

# Multimodal Prompts

Estimated time: 10 mins

## Introduction

In the evolving landscape of artificial intelligence, the ability to process and understand different types of data - such as text, images, audio, and video - is becoming increasingly vital. Traditional language models have focused predominantly on single-modal inputs, typically text. However, real-world communication is inherently multimodal: we interpret words, visuals, tone, and context simultaneously. **Multimodal prompts**, therefore, are a powerful development in AI that allow models to handle and reason across multiple types of input data simultaneously.

Multimodal prompting refers to the practice of using inputs from different modalities - for example, text and images - in a single prompt to an AI system. Multimodal models, such as OpenAI's GPT-4 (multimodal variant), Google's Gemini, or Meta's ImageBind, are designed to understand and generate responses that consider the interaction between modalities. This capability opens new frontiers in how AI systems interpret and interact with the world, offering more natural, rich, and insightful outputs.

## Advantages of multimodal prompts

### 1. Enhanced contextual understanding:

Multimodal prompts allow AI to interpret complex situations better. An image alone might be ambiguous, but combined with textual context, it becomes much more meaningful.

### 2. Versatility across tasks:

They enable a single model to perform a wide range of tasks—from image captioning to document summarization and even creative applications like storytelling from visuals.

### 3. Improved accuracy:

By integrating different data sources, models can cross-validate and refine their outputs, leading to more accurate and relevant responses.

### 4. Natural human-like interaction:

Humans communicate multimodally. We speak, point, and show. AI that understands multimodal input can interact more naturally, improving user experience in applications like virtual assistants, tutoring systems, and customer service.

## Examples of usage

Consider some of the common ways in which multimodal prompts can prove to be useful.

**Important Note:** All documents and images used in the examples that follow are synthetically generated for the purpose of this reading.

### Summarizing documents containing visual content

In domains such as healthcare, research, and financial analysis, documents are not purely textual—they include visuals like graphs, tables, and schematics that hold essential meaning. A traditional text-only analysis may ignore these, resulting in incomplete or misleading summaries or insights. Multimodal prompting enables AI to extract insights from both text and embedded visuals for richer, more accurate document understanding.

Consider the following example.

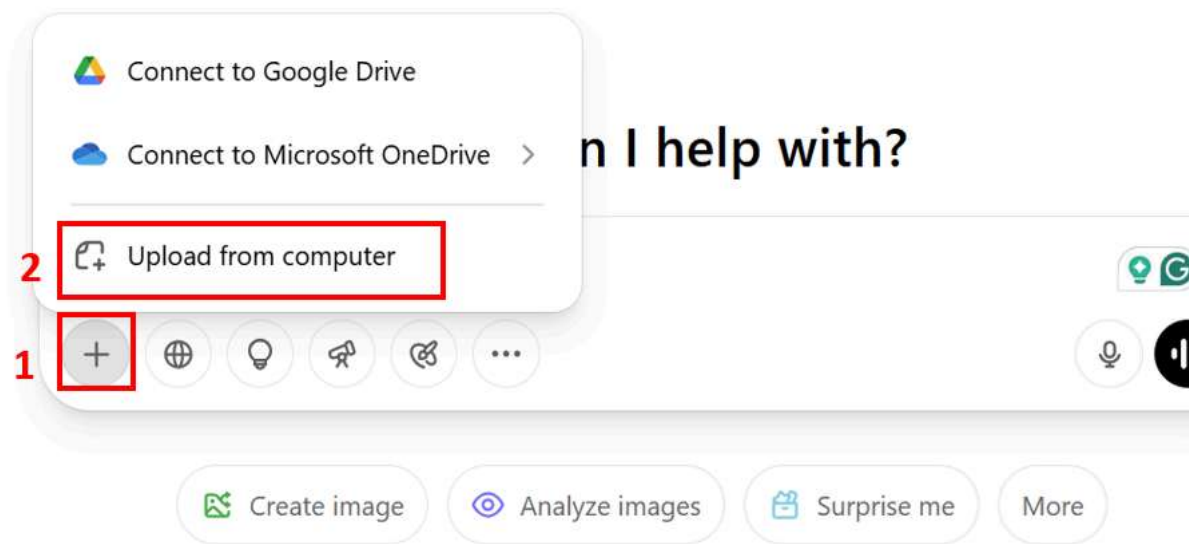
A company's mid-year sales report has been published, and you want to summarize the document to get the key takeaways.

You can download the report to be used using the link below.

[Sales Report PDF](#)

Log in to [chatgpt.com](https://chatgpt.com), and using the "Upload from computer" feature, upload the document to the prompting interface.

