Interpretation multimodal LLMs trained for medical domain

FINAL PROJECT

Authors: Iana Kulichenko,
Rinat Prochii,
Fedor Gubanov

TA: Polina Druzhinina

Supervisor: Maxim Sharaev

Team name: LLM Neuroimaging

PROJECT · PROJECT · PROJECT · PROJECT · PR

PROJECT · PROJECT · PROJECT · P

01

Problem description and solution

Problem

Multimodal AI models like LLaVA-Med, which integrate medical text and images, are difficult to interpret, posing risks to transparency and trust in clinical decision-making. Without clear understanding of how these models process and combine different inputs, their application in healthcare is limited.

Solution

This project focuses on interpreting the internal mechanisms of the LLaVA-Med model by applying classical interpretation methods and analyzing model representations (e.g., linearity, contextualization). This will enhance transparency and ensure safer, more reliable use in medical research and diagnostics.

02. Project Relevance and Importance

The growing use of AI in healthcare requires models to be transparent and reliable, especially when they are used for medical decision-making. Multimodal models like LLaVA-Med, which integrate both medical images and text, need to be interpreted clearly to ensure trust and safety in clinical practice. This project is crucial because it aims to improve the understanding of how these models work internally, which will enhance their transparency, reliability, and usability in medical research and diagnostics.

PROJECT · PROJECT · PROJECT · P

03

Goal and Objectives

Goal

The main goal of this project is to improve the interpretability of the multimodal LLaVA-Med model, ensuring its safe and transparent application in medical research and diagnostics.

Objectives

1) Evaluate the model using the LLaVA-Med dataset to assess performance on multimodal medical data.

2) Apply interpretation techniques like knowledge neurons to analyze how the model processes instructions and responses.

3) Analyze internal model properties like linearity and contextualization.

4) Explore how multimodal neurons manage text and images.

5) Report findings and propose improvements.

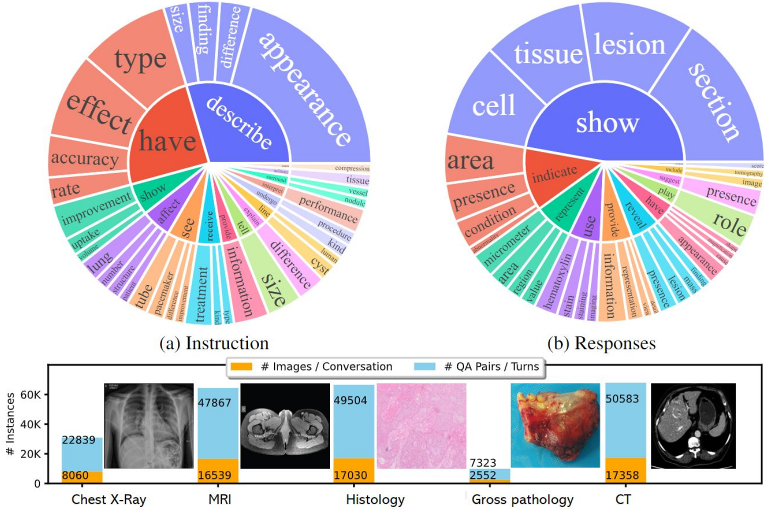
PROJECT · PROJECT · PROJECT · P

04

Dataset

LLaVA-Med Dataset

The is a multimodal biomedical dataset created to train AI models for instruction-following tasks. It includes 600K image-text pairs for concept alignment (image captioning) and 60K pairs for instruction-tuning (multi-round conversations). The data covers five key medical imaging modalities: CXR, CT, MRI, histopathology, and gross pathology. **LLaVA-Med uses GPT-4 to generate diverse biomedical multimodal instruction-following data using image-text pairs from PMC-15M.** This dataset enables models to perform tasks like VQA and medical image interpretation.



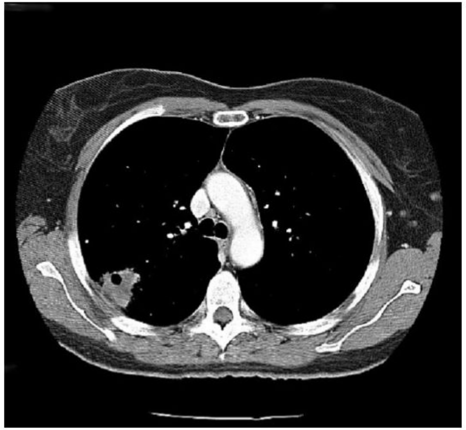
(a) Instruction (b) Responses

(c) Frequencies of images and QA pairs on the five domains.

Biomedical Visual Instruction-Following Example

Caption:

Figure 4: Contrast-enhanced CT scan of the chest for patient #1. A large cavitating lesion was identified in the right upper lobe with pleural involvement and likely extension into the chest wall. In addition, right hilar lymphadenopathy was suspected. This was the primary malignant lesion in this patient.



Sentences with in-line figure mentions:

- Computerized tomography (CT) scans of the chest (Figure 4) were obtained.
- (Figure 4) and demonstrated a large cavitating lesion in the posterior aspect of the right upper lobe with probable pleural involvement likely representing the primary lesion.

Conversation:

User: Where is the lesion located?

Assistant: The lesion is located in the posterior aspect of the right upper lobe of the lung.

