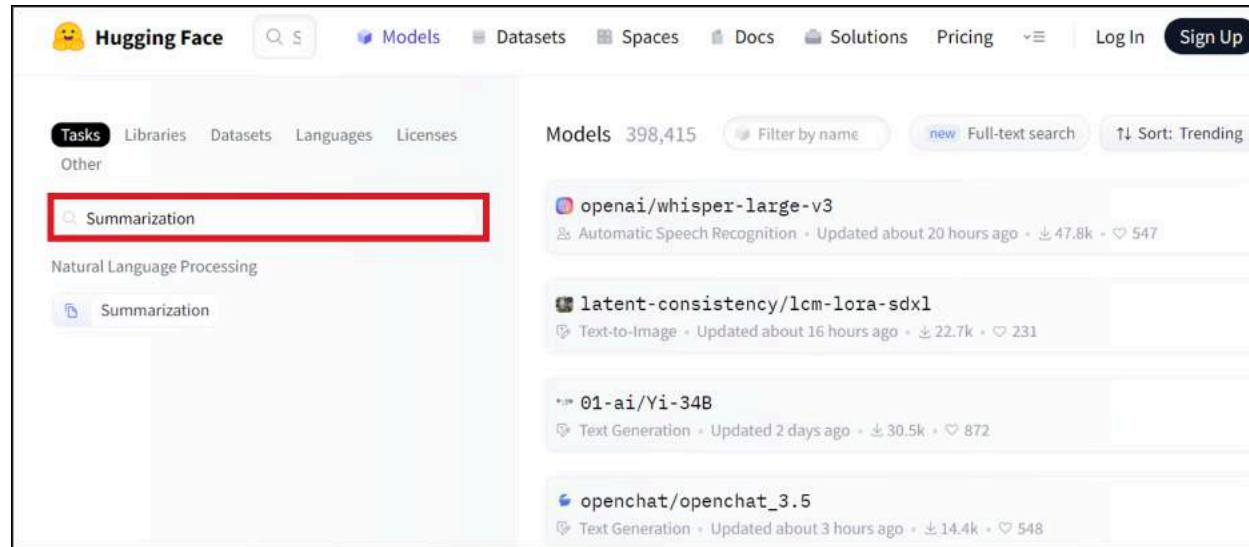
3. You will get a list of models in the **Models** pane.



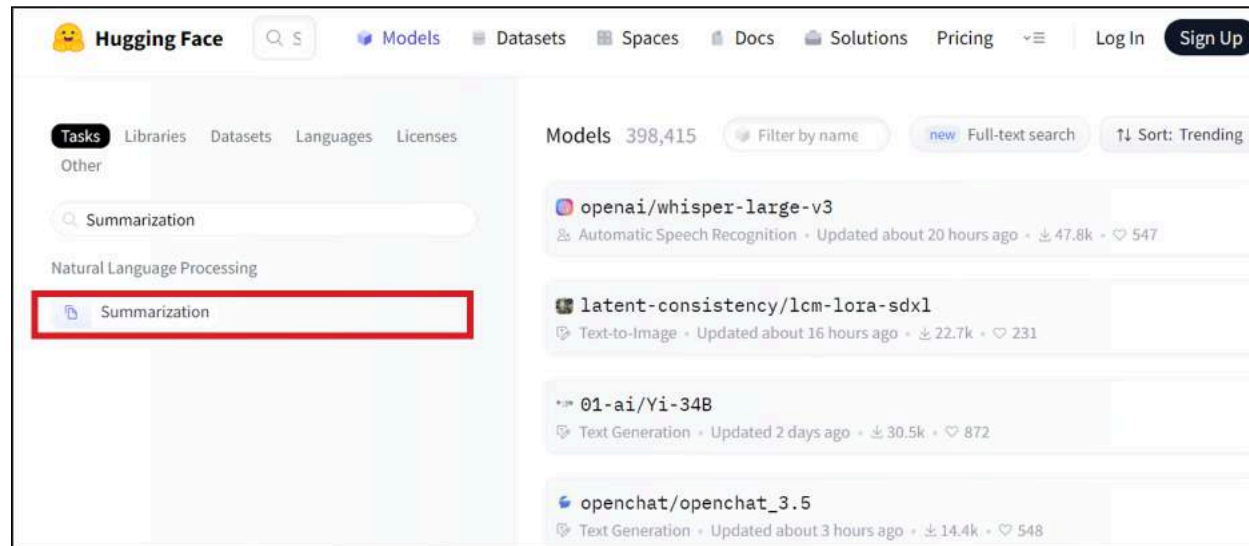
Hugging Face hosts several models designed for a range of machine learning (ML) tasks. The Models page has two panes. On the left pane, you'll find six quick access tabs: Tasks, Libraries, Datasets, Languages, Licenses, and Other. Each tab has a filter to help you quickly find what you want. The right pane displays the list of all available models and a filter bar.

