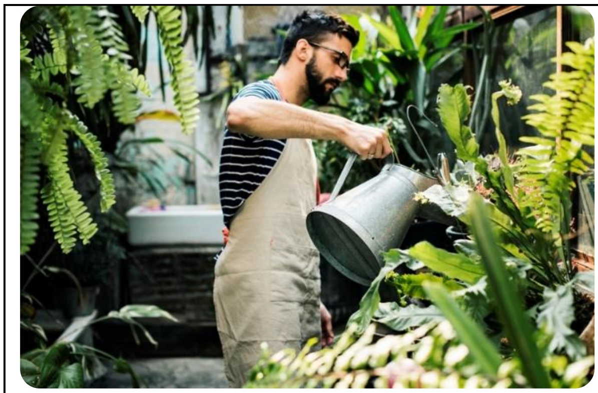
GPT-4 Turbo with Vision

A multimodal model that analyze images and provide textual responses to questions about them

Incorporates both natural language processing and visual understanding.

Detailed Image Descriptions:

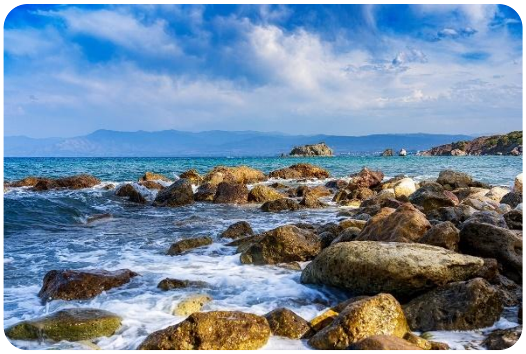
	**Vision AI** A person with a striped shirt and an apron is holding a metal watering can and watering the plants in the greenhouse. A large metal watering can with a long spout is being used by a person to water the green and leafy plants in the greenhouse. A greenhouse full of green and leafy plants is being watered by a person with a metal watering can. The person is wearing a striped shirt and an apron. A person is taking care of the plants in the greenhouse by watering them with a metal watering can. The person has a striped shirt and an apron on. The plants are green and leafy.

```
from util import getOpenAIClient

# Get OpenAI Client from Util.
client = getOpenAIClient()

response = client.chat.completions.create(
  model="gpt-4o-mini",
  messages=[
   {
     "role": "user",
     "content": [
       {"type": "text", "text": "What’s in this image?"},
        {
          "type": "image_url",
          "image_url": {
            "url": "https://upload.wikimedia.org/wikipedia/commons/thumb/d/dd/Gfp-wisconsin-madison-the-nature-boardwalk.jpg/2560px-Gfp-wisconsin-madison-the-nature-boardwalk.jpg",
          },
        },
      ],
    }
  ],
  max_tokens=300,
)
print(response.choices[0])
```

o/p: The image features a scenic outdoor landscape with a wooden path leading through lush green grass and foliage. The sky is blue with a few clouds scattered throughout. On either side of the path, there are various plants and trees, indicating a natural or park-like setting. Overall, it portrays a tranquil, picturesque environment.

Combine Text & Image Inputs

User: Write a six-line rhyming poem. Starting each line with the letter “c”.

The first three lines are around this image.  And the next three lines are around this image.	**Vision AI:** Calm and cool, the sea is blue, Crashing waves, a scenic view, Craggy rocks, a sight so true, Colorful and warm, the sun is low, Casting shadows, a golden glow, Cozy and calm, a peaceful show.

Example to compare images

```
from util import getOpenAIClient
client = getOpenAIClient()

response = client.chat.completions.create(
  model="gpt-4o-2024-05-13",
  messages=[
    {
      "role": "user",
      "content": [
        {
          "type": "text",
          "text": "What are in these images? Is there any difference between them?",
        },
        {
          "type": "image_url",
          "image_url": {
            "url":
"https://upload.wikimedia.org/wikipedia/commons/thumb/3/37/African_Bush_Elephant.jpg/1200px-African_Bush_Elephant.jpg",
          },
        },
        {
```

```
     "type": "image_url",
     "image_url": {
      "url": "https://www.elephantvoices.org/images/stories/field_studies/field_mr_nick_250kb.jpg",
     },
    },
   ],
  }
 ],
 max_tokens=300,
)
print(response.choices[0])
```

o/p:

```
The images both depict elephants, but there are notable differences between them:\n\n1. **Environment**:
The first elephant is set in a more arid landscape with dirt and sparse vegetation, while the second one is in a
greener, more grassy area.\n\n2. **Physical Features**: While both elephants appear to be of the same
species, they may have slightly different characteristics, such as the shape of their tusks or the texture of
their skin.\n\n3. **Posture and Position**: The elephants may have different poses; one might be more
upright while the other appears more relaxed.\n\nThese variations illustrate the diversity in environments and
individual elephants.
```

Limitations

While GPT-4 with vision is powerful and can be used in many situations, it is important to understand the limitations of the model. Here are some of the limitations we are aware of:

- **Medical images:** The model is not suitable for interpreting specialized medical images like CT scans and shouldn't be used for medical advice.
- **Non-English:** The model may not perform optimally when handling images with text of non-Latin alphabets, such as Japanese or Korean.
- **Small text**: Enlarge text within the image to improve readability but avoid cropping important details.
- Rotation: The model may misinterpret rotated / upside-down text or images.
- **Visual elements:** The model may struggle to understand graphs or text where colors or styles like solid, dashed, or dotted lines vary.
- **Spatial reasoning:** The model struggles with tasks requiring precise spatial localization, such as identifying chess positions.
- **Accuracy**: The model may generate incorrect descriptions or captions in certain scenarios.

- **Image shape**: The model struggles with panoramic and fisheye images.
- **Metadata and resizing**: The model doesn't process original file names or metadata, and images are resized before analysis, affecting their original dimensions.
- **Counting**: May give approximate counts for objects in images.
- **CAPTCHAS**: For safety reasons, we have implemented a system to block the submission of CAPTCHAs.