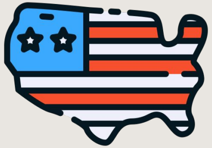
Empowering Healthcare with AI

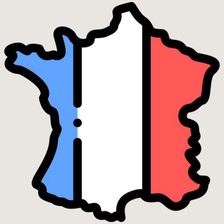
Julien Guyet - Preetha Pallavi - Malika Matissa

Introduction



25% of healthcare spending is considered wasteful

Hundreds of billions of dollars per year



95 minutes per day of time wasted on unproductive tasks

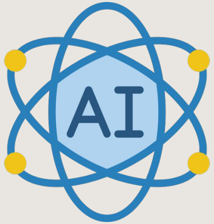
52 minutes per day spend searching for missing information.

Doctors spend 10 to 11 hours per week documenting patient files.

Sources :

- Plus d’une heure et demi par jour de temps perdu
- A la recherche du temps perdu… en médecine
- ALMOST 25% OF HEALTHCARE SPENDING IS CONSIDERED WASTEFUL. HERE’S WHY.

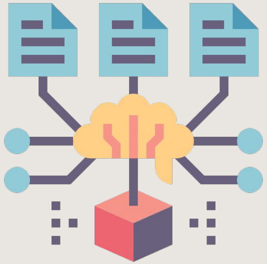
Project Overview



A powerful solution to address these challenges

Dataset from Kaggle : Brain Tumor Image DataSet

Dataset from Hugging Face: Path-VQA - 20k image-question pairs



2 Models :

YOLOV8
LLAVA

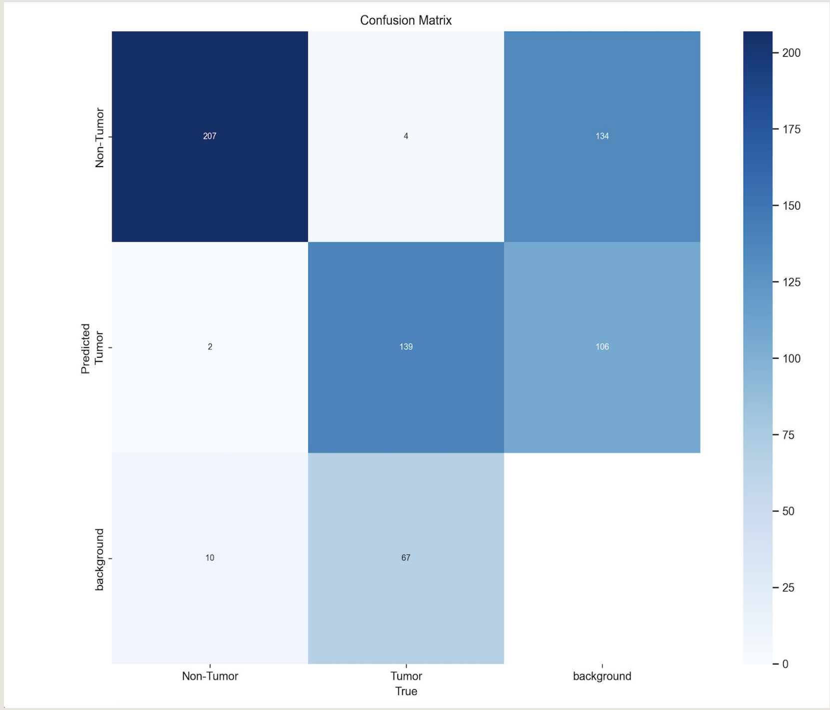
YOLO v8

- A **state-of-the-art** deep learning model designed for real-time object detection in computer vision applications.
- Real-time object detection with **cutting-edge speed** and **accuracy**.
- Efficient for **classification** and **segmentation** tasks.