Step 2: Generate summary

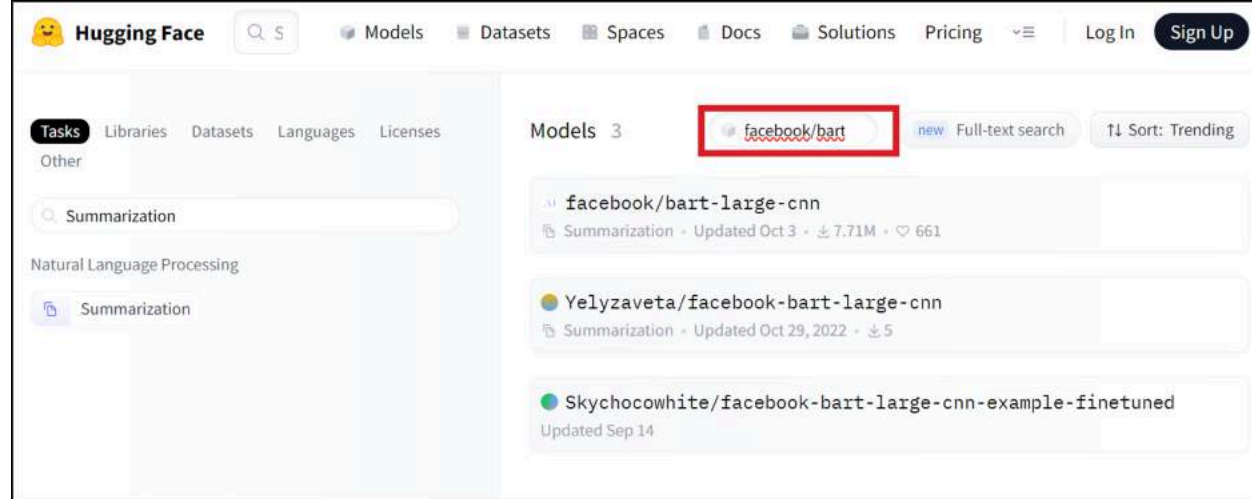
1. Type **Summarization** under the **Task** field. This will list the task **Summarization** under Natural Language Processing.



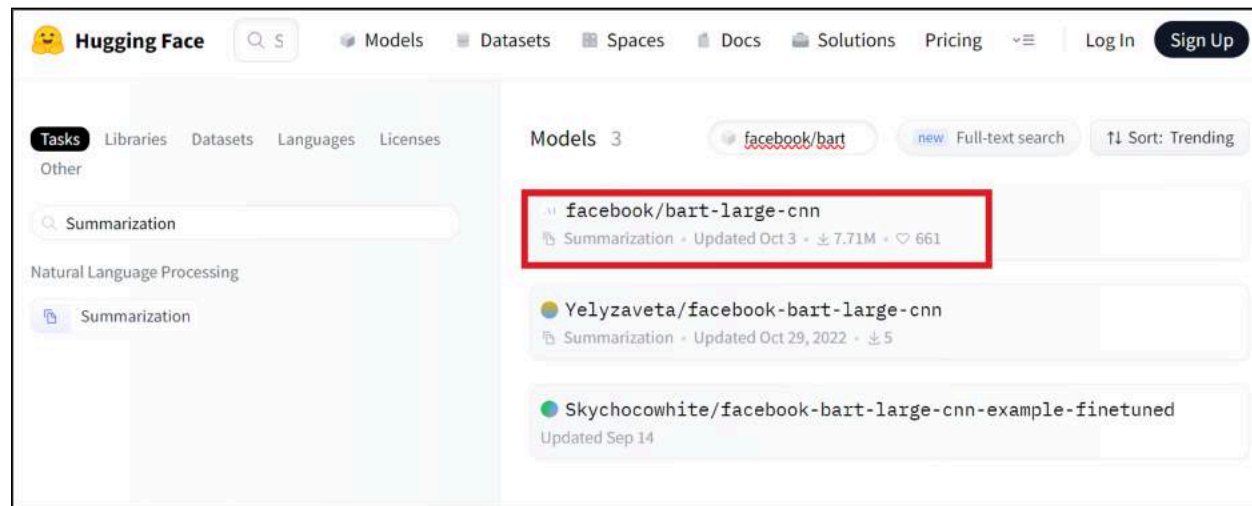
2. Click on the **Summarization** task under Natural Language Processing.



3. Next, you need to select a model say, **facebook/bart-large-cnn** from the list of models in the right panel. So, type facebook/bart-large-cnn in the **Filter by name** box.



4. Select **facebook/bart-large-cnn** from the list of displayed models.



5. The specific model page opens. You can see an example of a summary in the **Inference API** section.

facebook / bart-large-cnn like 661

Summarization Transformers PyTorch TensorFlow JAX Rust cnn_dailymail English bart text2text-generation Eval Results

Inference Endpoints arxiv:1910.13461 License: mit

Model card Files and versions Community 61

Edit model card

BART (large-sized model), fine-tuned on CNN Daily Mail

BART model pre-trained on English language, and fine-tuned on [CNN Daily Mail](#). It was introduced in the paper [BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension](#) by Lewis et al. and first released in [this repository] (<https://github.com/pytorch/fairseq/tree/master/examples/bart>).

Disclaimer: The team releasing BART did not write a model card for this model so this model card has been written by the Hugging Face team.

Model description

BART is a transformer encoder-encoder (seq2seq) model with a bidirectional (BERT-like) encoder and an autoregressive (GPT-like) decoder. BART is pre-trained by (1) corrupting text with an arbitrary noising function, and (2) learning a model to

Downloads last month
7,714,534

Inference API

Summarization Examples

The tower is 324 metres (1,063 ft) tall, about the same height as an 81-storey building, and the tallest structure in Paris. Its base is square, measuring 125 metres (410 ft) on each side. During its construction, the Eiffel Tower surpassed the Washington Monument to become the tallest man-made structure in the world, a title it held for 41 years until the Chrysler Building in New York City was finished in 1930. It was the first structure to reach a height of 300 metres. Due to the addition of a broadcasting aerial at the top of the tower in 1957, it is now taller than the Chrysler Building by 5.2 metres (17 ft). Excluding transmitters, the Eiffel Tower is the second tallest free-standing structure in France after the Millau Viaduct.

6. Delete the existing text under the input field and paste the following text lines:

1. 1

1. The company reported a modest increase in profits this quarter. They attribute this growth to their successful marketing campaigns and new product launches.

Copied!