User: Is it extending into the chest wall?

Assistant: Yes, the CT scan suggests that the large cavitating lesion in the right upper lobe is likely extending into the chest wall.

User: Can you explain more about the cavitating lesion?

Assistant: Certainly. A cavitating lesion refers to an abnormal mass or growth within the lung tissue that has a cavity or hollow space inside it. This cavity can be formed due to the necrosis (death) of the tissue inside the lesion, leading to the breakdown and disintegration of the affected tissue. Cavitating lesions can be caused by various factors, such as infections (like tuberculosis or fungal infections), autoimmune diseases, or malignancies (cancer). In the context of the CT scan provided, the large cavitating lesion in the right upper lobe of the lung is suggestive of a malignant (cancerous) growth. The cavity within the lesion may be a result of the rapid and uncontrolled growth of cancer cells, which can lead to the destruction of surrounding healthy tissue. It's important to note that further diagnostic tests, such as a biopsy, would be required to confirm the malignancy and determine the specific type of cancer involved.

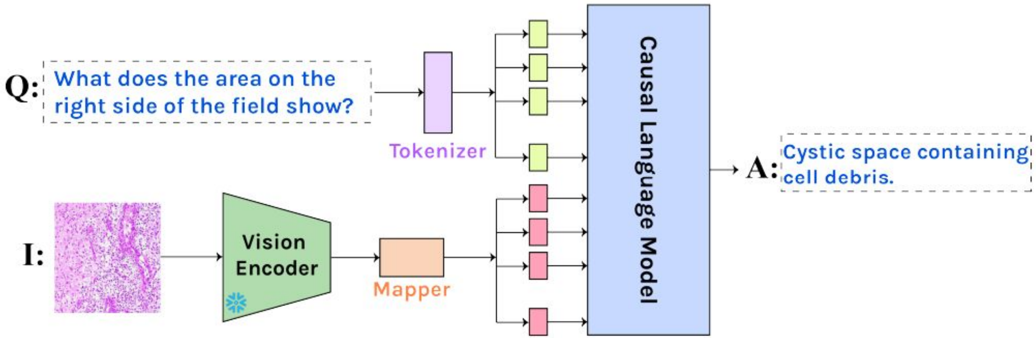
PROJECT · PROJECT · PROJECT · P

05

Pipeline

LLaVA-Med Architecture

VQA model is designed as an encoder-decoder architecture, with a two stream encoder and LM as a decoder. Two streams encode the two input modalities, namely the image I and the question Q. The language model generates the answer A in an autoregressive manner. **Linear projector maps the visual features into a visual prefix for the language model.** It closely follows the prefix tuning technique for prompting a language model to produce an output of a particular style.



06. Tasks and Expected Outcomes

Tasks:

1. Evaluate the pre-trained LLaVA-Med model on open datasets.
2. Apply interpretation methods to gain insights into model behavior.
3. Analyze internal representations like linearity and contextualization.
4. Explore multimodal neurons and their interaction across domains.

Expected Results:

1. Improved understanding of LLaVA-Med model structure and decision-making process.
2. Deeper insights into the interaction between medical text and images.
3. Recommendations for enhancing the model's interpretability in medical applications.

07. Interpretation methods

1. **Anisotropy** - quantitative measure that describes the extent to which a point cloud is stretched or elongated in a particular direction, indicating the degree of heterogeneity in the distribution.

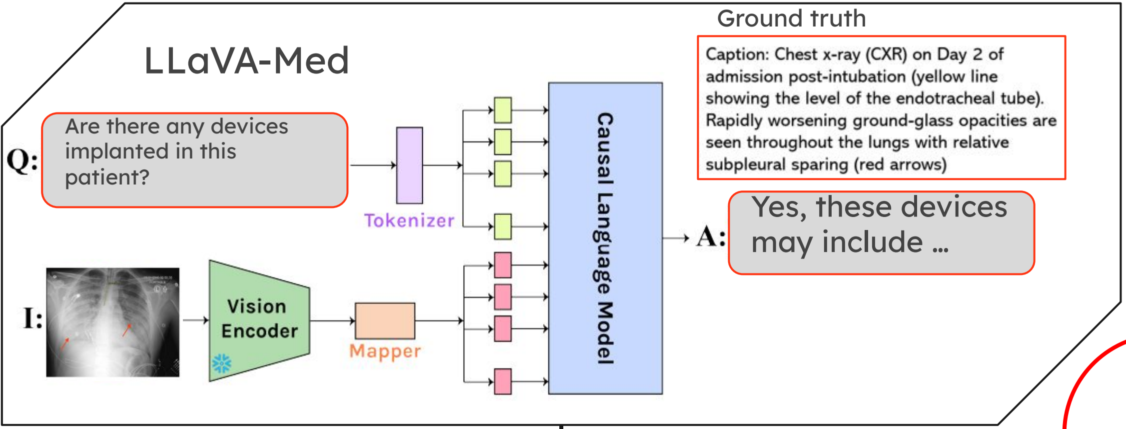
2. **Intrinsic dimension** - measure of the effective data dimensionality, highlighting the essence of information captured by the embeddings.

3. **Contextualisation** - assesses how language models dynamically adjust embeddings based on surrounding text, revealing the model's capacity to capture and represent varying semantic nuances across different contexts.