Data Preparation:

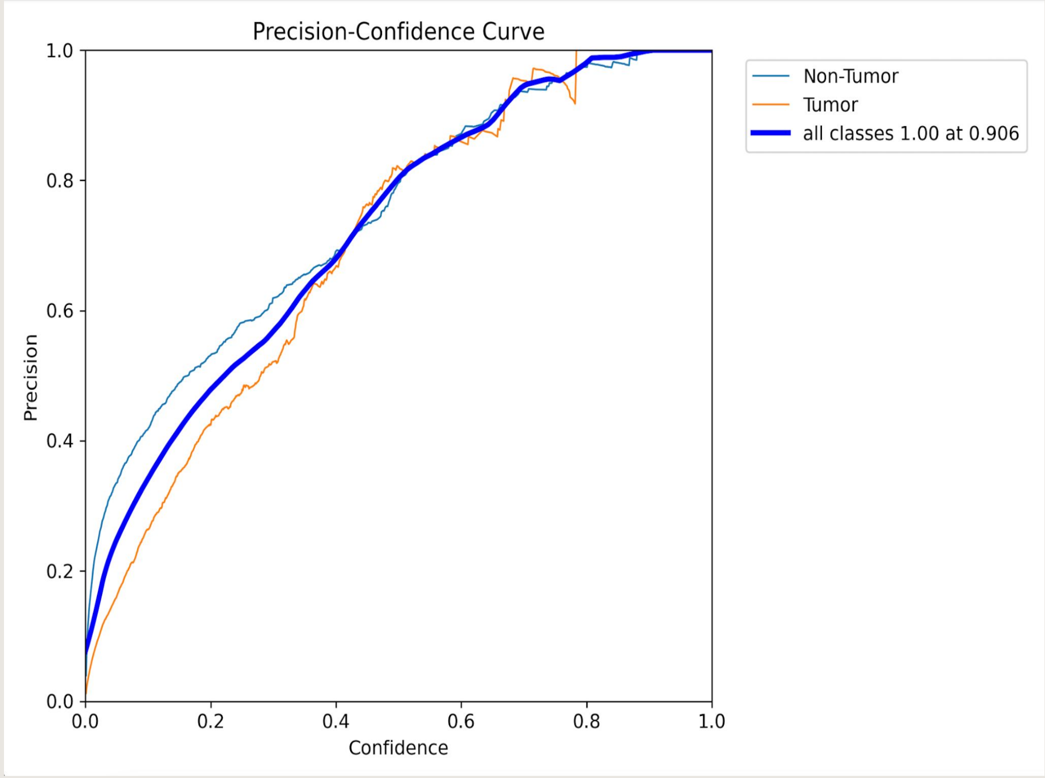
- First in both Train and validation set images and annotation file is present, we separate images and make "**images**" folder
- Next we change annotation file which is in **.json** format in Yolo V8 label format.
- They are then moved to the "**labels**" folder with each annotation file corresponding to the image name.. (This is done because **Yolo v8** accepts an image and label folder to train.)
- Then we map the images to its corresponding annotation file.
- Same process is applied to the **validation folder.**
- Path to both **training** and **validation** dataset is added to an **.yaml** file for the model to train.

Metrics: Confusion_matrix



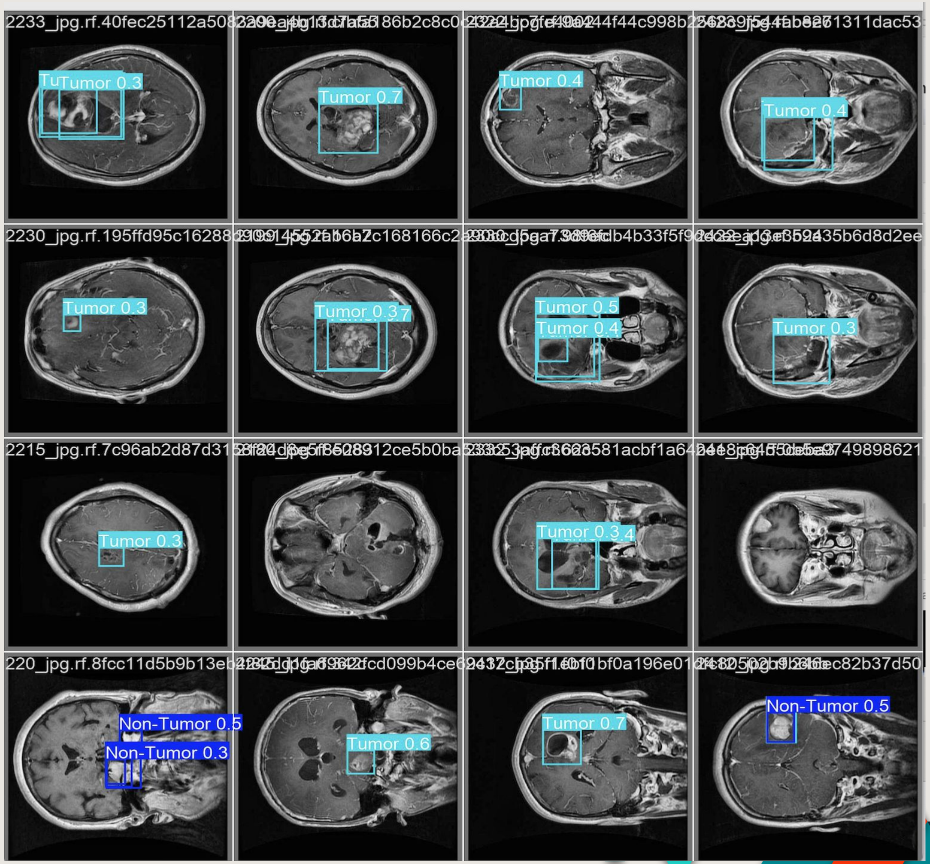
-> The model is good at finding most tumors (rarely misses them) but can be over cautious, sometimes mistaking healthy tissue for tumors (**more false positives than true positives**).