You can copy the text using the Copy icon.

facebook / bart-large-cnn like 661

Summarization Transformers PyTorch TensorFlow JAX Rust cnn_dailymail English bart text2text-generation Eval Results

Inference Endpoints arxiv:1910.13461 License: mit

Model card Files and versions Community 61

Edit model card

BART (large-sized model), fine-tuned on CNN Daily Mail

BART model pre-trained on English language, and fine-tuned on [CNN Daily Mail](#). It was introduced in the paper [BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension](#) by Lewis et al. and first released in [this repository] (<https://github.com/pytorch/fairseq/tree/master/examples/bart>).

Disclaimer: The team releasing BART did not write a model card for this model so this model card has been written by the Hugging Face team.

Model description

BART is a transformer encoder-encoder (seq2seq) model with a bidirectional (BERT-like) encoder and an autoregressive (GPT-like) decoder. BART is pre-trained by (1) corrupting text with an arbitrary noising function, and (2) learning a model to

Downloads last month
7,714,534

Inference API

Summarization Examples

The company reported a modest increase in profits this quarter. They attribute this growth to their successful marketing campaigns and new product launches.

Compute

Computation time on Intel Xeon 3rd Gen Scalable cpu: cached

The tower is 324 metres (1,063 ft) tall, about the same height as an 81-storey building. Its base is square, measuring 125 metres (410 ft) on each side. During its construction, the Eiffel Tower surpassed the Washington Monument to become the tallest man-made structure in

7. Then click on **Compute**. A summary will be generated.

The screenshot shows the Hugging Face model card for BART (large-sized model), fine-tuned on CNN Daily Mail. The card includes a description of the model, a disclaimer, and a model description. On the right, the Inference API section shows a 'Compute' button highlighted with a red box. Below the button, the generated summary is displayed, also highlighted with a red box. The summary contains hallucinated information not present in the original text.

BART (large-sized model), fine-tuned on CNN Daily Mail

BART model pre-trained on English language, and fine-tuned on [CNN Daily Mail](#). It was introduced in the paper [BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension](#) by Lewis et al. and first released in [this repository] (<https://github.com/pytorch/fairseq/tree/master/examples/bart>).

Disclaimer: The team releasing BART did not write a model card for this model so this model card has been written by the Hugging Face team.

Model description

BART is a transformer encoder-encoder (seq2seq) model with a bidirectional (BERT-like) encoder and an autoregressive (GPT-like) decoder. BART is pre-trained by (1) corrupting text with an arbitrary noising function, and (2) learning a model to reconstruct the original text.

BART is particularly effective when fine-tuned for text generation (e.g. summarization, translation) but also works well for comprehension tasks (e.g. text classification).

Downloads last month: 7,714,534

Inference API

Summarization Examples

The company reported a modest increase in profits this quarter. They attribute this growth to their successful marketing campaigns and new product launches.

Compute

Computation time on Intel Xeon 3rd Gen Scalable cpu: 4.133 s

The company reported a modest increase in profits this quarter. They attribute this growth to their successful marketing campaigns and new product launches. The company also reported a small increase in sales and a slight increase in revenue. The growth was attributed to a successful marketing campaign and a new product launch.

JSON Output Maximize

In this example, the desired summary should accurately reflect the information in the original text. However, the hallucinated summary exaggerates the information like, **The company also reported a small increase in sales and a slight increase in revenue. The growth was attributed to a successful marketing campaign and a new product launch.** This information is not consistent with the source text. This is an example of hallucination in text summarization, where the summary contains information not supported by the source content.

Exercise 2: Hallucination of Image Generating LLMs

Image hallucination models produce visuals that are either distorted, surreal, or unrealistic. These images contain strange patterns or meaningless combinations of objects.

Step 1: Access Hugging Face and navigate to the Models pane

1. Click [Hugging Face - The AI community building the future](#) to launch the Hugging Face platform.

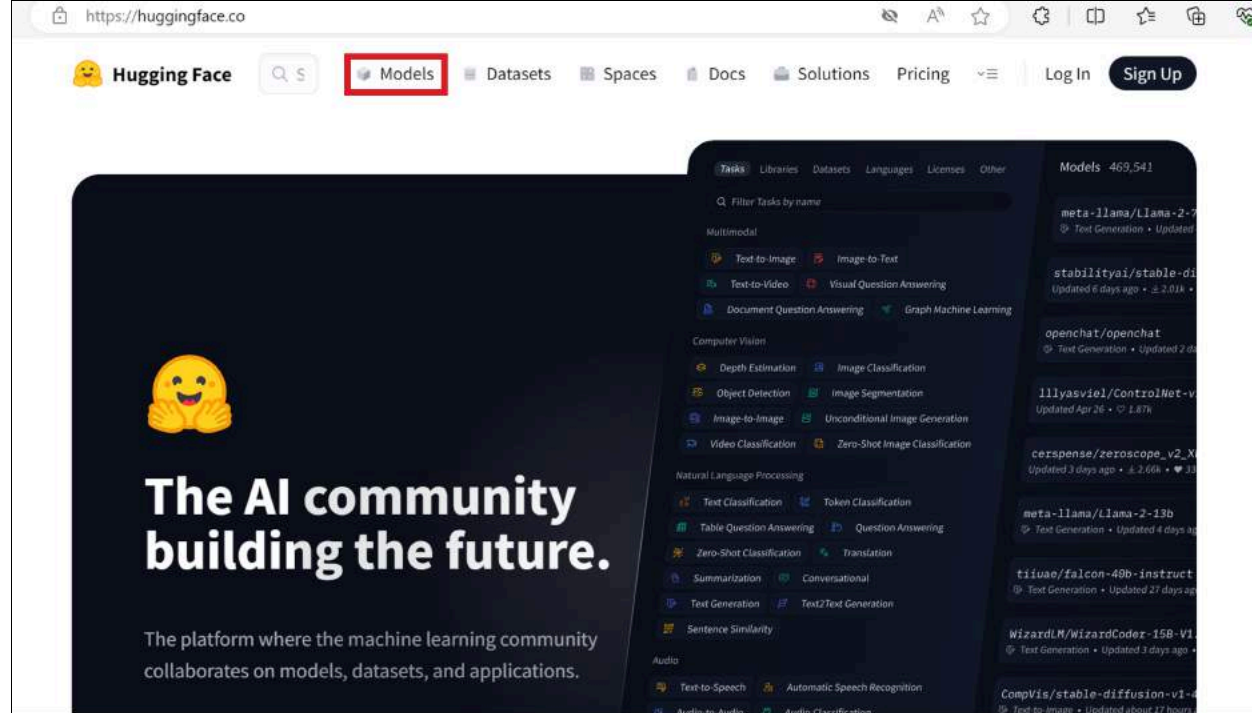
Note: If you cannot access Hugging Face by clicking on the link, copy the following URL and paste it in a browser.

