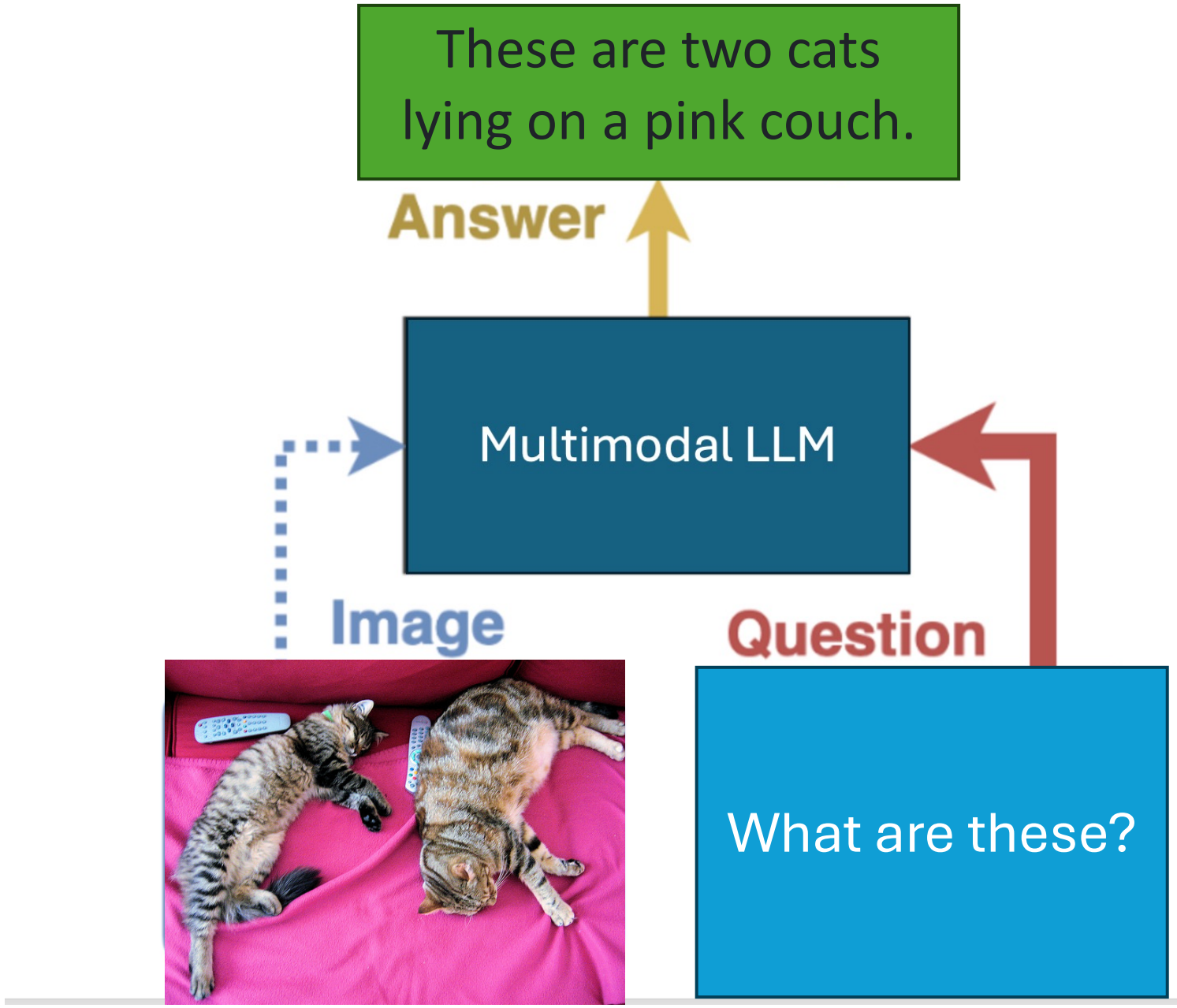
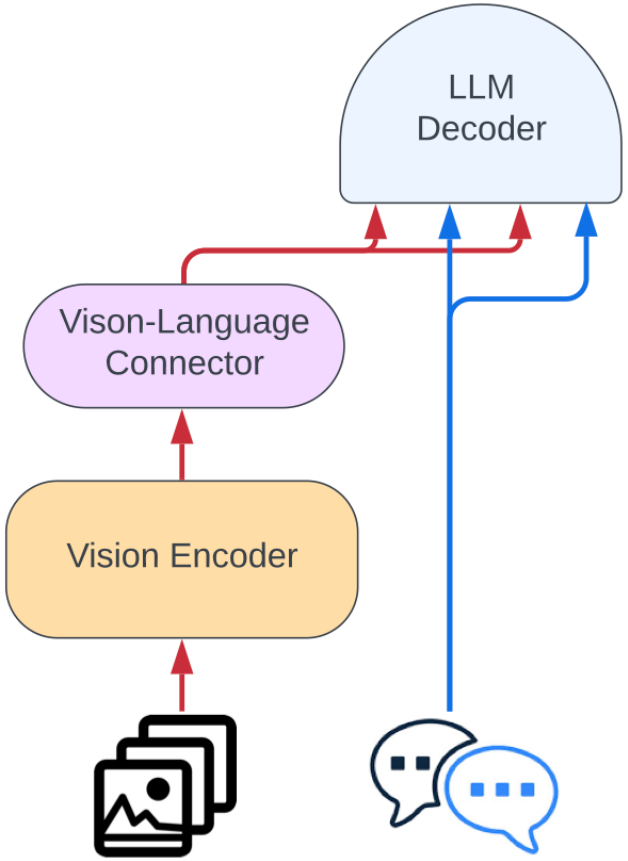
Idefics2: A Powerful 8B Vision-Language Model


NECTEC
a member of NSTDA


These are two cats lying on a pink couch.
Answer
Multimodal LLM
Image
Question
What are these?