4. **Linearity score** - characterizes degree of linearity in embedding transformations between sequential layers using Procrustes Similarity.

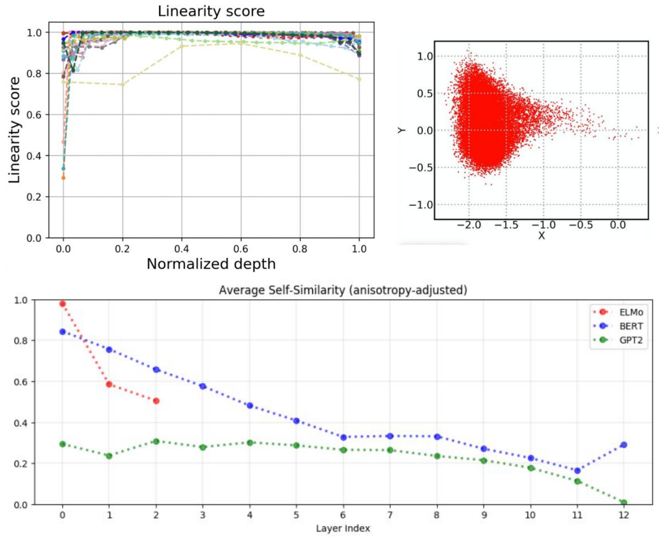
5. **Multimodal neurons** - specialized neurons that convert visual representations into corresponding text. These neurons can integrate and process information from diverse sources, allowing the model to recognize and interpret complex, multimodal patterns.

08. Methodology



Interpretation methods:
anisotropy, intrinsic dimension, contextualisation, linearity score

Results:



PROJECT · PROJECT · PROJECT · PROJECT · P

09. Literature Overview

1) LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day

- Trains a multimodal model using PMC-15M and GPT-4, achieving top results in biomedical VQA.
- Relevance: Key foundation for our project to interpret and improve LLaVA-Med's transparency.

2) Your Transformer is Secretly Linear

- Shows high linearity between transformer layers, which can be reduced for better performance.
- Relevance: Helps us analyze how LLaVA-Med processes text and image data.

3) Multimodal Neurons in Pretrained Text-Only Transformers

- Identifies neurons converting visual inputs into text in transformers.
- Relevance: Supports our exploration of how LLaVA-Med integrates visual and textual data.

4) From Redundancy to Relevance

- Proposes reducing redundant image tokens to improve efficiency.
- Relevance: Useful for enhancing LLaVA-Med's efficiency and interpretability.

PROJECT · PROJECT · PROJECT · P

Model Validation & Setup

10

Progress

Model: LLAVA-Med-v1.5-mistral-7b

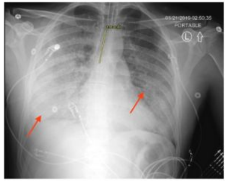
Environment setup: We prepared and validated the model independently to ensure correctness.

Inference: Completed inference testing using the notebook. Results match expectations. Dialog interface with a question, image, and model response.

Example Output:

Got file answer-file_test.jsonl with answers in format:

{"question_id": 0, "prompt": "What is the organ in the image?", "text": "The organ in the image is the gallbladder, which is a small, pear-shaped organ located beneath the liver in the upper right abdomen.", "answer_id": "GoStcLQWFigAzMuMZnexu5", "model_id": "llava-med-v1.5-mistral-7b", "metadata": {}}



Caption: Chest x-ray (CXR) on Day 2 of admission post-intubation (yellow line showing the level of the endotracheal tube). Rapidly worsening ground-glass opacities are seen throughout the lungs with relative subpleural sparing (red arrows)

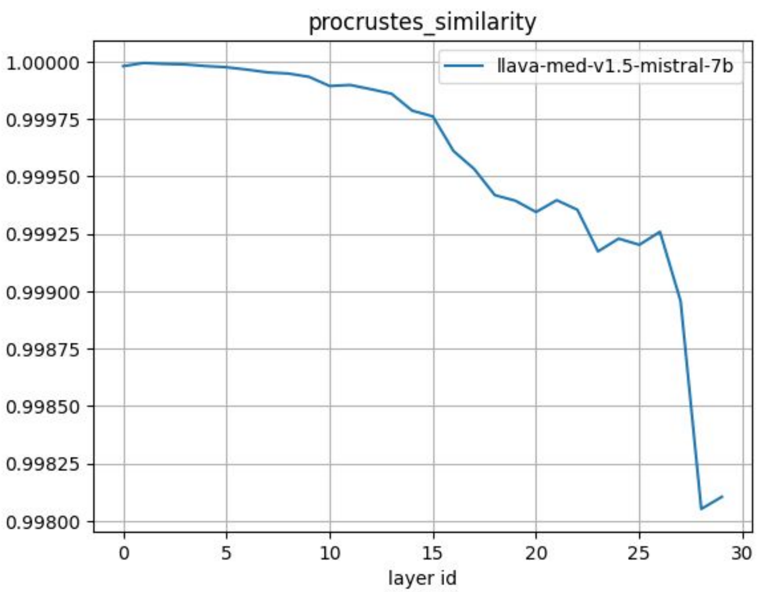
Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7398722/