1. 1
1. <https://huggingface.co/>

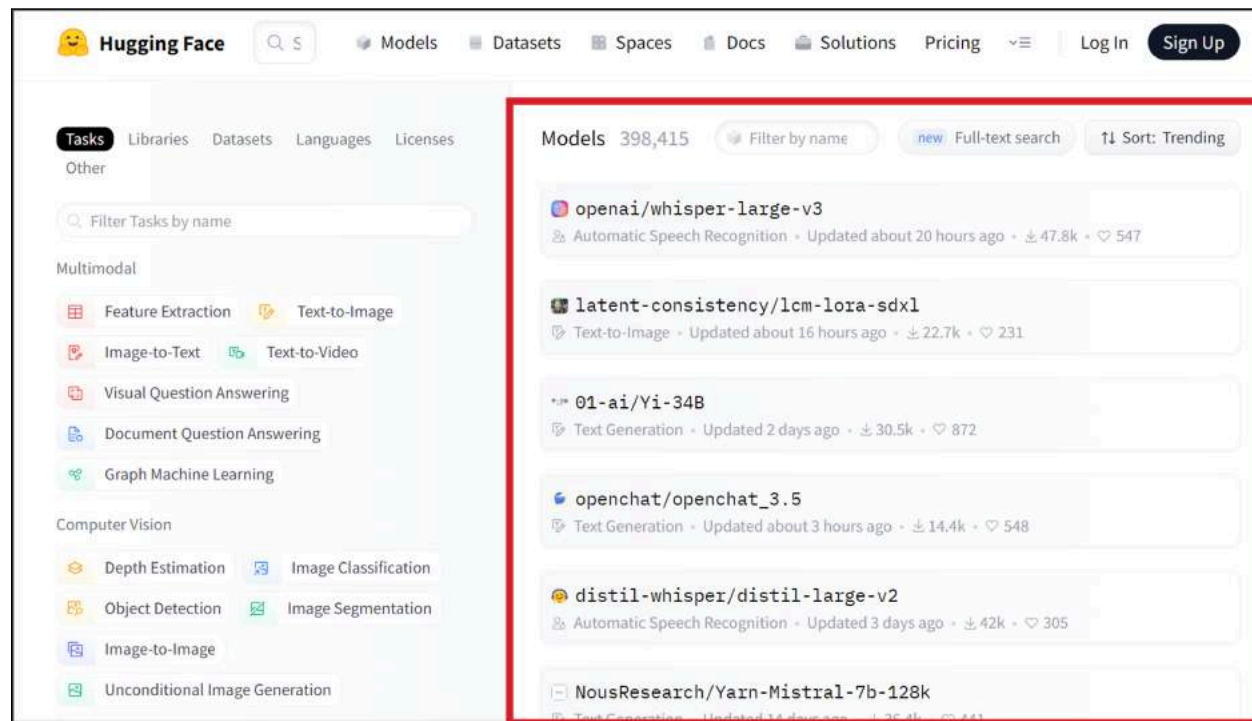
Copied!

Although you can use the platform without logging in, we recommend logging in for the best experience.