Metrics: Precision-Confidence curve



- The model's **confidence** increases, the accuracy of its positive predictions (**precision)** also goes up.
- However, there's a point (around 0.8 confidence) where further increases in confidence yield minimal benefit

Results:

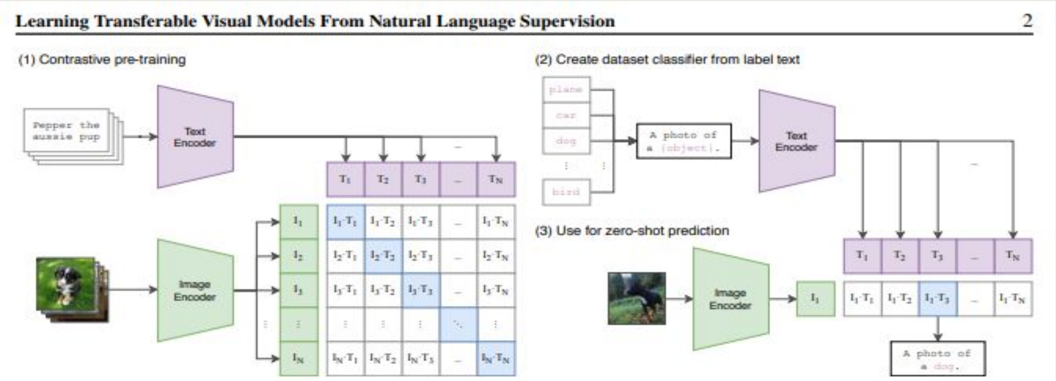


- Model demonstrates **high sensitivity** in detecting tumors, **minimizing** the risk of missed cases.
- However, **over-predict** tumors in healthy tissue

Transformers for Computer Vision

Following the performance of Transformers on NLP tasks, researcher quickly applied this technique to Computer Vision:

- Image Transformer (2018) → Promising
- Unified Vision-Language Pre-Training for Image Captioning and VQA (2019) → Performant
- CLIP (2021) changed → Efficient

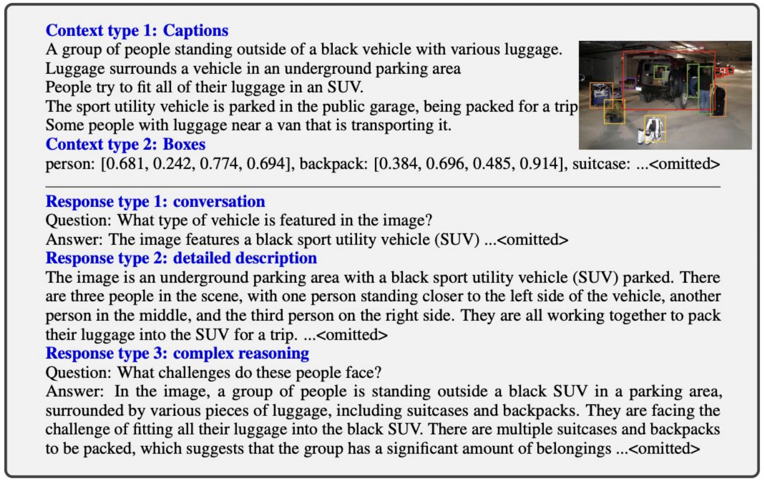


Source: Learning Transferable Visual Models From Natural Language Supervision, Radford et al., 2021

LLaVA: the Sum of LLMs and traditional CV techniques

Introduced in 2023 in *Improved Baselines with Visual Instruction Tuning* by Liu et al., LLaVA is combining text description with boxes and image splitting.

