**Llama-3.2-11B-Vision**

**Introduction:**

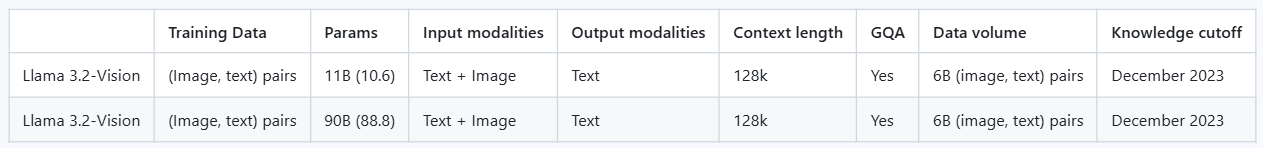
* The Llama 3.2-Vision collection of multimodal large language models (LLMs) is a collection of instruction-tuned image reasoning generative models in 11B and 90B sizes (text + images in / text out).

**Application:**

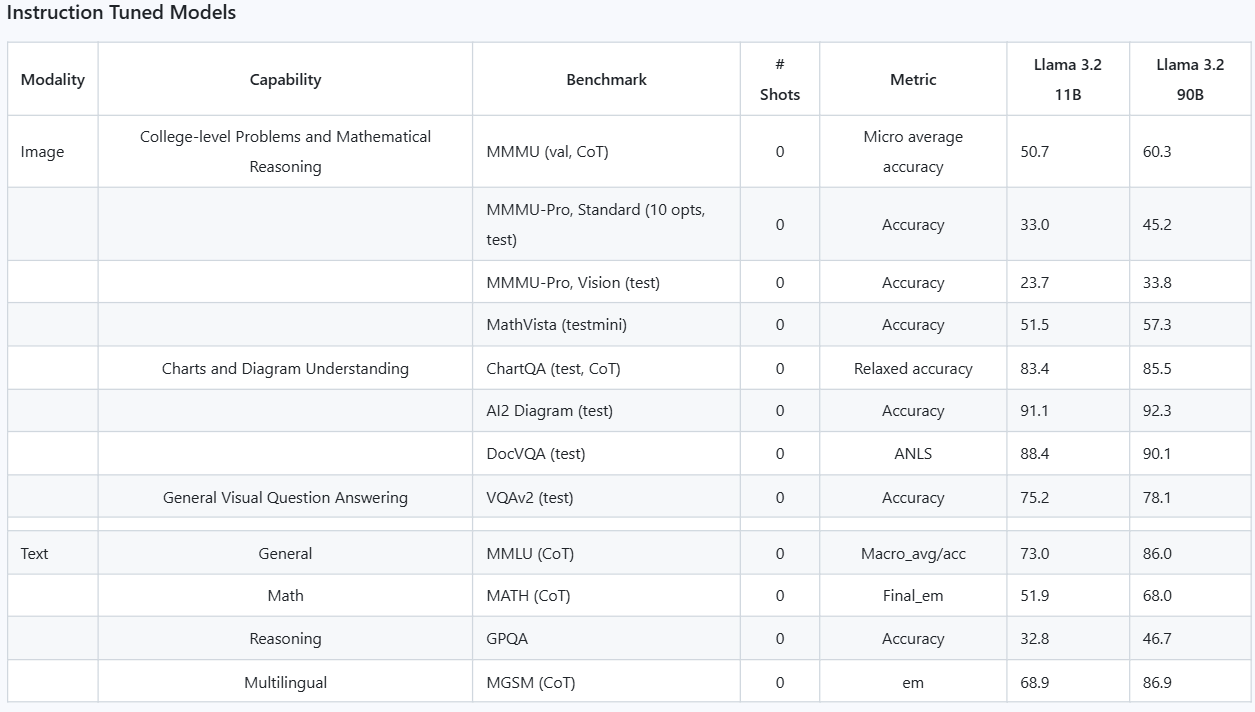
1. Visual recognition,
2. Image reasoning,
3. Captioning
4. Answering general questions about an image.

* **Supported Languages:** For text only tasks, English, German, French, Italian, Portuguese, Hindi, Spanish, and Thai are officially supported.
* Llama 3.2 has been trained on a broader collection of languages than these 8 supported languages. Note for image+text applications, English is the only language supported.

**Training**



Llama 3.2-Vision was pre-trained on 6B image and text pairs.





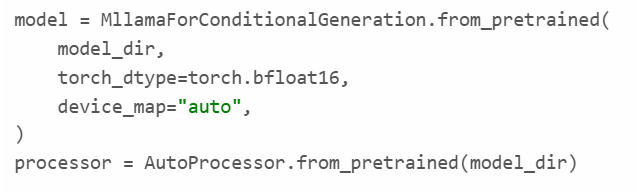
**Usage using the transformers**

Make sure to update your transformers installation via pip install --upgrade transformers.

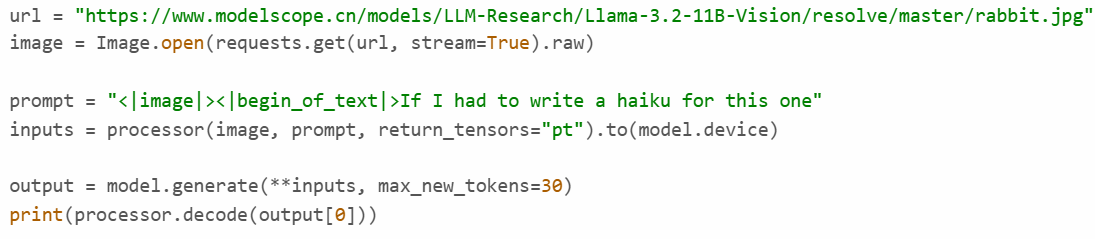
**Download Model from Modelscope**

from modelscope import snapshot\_download model\_dir = snapshot\_download('LLM-Research/Llama-3.2-11B-Vision', ignore\_file\_pattern=['\*.pth'])

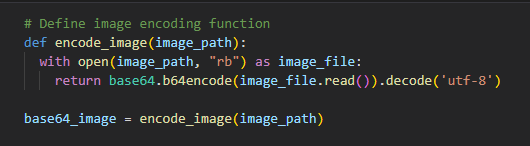




**Prompt and upload image and generate result:**



**Image Encoding**



**Image to text function:**