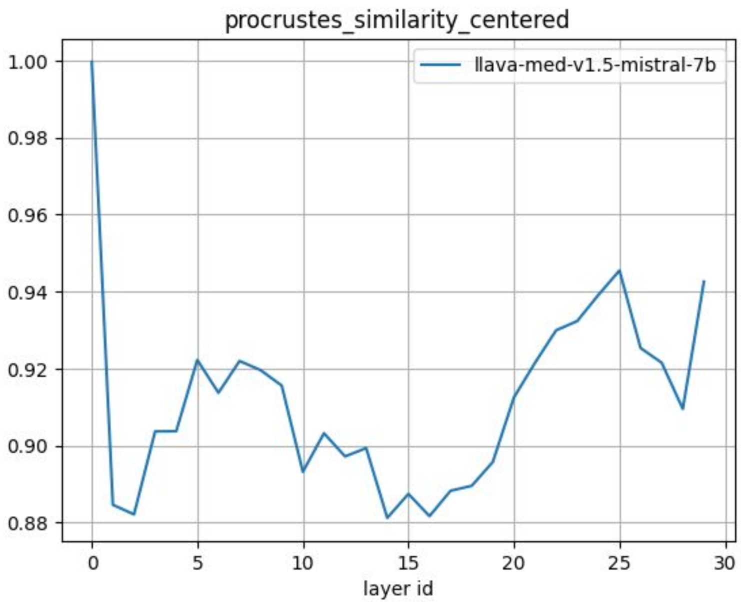
User	What is shown in this image?
LLaVA-Med	The image is a chest X-ray (CXR) that shows bilateral patchy infiltrates, which are areas of increased opacity in the lungs. These infiltrates can be indicative of various lung conditions, such as infections, inflammation, or other lung diseases.
User	Are there any devices implanted in this patient?
LLaVA-Med	Yes, it appears that the patient has several devices implanted. These devices may include endotracheal tubes, central venous catheters, and/or pulmonary artery catheters. These devices are typically used for monitoring and managing various medical conditions, as well as for providing support during surgical procedures.

10. Linear score (full embeddings)

Embedding collection: For 64 validation examples, embeddings (~600*64 x 4096 matrix) were collected from all layers.
Statistics: Procrustes similarity, Procrustes similarity centered

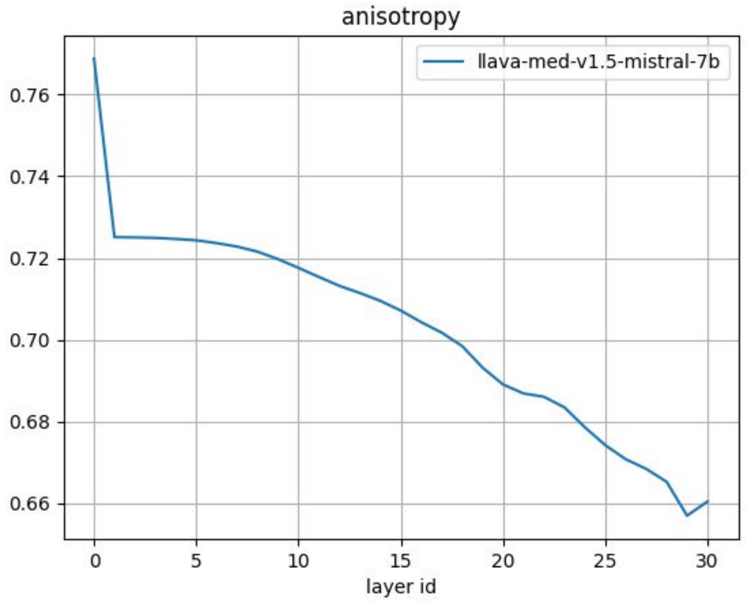
$$\text{linearity\_score} := 1 - \min_{A \in R^{d \times d}} ||\tilde{X}A - \tilde{Y}||_2^2$$

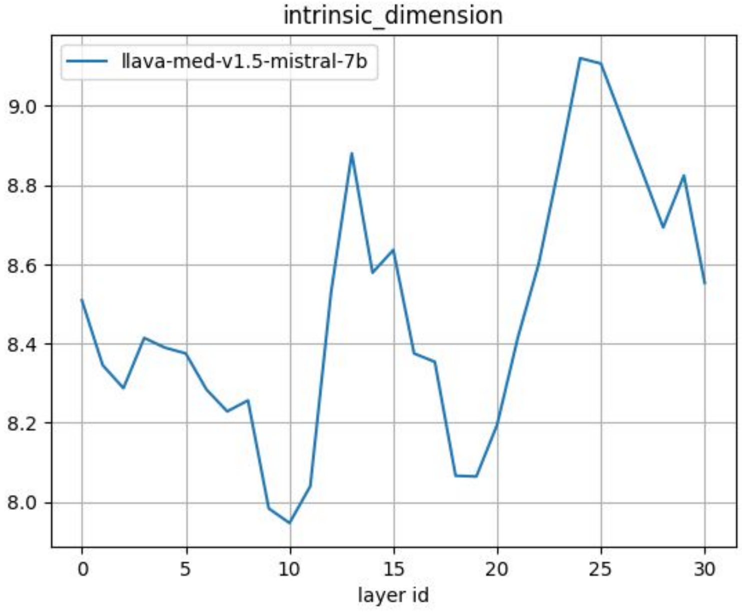




11. Anisotropy & Intrinsic dimension (full embeddings)

Anisotropy: Uniformity of embeddings across layers, showing gradual increase
Intrinsic dimension: Fluctuation between 8 - 9 values