ChatGPT ▾



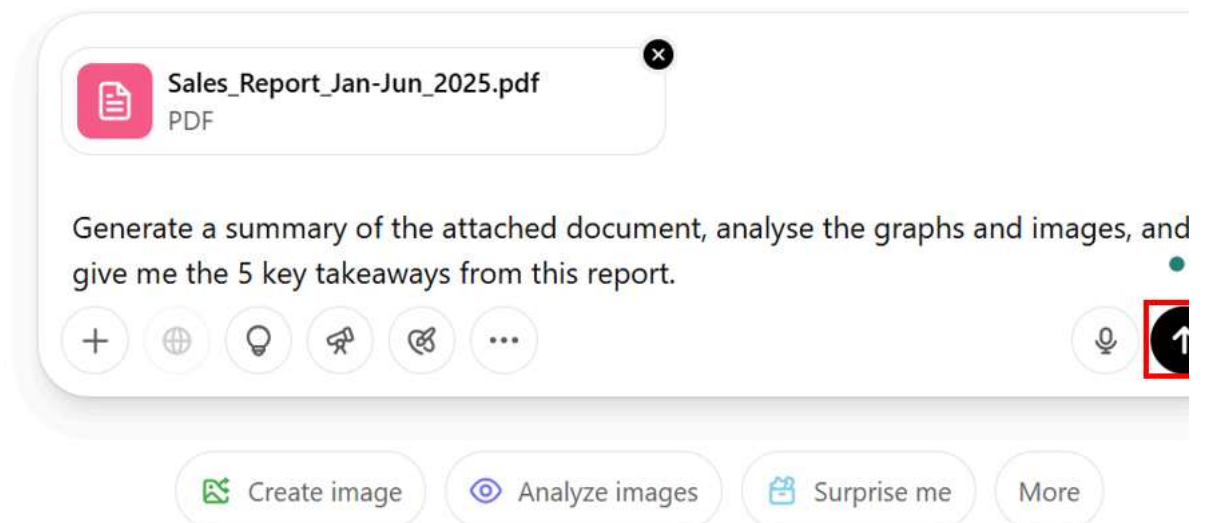
Further, you can add the text prompt indicating the action you want performed.

Let us assume here that we are trying to generate a summary of the document, identifying the five key takeaways from it. We can simply use the following prompt.

Generate a summary of the attached document, analyze the graphs and images, and give me the 5 key takeaways from this report.

ChatGPT ▾

## What can I help with?



As a response, the system will study the file and generate an appropriate response. It is important to note that LLM models like ChatGPT can generate different responses for the same query. The sample response generated for us is shared below.

► [Click here to display the sample response](#)

### Creative content generation

Multimodal capability of generative AI, or GenAI, models can also be leveraged for generating creative content inspired by an image.

Consider, for example, the following image.



Now, let's use ChatGPT to generate a short story inspired by this image. You can simply copy this image by right clicking it, and pasting it to the ChatGPT querying interface. Use the following text prompt to generate it.

Write a short story, inspired by the image attached, focusing on the struggles of the painter and what he intends to convey using his art.

Inspired by the image, the model generates a story for us as shared below.

► [Click here to show the response](#)

### Creative captioning and tagging

GenAI models are valuable tools for social media marketing, especially when it comes to crafting creative captions and smart tagging. They can quickly generate engaging, on-brand captions that match the tone and message of a post—whether it's witty, professional, or emotionally driven. This helps save time while keeping content fresh and relatable. GenAI models can also suggest effective hashtags and user tags to boost discoverability and reach. By analyzing image content and context, they create captions that tell a story and tagging strategies that connect with the right audience. In a crowded social media space, GenAI empowers marketers to stand out with consistent, creative, and targeted posts that drive engagement and build brand presence.

Let us now try and use [Google's Gemini](#) for generating a creative caption and relevant tags for the image below.



Once you have logged in to the Gemini interface, copy and paste the image shared above and add the following prompt.


Create an appropriate caption and hashtags for the product shared in the image, assuming that it is to be posted as part of a product marketing pitch

Gemini

2.5 Flash (experimental)

Try Gemini Advanced

Hello,



Create an appropriate caption and hashtags for the product shared in the image, assuming that it is to be posted as part of a product marketing pitch.

+

Canvas

The response generated by Gemini is shared below.

► [Click here to see the sample response](#)

## Conclusion

Multimodal prompts represent a pivotal advancement in artificial intelligence. By integrating and reasoning over various data types, they allow AI systems to interpret the world more holistically—much like humans do. As multimodal models become more sophisticated, they will drive innovation in fields like education, healthcare, journalism, and enterprise analytics. Whether it's summarizing complex reports, generating human-like image captions, or extracting deep insights from visuals, the power of multimodal prompting is setting the stage for a more intelligent, adaptable future.

## Author(s)

[Abhishek Gagneja](#)



# Skills Network