2. Once launched, click **Models** in the top menu bar.

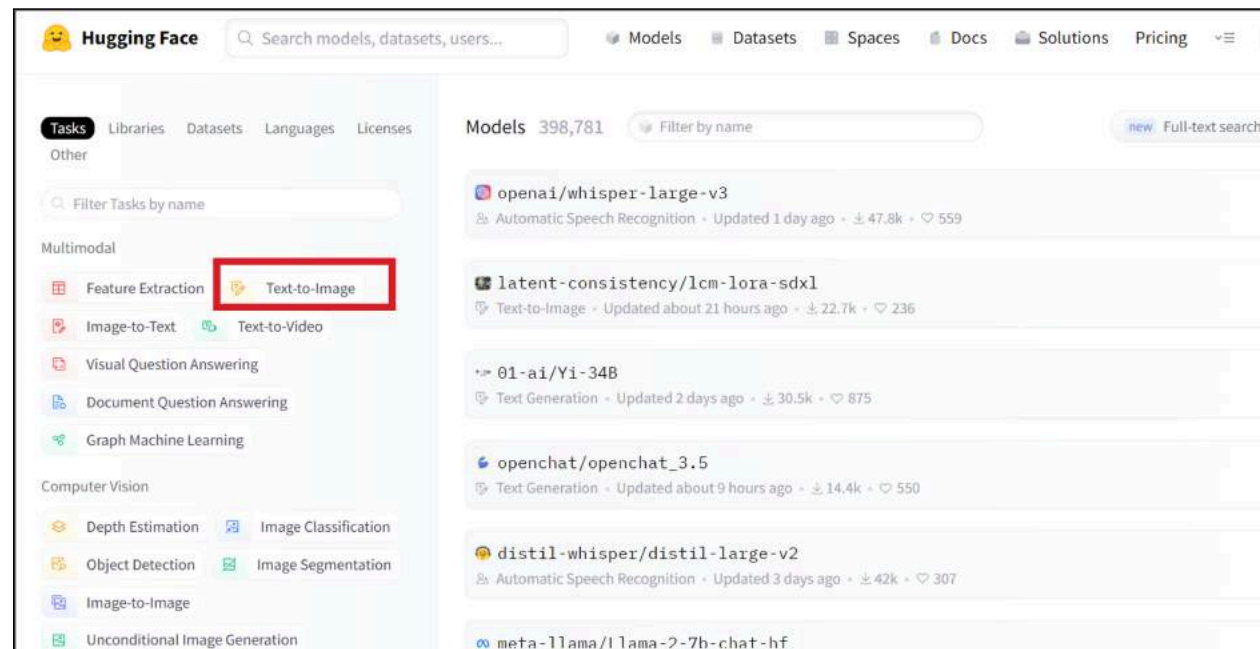


3. You will get a list of models in the Models pane.



Step 2: Generate the image

1. Click on the **Text-to-Image** button under the **Multimodal** category in the left pane. This will display the list of models available for text-to-image generation in the right pane.



2. Select **stabilityai/stable-diffusion-xl-base-1.0** from the list of models in the left pane.



3. You have the model card for the **stabilityai/stable-diffusion-xl-base-1.0** displayed.

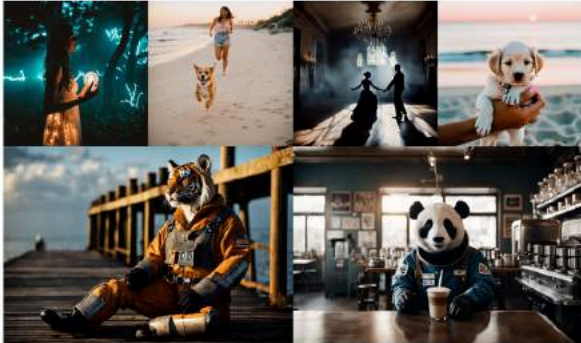
stabilityai/stable-diffusion-xl-base-1.0 like 3.52k

Text-to-Image Diffusers ONNX StableDiffusionXLPipeline stable-diffusion Inference Endpoints arxiv:2307.01952 arxiv:2211.01324 arxiv:2108.01073 arxiv:2112.10752 License: openrail++

Model card Files and versions Community 116 Deploy Use in Diffusers

Edit model card

SD-XL 1.0-base Model Card



Downloads last month
9,947,918

Inference API

Text-to-Image

Your sentence here... Compute

This model can be loaded on the inference API on-demand.

JSON Output Maximize

Spaces using stabilityai/stable-diffusion-xl-base-1.0 598

- google/sd-xl
- Shopify/background-replacement
- hysts/SD-XL
- Manjushri/SDXL-1.0
- diffusers/stable-diffusion-xl-inpainting
- songweig/rich-text-to-image
- TencentARC/T2I-Adapter-SDXL

4. Under the Inference API section, provide a text prompt for image generation. Let's enter the following prompt in the **Text-to-Image** field and then click on **Compute**.

1. 1

1. Architecture Diagram of Computer

Copied!

⚡ Inference API ⓘ

📄 Text-to-Image

Architecture Diagram of Computer

This model can be loaded on the Inference API on-demand.