PROJECT · PROJECT · PROJECT · PROJECT · P

12. Linear score (text / img)

Embedding preparation: Split embeddings related to image and text tokens

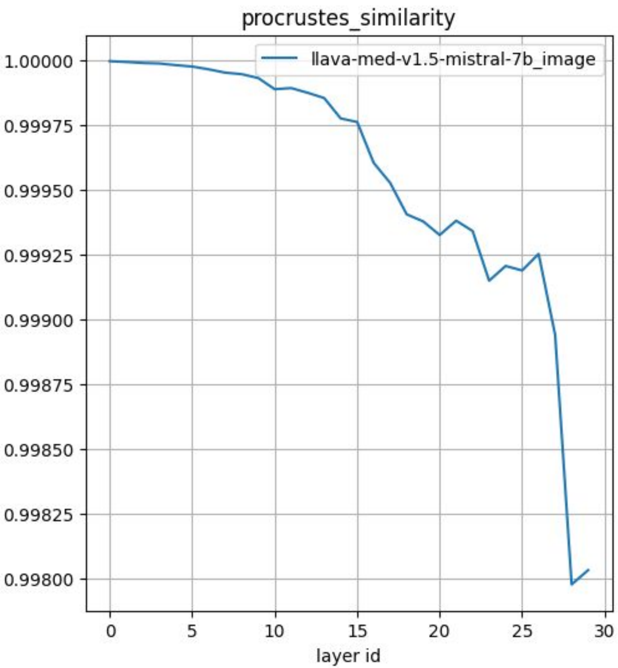
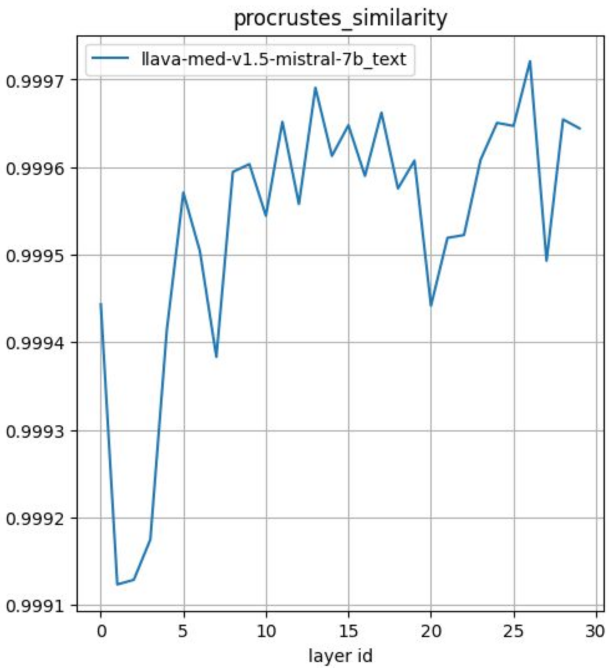


Image Tokens



Text Tokens

13. Linear score (text / img)

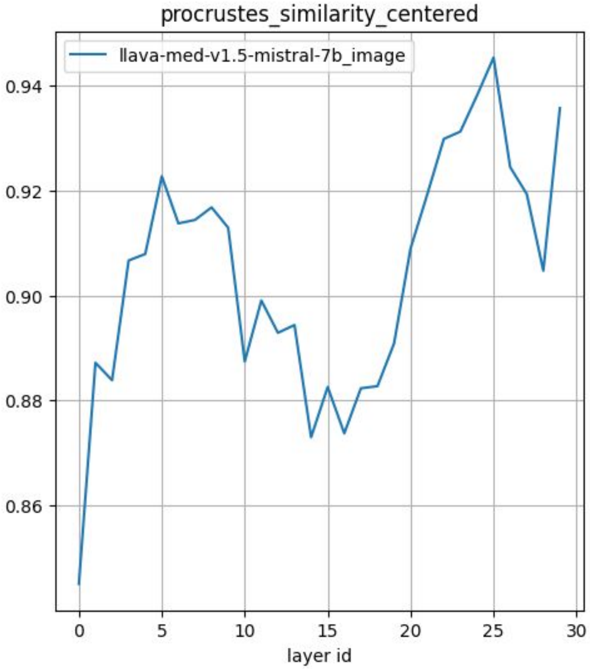
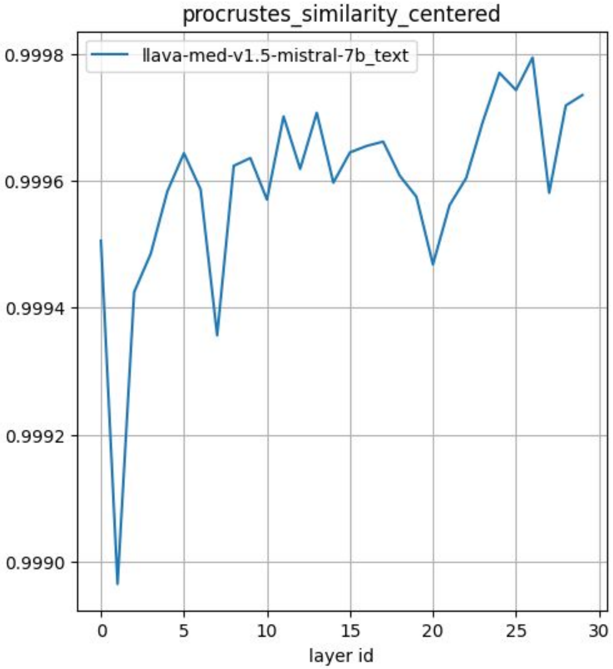