Idefics2



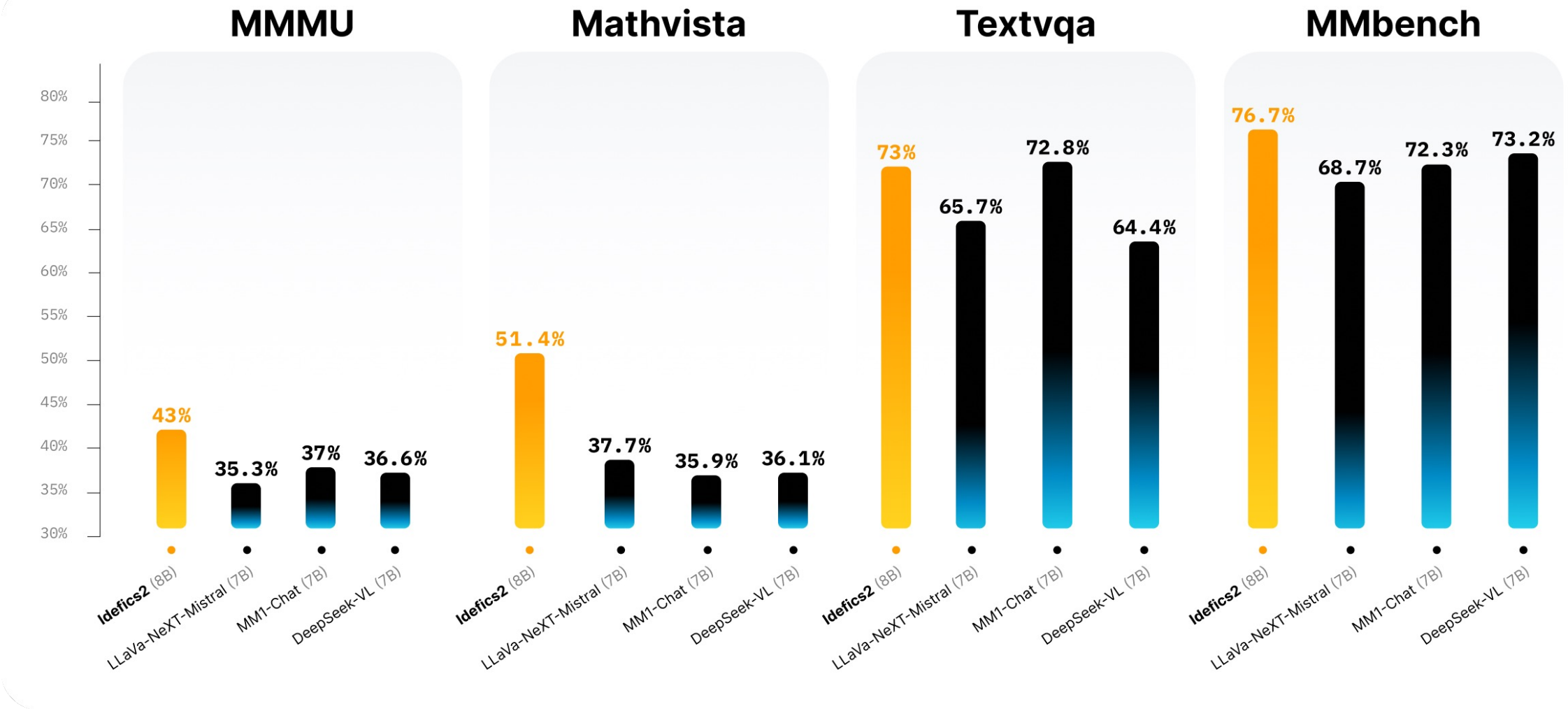
The model is built on top of two pre-trained models:

- Mistral-7B-v0.1 and
- siglip-so400m-patch14-384

https://huggingface.co/blog/idefics2

Idefics2

- 8B parameters
- Open license (Apache 2.0)
- Enhanced OCR (Optical Character Recognition) capabilities
- Achieves top-of-class performance on Visual Question Answering benchmarks, competing with much larger models.
  - LLava-Next-34B
  - MM1-30B-chat


MMMU
Mathvista
Textvqa
MMbench
80%
75%
70%
65%
60%
55%
50%
45%
40%
35%
30%
43%
35.3%
37%
36.6%
51.4%
37.7%
35.9%
36.1%
73%
65.7%
72.8%
64.4%
76.7%
68.7%
72.3%
73.2%
Idefics2 (8B)
LLaVa-NeXT-Mistral (7B)
MM1-Chat (7B)
DeepSeek-VL (7B)

Training Data

- Interleaved web documents (Wikipedia,OBELICS)
- Image-caption pairs (Public Multimodal Dataset, LAION-COCO)
- OCR data (PDFA (en), IDL and Rendered-text, and image-to-code data (WebSight))

Colab

- https://github.com/NielsRogge/Transformers-Tutorials/blob/master/Idefics2/Fine_tune_Idefics2_for_JSON_extraction_use_cases_(PyTorch_Lightning).ipynb
- https://colab.research.google.com/drive/1NtcTgRbSBKN7pYD3Vdx1j9m8pt3fhFDB?usp=sharing