</> JSON Output

🖥 Maximize

🏠 Spaces using stabilityai/stable-diffusion-xl-base-1.0 597

🔥 google/sdxl

🌐 hysts/SD-XL

⚡ Manjushri/SDXL-1.0

🔥 diffusers/stable-diffusion-xl-inpainting

🌐 songweig/rich-text-to-image

🚀 TencentARC/T2I-Adapter-SDXL

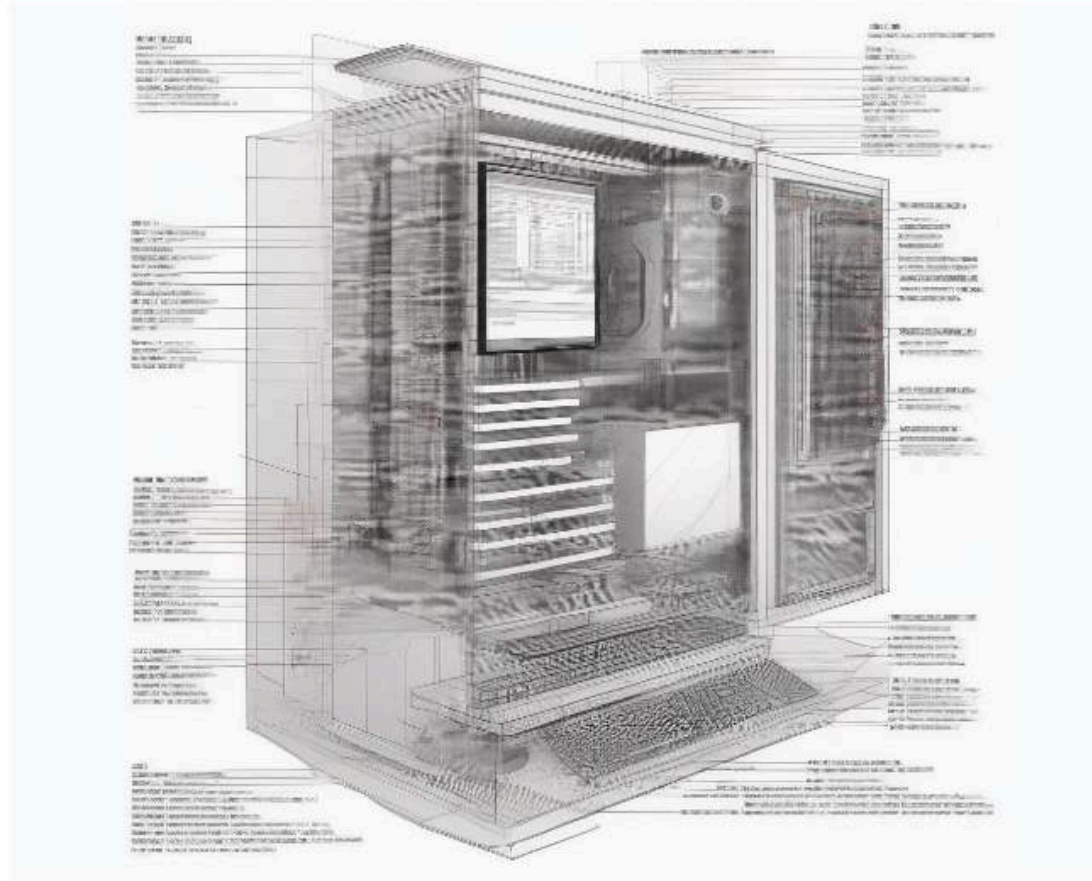
🚀 TencentARC/T2I-Adapter-SDXL-Sketch

5. The following image is generated by the model.

Architecture Diagram of Computer

Compute

Computation time on gpu: 6.741 s



The image generated here is hallucinated, as this is not a labeled representation of the 'Architecture Diagram of a Computer'. This is because the model is not trained on the relevant data required to generate in-context images. Instead, it trained on massive data sets. Therefore, it generated a hallucinated image.

Exercise 3: Experimenting with prompts to reduce hallucination

A common cause of the hallucination of LLMs can relate to the input context. Here, context refers to the information that is given to the models as an input prompt. Context can help guide the model to produce relevant and accurate outputs. However, it can also confuse or mislead the model if it's unclear, inconsistent, or contradictory.

As a user, we can reduce hallucinations in our conversations with LLMs by providing clear and specific prompts. The more precise and detailed the input prompt, the more likely the LLM will generate relevant and accurate outputs. Another way to reduce hallucinations can be through multi-shot prompting, where you can provide examples of the desired output, format, or context.

Let's experiment with ChatGPT (based on GPT 3.5) to understand how LLMs can produce inaccurate or irrelevant output based on the lack of clarity of context and how providing clearer and specific prompts can reduce hallucinations.

Note: The output produced at your end may be different than the output shown in this exercise. The basic idea to comprehend is that the output of a model can be modified by modifying the prompts.

Step 1: Log in to ChatGPT

1. Launch [ChatGPT](https://chat.openai.com/) and sign in.

Note: If you cannot access ChatGPT by clicking on the link, copy the following URL and paste it into a browser.

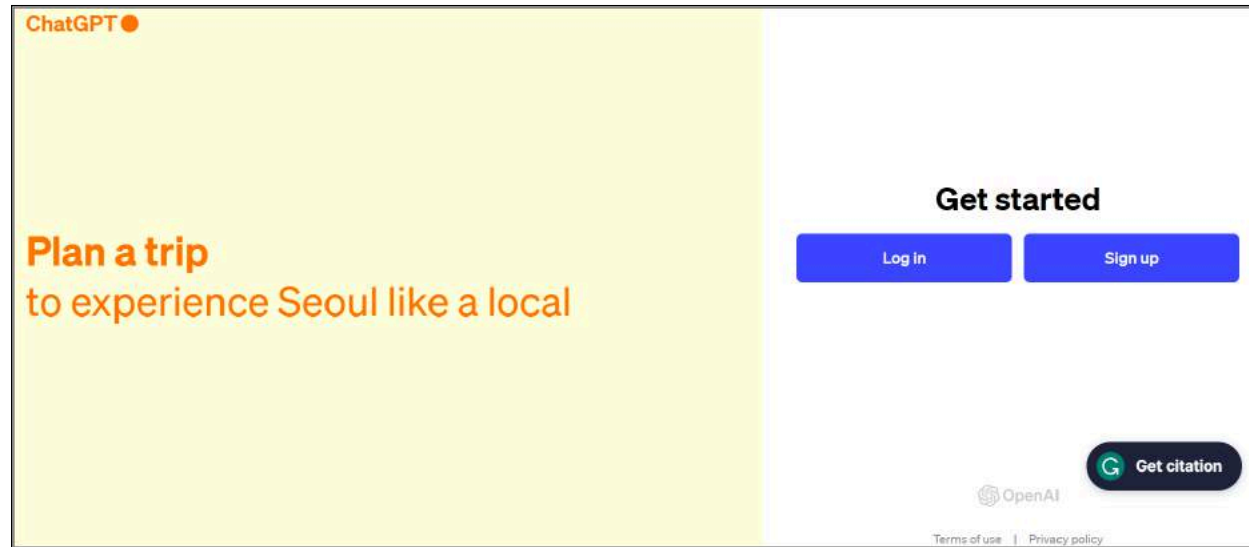
1. 1

1. <https://chat.openai.com/>

Copied!

2. Once launched, you must sign up or login.

Note: You can create a free account on ChatGPT and use GPT-3.5 features for free.



a. To sign up, you will be required to enter your email ID and password. You will receive an automated verification email. After you click verification, you will need to enter your personal details and follow the steps to complete the sign up on your first login.

b. After sign up, you will reach the login page, as shown below.