Image Tokens



Text Tokens

PROJECT · PROJECT · PROJECT · P

14. Anisotropy (text / img)

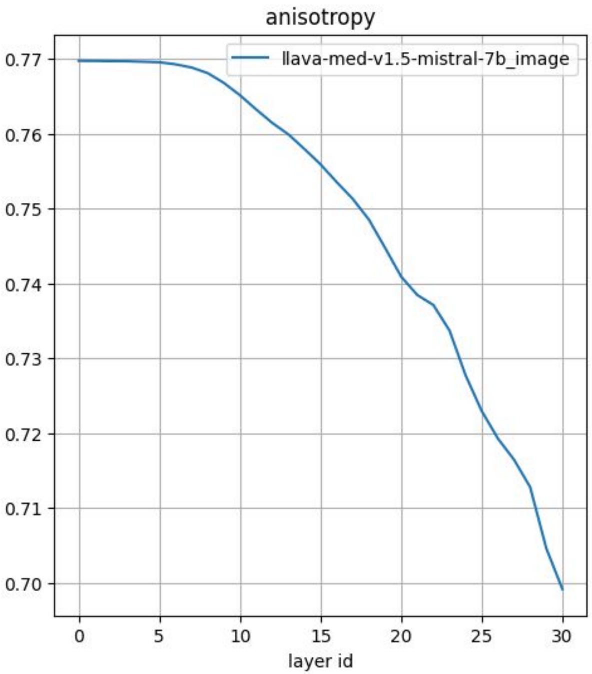
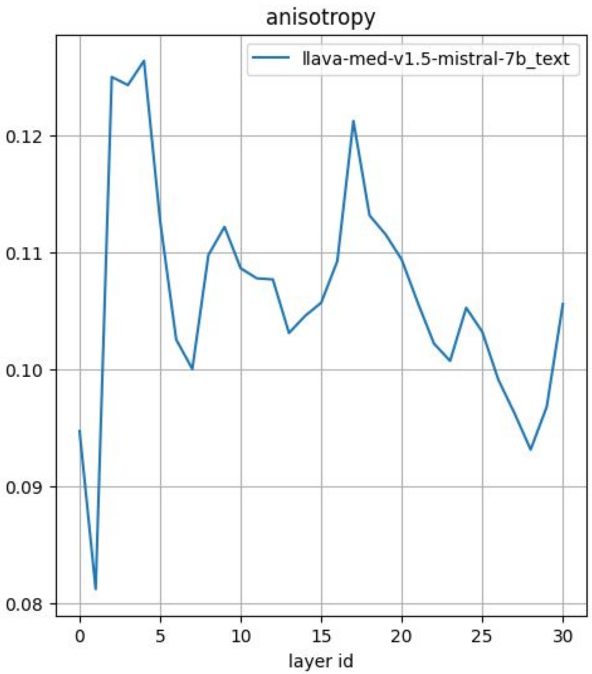