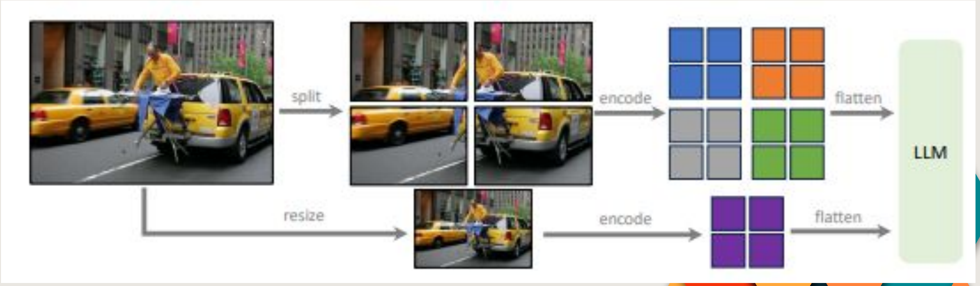
It is using CLIP for image processing and Vicuna (Llama fine-tuned on ShareGPT data). It is available on HuggingFace.



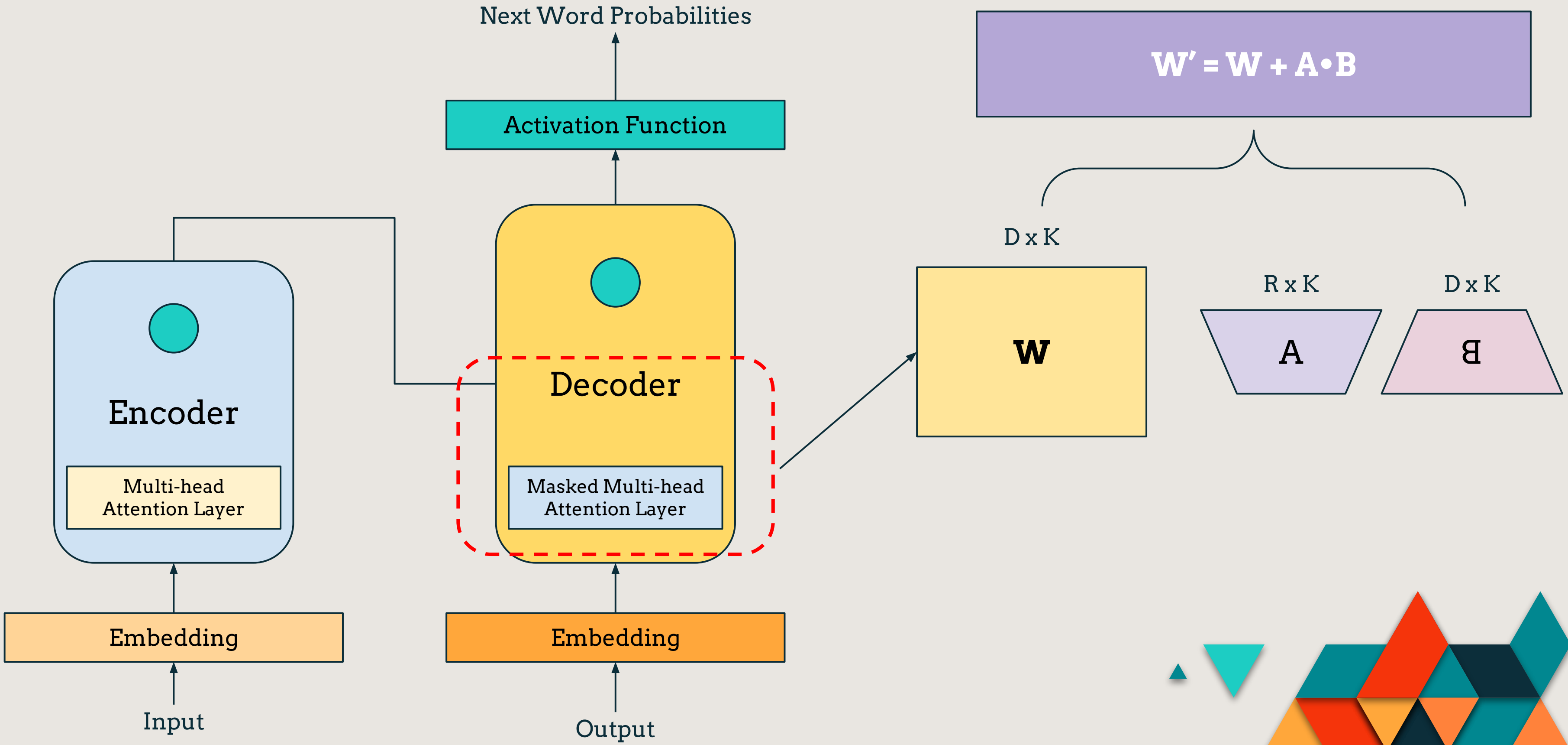
Context type 1: Captions
A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.
Context type 2: Boxes
person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV) ...<omitted>
Response type 2: detailed description
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>
Response type 3: complex reasoning
Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