Welcome back

Email address

Continue

Don't have an account? [Sign up](#)

OR



Continue with Google

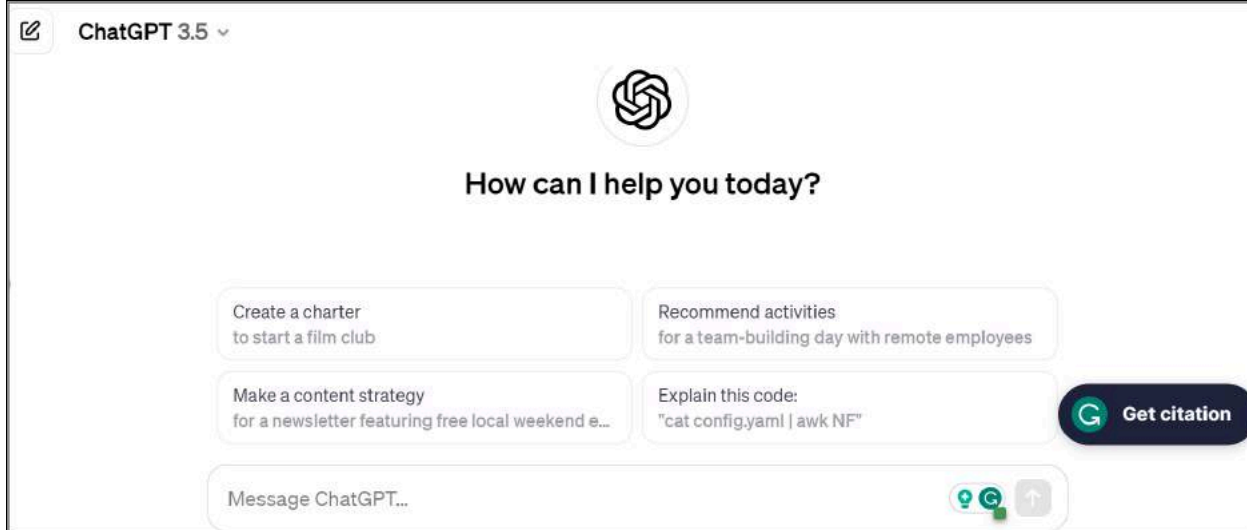


Continue with Microsoft Account



Continue with Apple

3. After login, you will view the OpenAI chatbot platform, ChatGPT, as shown below.



Step 2: Generate the text

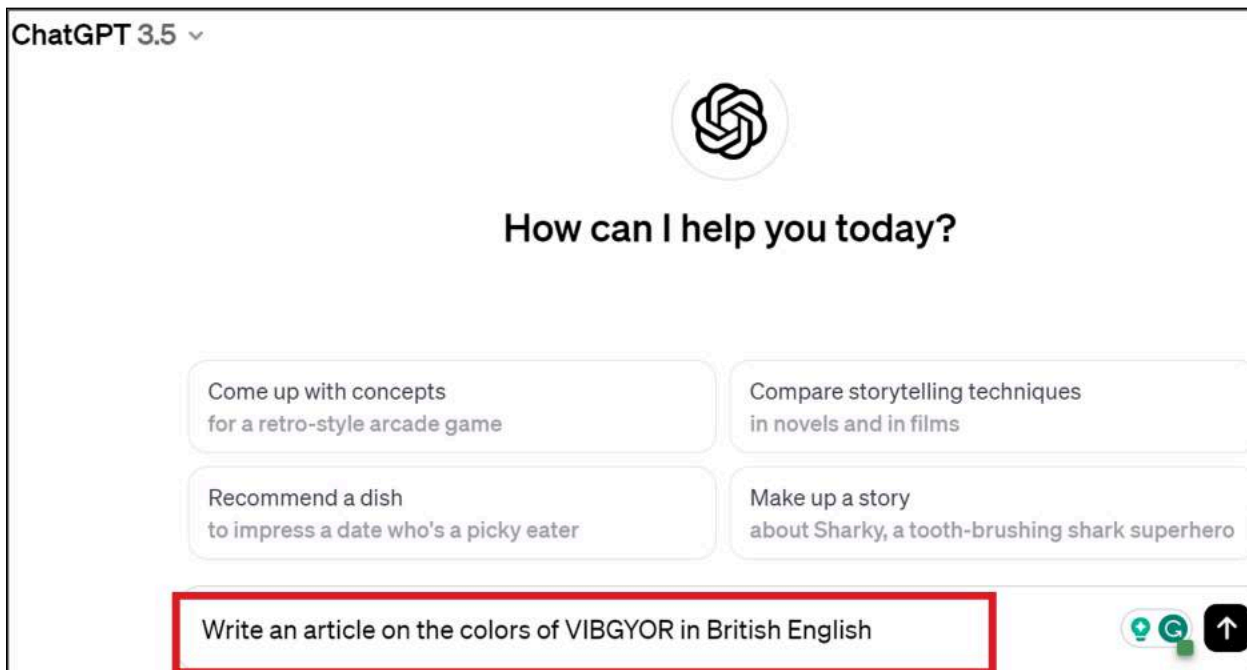
In this exercise, our purpose is to generate an article on 'Colors of VIBGYOR' and we want this article to be written in British English (the English language as it is spoken and written in the U.K.)

1. Let's enter the following prompt in the Message ChatGPT field.

1. 1

1. Write an article on the colors of VIBGYOR in British English

Copied!



2. Click the **Send Message** icon to generate the output.



How can I help you today?

Come up with concepts
for a retro-style arcade game

Compare storytelling techniques
in novels and in films

Recommend a dish
to impress a date who's a picky eater

Make up a story
about Sharky, a tooth-brushing shark superhero

Write an article on the colors of VIBGYOR in British English



3. We received the output shown in the following image.



ChatGPT

Title: "Exploring the Vibrant Spectrum: The Colours of VIBGYOR in British English"

Introduction:

The world of color is a fascinating realm that captures the essence of beauty, diversity, and the sheer wonder of light. One of the ways we perceive this spectrum is through the acronym VIBGYOR, which represents the sequence of colors in a rainbow: Violet, Indigo, Blue, Green, Yellow, Orange, and Red. In this article, we will delve into the distinctive characteristics of each color, exploring their significance and impact in the context of British English.