Image Tokens



Text Tokens

15. Intrinsic dimension (text/ img)

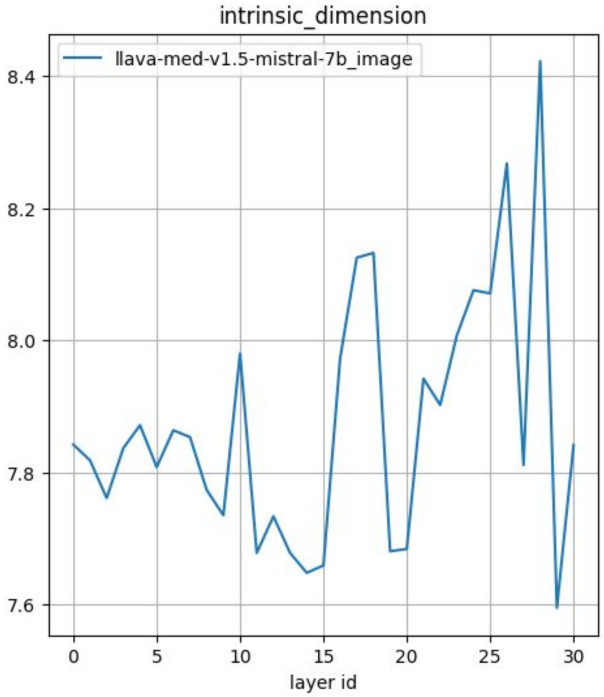
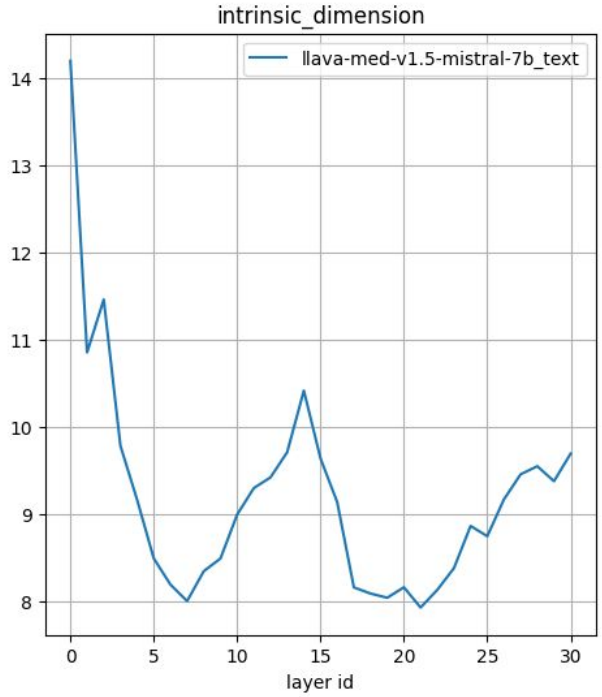
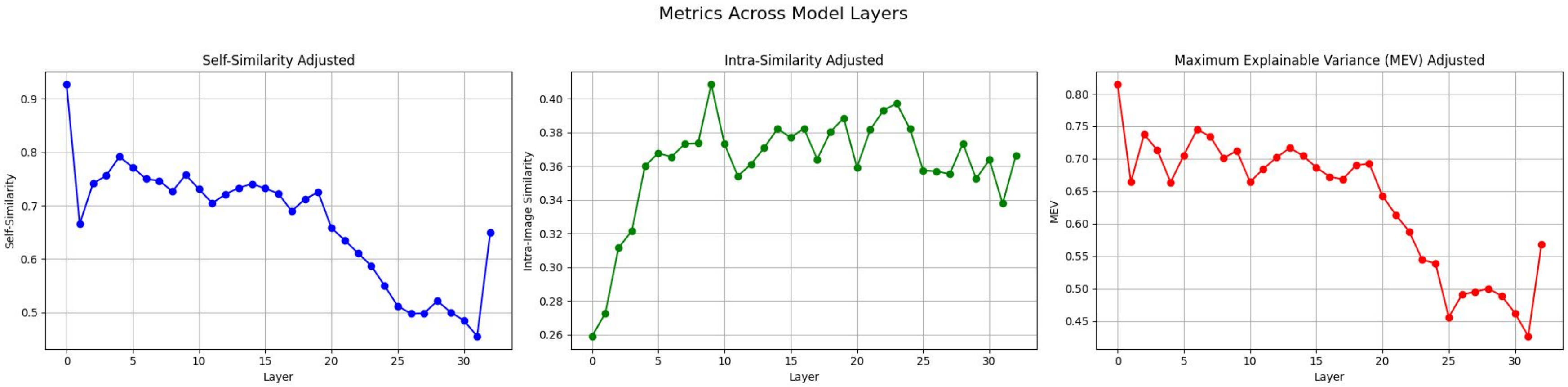


Image Tokens



Text Tokens

16. Contextualization



$$SelfSim_\ell(w) = \frac{1}{n^2 - n} \sum_j \sum_{k \neq j} \cos(f_\ell(s_j, i_j), f_\ell(s_k, i_k))$$

$$IntraSim_\ell(s) = \frac{1}{n} \sum_i \cos(\vec{s}_\ell, f_\ell(s, i))$$

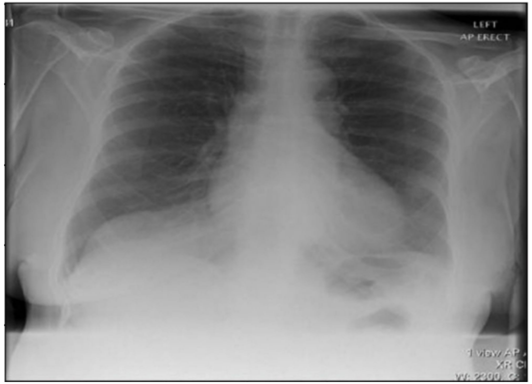
$$\text{where } \vec{s}_\ell = \frac{1}{n} \sum_i f_\ell(s, i)$$

$$MEV_\ell(w) = \frac{\sigma_1^2}{\sum_i \sigma_i^2}$$

PROJECT · PROJECT · PROJECT · P

17. Image Tokens Occurrence

Images



Top 10 Tokens by Occurrence

5th Layer

```
most_frequent_words

[('_kennis', 440),
 ('_/******/', 410),
 ('_‘', 233),
 ('_ingår', 170),
 ('_[...]', 133),
 ('_sep', 66),
 ('_scan', 35),
 ('_database', 26),
 ('_FBI', 25),
 ('_cards', 24)]
```

30th Layer

```
most_frequent_words

[('_/******/', 1071),
 ('_', 302),
 ('_bekan', 292),
 ('_Fuß', 216),
 ('_kennis', 170),
 ('_Hug', 168),
 ('_question', 154),
 ('_PARTICULAR', 105)
 ('_the', 90),
 ('_hint', 79)]
```

5th Layer

```
most_frequent_words

[('_kennis', 522),
 ('_/******/', 337),
 ('_ingår', 296),
 ('_‘', 153),
 ('_sep', 122),
 ('_[...]', 116),
 ('_therefore', 93),
 ('_surgery', 70),
 ('_Nah', 55),
 ('_patients', 48)]
```