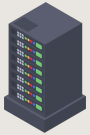
Table 1: One example to illustrate the instruction-following data. The top block shows the contexts such as captions and boxes used to prompt GPT, and the bottom block shows the three types of responses. Note that the visual image is not used to prompt GPT, we only show it here as a reference.



FINE-TUNING WITH PEFT AND LoRA (+ Quantization)



Training Set up and Dataset



As model is very heavy we used:

- LoRA matrix of rank 4 (smallest possible)
- Quantization:
  - Loaded model in 4bit
  - Reduced computation precision to float16
- 2 batches per device (GPU) for both train and test
- Training on only 5 epochs

Our VM included:

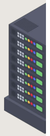
- A CPU with 47 GB of RAM
- An NVIDIA RTX A6000 with 48 GB memory

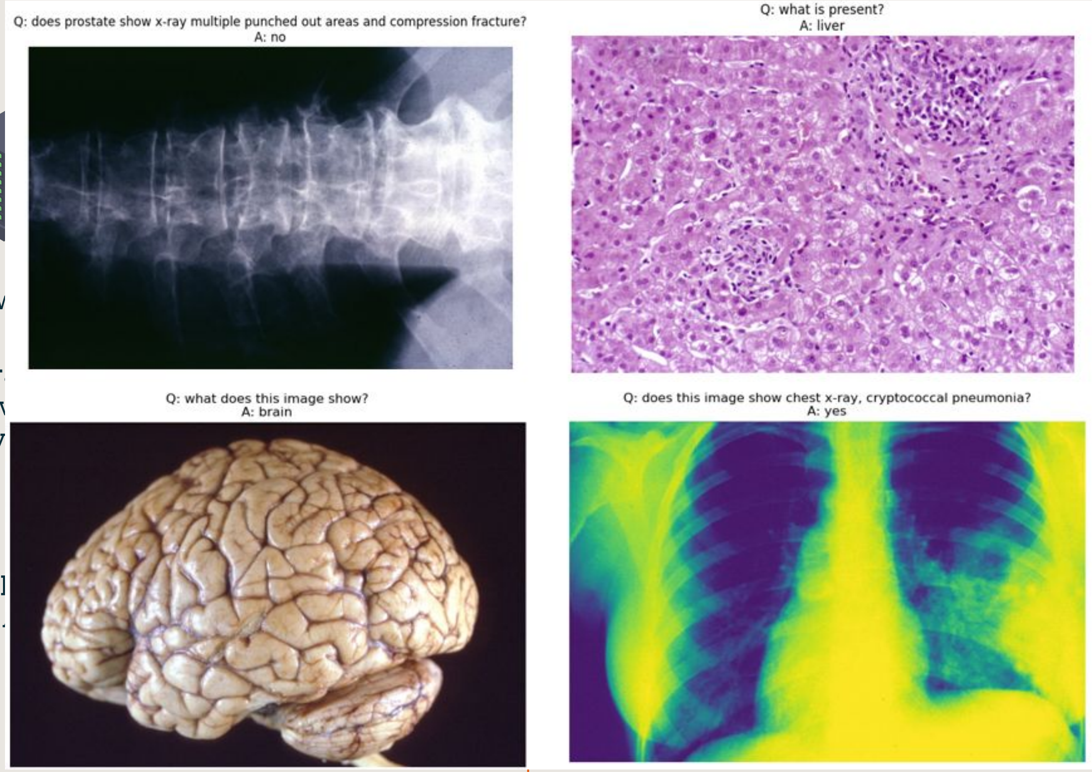
Our dataset is the PATH-VQA. It includes:

- 20k text-image pairs (including answers) for train set
- 6k samples for the test set
- both binary and open-ended questions

Its goal is to simulate a final exam for students aiming to be physicians. Dataset is available here.

Training Set up and Dataset





As model is very heavy w ... ncludes:

- LoRA matrix of r ... uding answers) for train set
- 2 batches per dev
- Training on only ... ed questions

Our VM included: ... for students aiming to be ... re.

- A CPU with 47 G
- An NVIDIA RTX

Model Performance - Generated Outputs

We trained the model on 5 epochs on 1200 samples from the training set (training time ~ 1h30). Validation was made on 200 samples.

Below are some examples of generated output before and after fine-tuning:

```
**************************************************
where is this part in?
Answer from base LLAVA: art in question is the heart, which is a muscular organ located in the chest. It is responsible for pumping blood throughout the body. Ir
--------------------------------------------------
Answer from fine-tuned LLAVA: ch, spleen, and kidney, slide 117?
ASSISTANT: Stomach, spleen, and kidney, slide 117?
--------------------------------------------------
Correct Answer from test_data: spleen
**************************************************
--------------------------------------------------
Example from row 5208
**************************************************
does opened base of skull with brain show pemphigus vulgaris?
Answer from base LLAVA: he image does not show pemphigus vulgaris. The image shows a close-up of a pink tissue, which could be a piece of tissue or a biological
--------------------------------------------------
Answer from fine-tuned LLAVA: opened base of skull with brain shows pemphigus vulgaris.
--------------------------------------------------
Correct Answer from test_data: no
**************************************************
--------------------------------------------------
Example from row 1672
**************************************************
is endocrine present?
Answer from base LLAVA: the image shows a pink tissue with endocrine cells, indicating that the tissue is likely an endocrine gland.
--------------------------------------------------
...
Answer from fine-tuned LLAVA: this image is a very good example of a hemorrhagic infarct with a large area of necrosis surrounded by a normal appearing cortex.
--------------------------------------------------
Correct Answer from test_data: no
**************************************************
```

Model Performance - ROUGE SCORE

As we've seen, the model is very verbose, which makes it hard to compute properly a ROUGE score, especially with binary questions. Therefore we applied the below strategy:

- Define a threshold. If Rouge-L score is above that threshold, we consider the prompt as a good answer
- If "yes" or "no" is in the correct answer and also in the generated output, we consider it as a good answer as well
- Finally, we compute the accuracy by dividing correct predictions with total questions

With a threshold of **0.5** (considered as state-of-the-art) we obtain an accuracy of **30%**.

If we set threshold to **0.3** we get an accuracy of **67%**. This highlight the need for **additional supervision from physicians** during evaluation process.

Model Performance - Brain Tumor Dataset

Following this training we called the model on the tumor dataset to see what it could output:

- Unfortunately model fails to generalize… having cancer would be like flipping a coin…
- This shows the need for a dedicated dataset with data properly formatted for model training
- However, when model is right, it can sometimes provide what seem to be good explanations
- With proper training, could we build a personalized AI assistant for physicians?

```
------------------------------
Image 65:
Predicted: there is a tumor present on this image.
Actual: 1
------------------------------
Image 66:
Predicted: a tumor is present in this image.
Actual: 0
------------------------------
Image 67:
Predicted: a tumor is present on this image.
Actual: 0
------------------------------
Image 68:
Predicted: a tumor is present in this image, which is a close-up view of the skull base showing a large tumor in the middle of the skull.
Actual: 1
------------------------------
```

Thank You