Violet:

Violet, the first color in the VIBGYOR sequence, is a regal and sophisticated hue. In British English, it is associated with elegance, luxury, and creativity. The term 'violet' itself is derived from the Latin word 'viola,' signifying the violet flower. This color is often linked to royalty, symbolizing power and opulence.



Indigo:

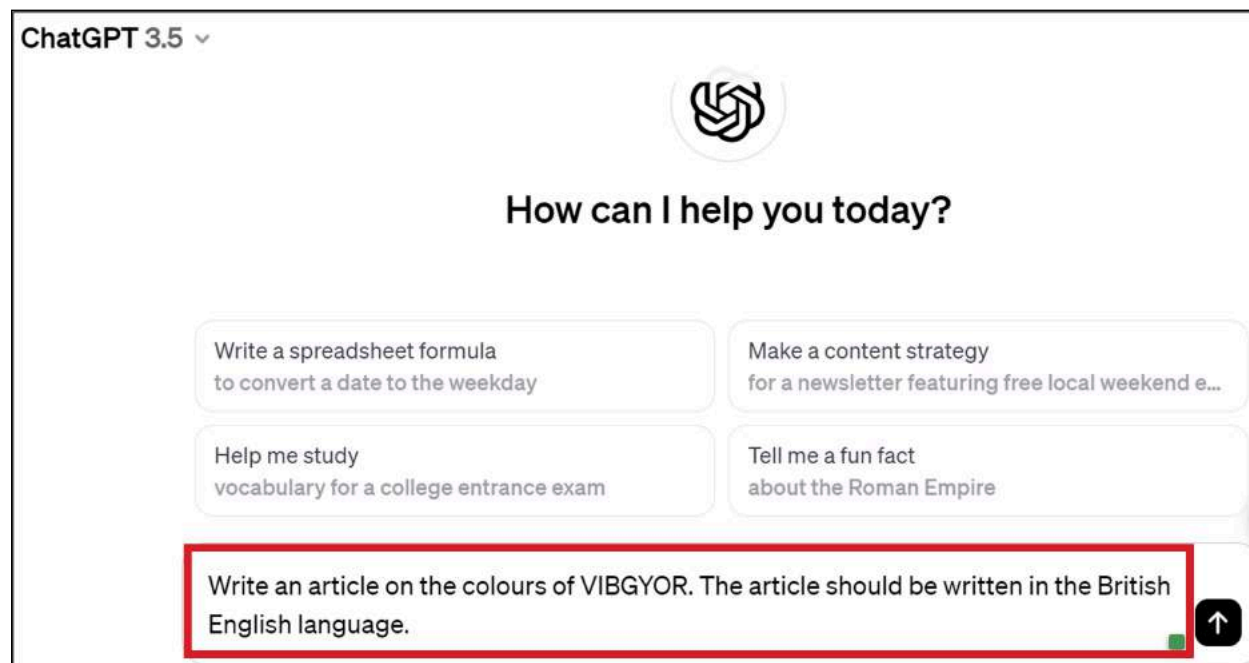
In the response, you can see that instead of providing the article on 'Colors of VIBGYOR' written in the British English language, the system misunderstood the context and has provided the article stating the significance and impact of the colors in VIBGYOR in the 'context of British English'.

4. Let's modify the prompt to specify the context in a clearer way. Also, as we want the output to be in British English, let's provide the prompt as well in British English; so, let's change 'colors' to 'colours' in the prompt. In a new session of ChatGPT, let's enter the following prompt.

1. 1

1. Write an article on the colours of VIBGYOR. The article should be written in British English language.

Copied!



5. Click the **Send Message** icon. Based on the modified prompt, we received the output shown in the following image.



How can I help you today?

Write a spreadsheet formula
to convert a date to the weekday

Make a content strategy
for a newsletter featuring free local weekend e...

Help me study
vocabulary for a college entrance exam

Tell me a fun fact
about the Roman Empire

Write an article on the colours of VIBGYOR. The article should be written in the British English language.



In this output, although the significance and impact of the colors are not specified in the context of 'British English', the output still is not written explicitly in British English. For example, note that the word 'colour' is written as 'color', representing American English.

6. Let's make the prompt more precise and specify an example to help the system understand the context and the required modifications in the previous output. Let's enter the following prompt.

1. 1

1. The article should be written in 'British English as the language. For example, the word 'colors' should be written as 'colours' in British English. Similarly, convert other words and punctuation

Copied!



ChatGPT

Title: The Spectacular Palette of VIBGYOR: Unveiling the Rainbow's Rich Tapestry

Introduction

In the realm of light and optics, the breathtaking display of colors finds its most enchanting expression in the formation of a rainbow. This natural phenomenon, often described as a celestial bridge between Earth and sky, reveals the mesmerizing spectrum known as VIBGYOR. Each letter in this acronym represents a distinct hue, and together they c ↓ e a visual symphony that captivates the human eye.

The article should be written in 'British English as the language. For example, the word 'colors' should be written as 'colours' in British English. Similarly, convert other words and punctuation based on the rules of British English.



7. Click the **Send Message** icon to generate the output.

Summary

Congratulations on completing the hands-on lab, Hallucination of LLMs.

In this lab, you explored the popular multi-modal open-source tool Hugging Face to identify instances of hallucinations of text and image-generating LLMs. You also identified how you, as a user of generative Ai, can reduce Hallucinations.

© IBM Corporation. All rights reserved.



Skills Network