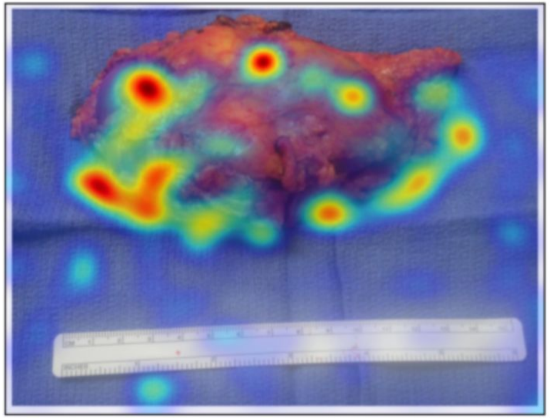
30th Layer

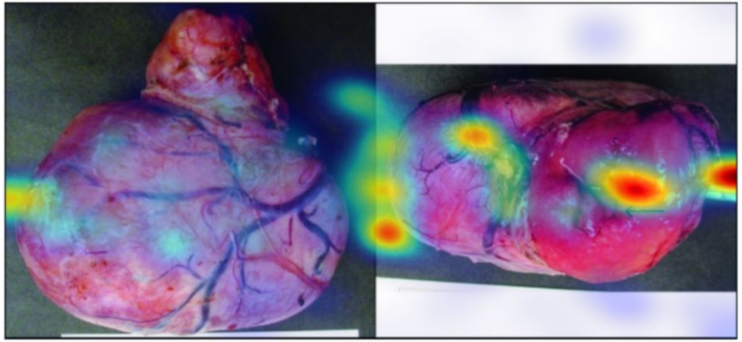
```
most_frequent_words

[('_/******/', 1012)
 ('_bekan', 255),
 ('_Fuß', 224),
 ('_', 216),
 ('_kennis', 165),
 ('_Hug', 145),
 ('_IC', 115),
 ('_question', 110),
 ('_PARTICULAR', 73)
 ('_Pred', 66)]
```

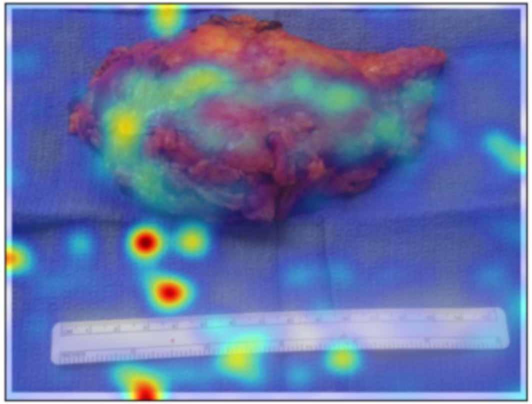
PROJECT · PROJECT · PROJECT · P

18. GradCam Interpretation (Question)

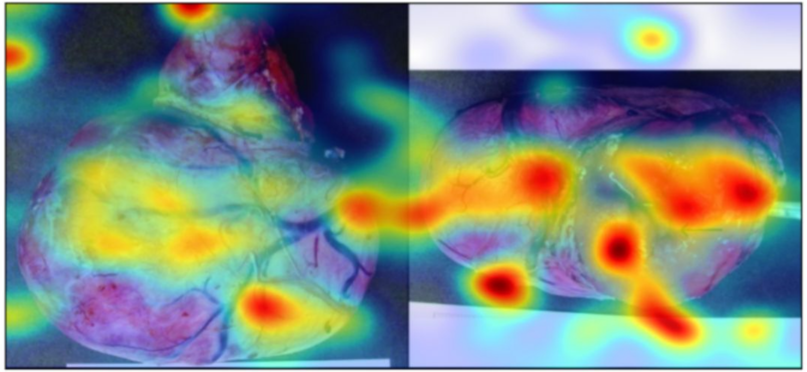




19. GradCam Interpretation (Predict)



</s>



</s>

PROJECT · PROJECT · PROJECT · P

20. Conclusion

- Analyzed embedding space and properties in a multimodal medical model, focusing on linearity scores across layers, anisotropy, and intrinsic dimension
- Explored embedding contextualization abilities and Grad-CAM, with initial study on multimodal neurons
- Key findings:
  - The model is not overfitted after fine-tuning on VQA medical data; rather, visual embeddings may be undertrained compared to text embeddings
  - Contextualization capabilities improve across deeper layers
  - Grad-CAM shows the model’s focus on relevant regions in medical images
- Presented approaches require deeper research

PROJECT · PROJECT · PROJECT · P

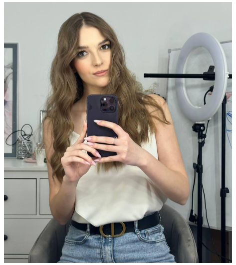
PROJECT · PROJECT · PROJECT · PROJECT · P

21

Project Team

Fedor Gubanov: Experienced in deep learning (torch), focusing on literature review and model interpretation.

Rinat Prochii: Expertise in ultrasound medical images, bash, git, docker, working on environment setup and model evaluation.



Iana Kulichenko: Skilled in deep learning (torch), running validations, checking metrics, and preparing the report.

Thank you

PROJECT · PROJECT · PROJECT · P

PROJECT · PROJECT · PROJECT · P