MediConfusion: Can you trust your AI radiologist?

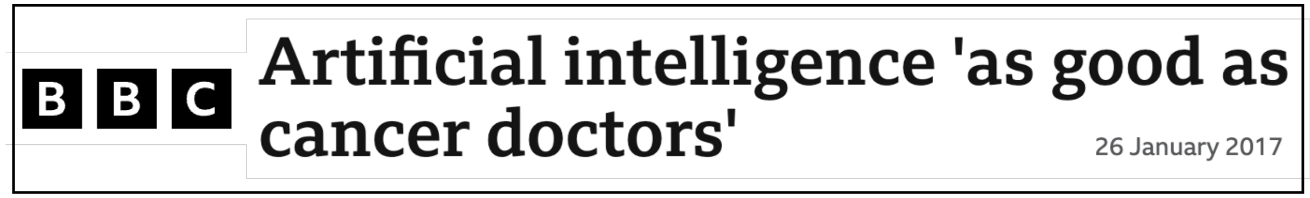
USC Center on AI Foundations for the Sciences (AIF4S)

ICLR
International Conference On Learning Representations

Mohammad Shahab Sepehri, Zalan Fabian, Maryam Soltanolkotabi, Mahdi Soltanolkotabi

Is AI the future of health care?

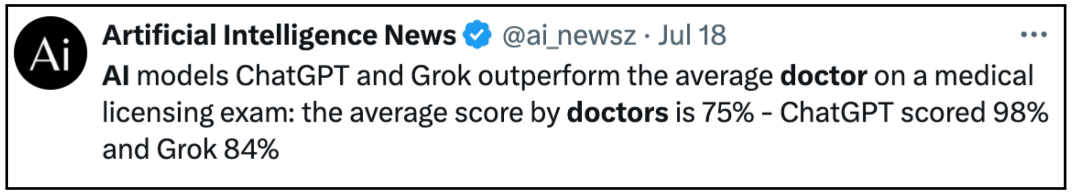
Recently AI models has achieved impressive performance



BBC

Artificial intelligence 'as good as cancer doctors'

26 January 2017



Artificial Intelligence News @ai_newsz · Jul 18

AI models ChatGPT and Grok outperform the average **doctor** on a medical licensing exam: the average score by **doctors** is 75% - ChatGPT scored 98% and Grok 84%

But there is still some concerns

HAI Stanford University Human-Centered Artificial Intelligence

The Shaky Foundations of Foundation Models in Healthcare

Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging

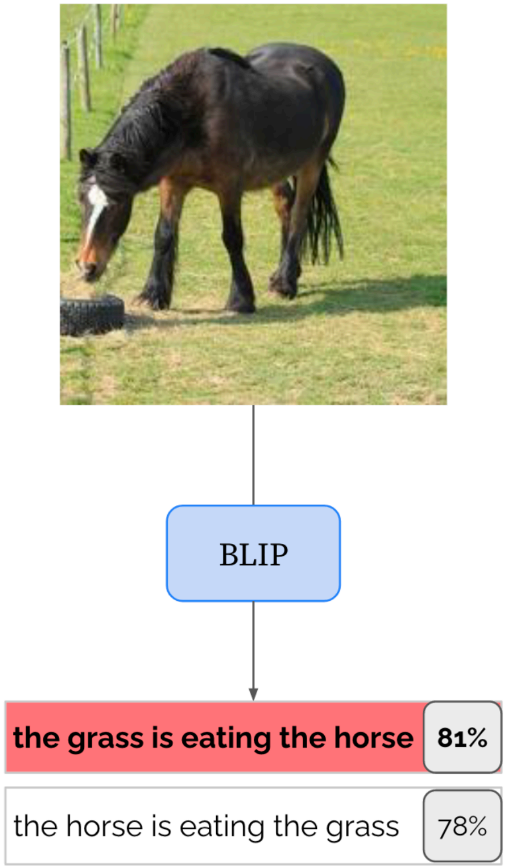
Limitations of Existing MLLMs

Known issues with MLLMs visual encoders

Detecting relations between objects

Capturing spatial information

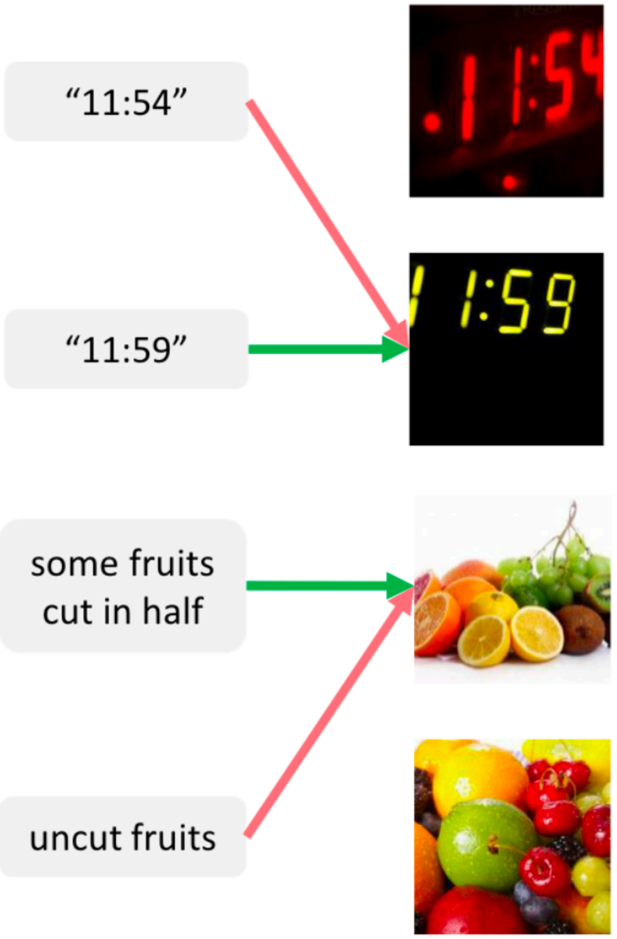
...



BLIP

the grass is eating the horse 81%

the horse is eating the grass 78%



"11:54"

"11:59"

some fruits cut in half

uncut fruits

Idea

Search for images with similar encoding but clear visual differences

New Eval Benchmark: MediConfusion

0	Gemini 2	% 28.41
1	Random Guessing	% 25
2	o1	% 24.43
3	Gemini 1.5 Pro	% 19.89
4	GPT-4o	% 18.75
5	Llama 3.2	% 15.34
6	InstructBLIP	% 12.50
7	Molmo 2	% 9.66
8	LLaVA	% 9.09
9	Claude 3 Opus	% 8.52
10	BLIP-2	% 6.82
11	Molmo 72B	% 6.82
12	RadFM	% 5.68
13	Med-Flamingo	% 4.55
14	LLaVA-Med	% 1.14

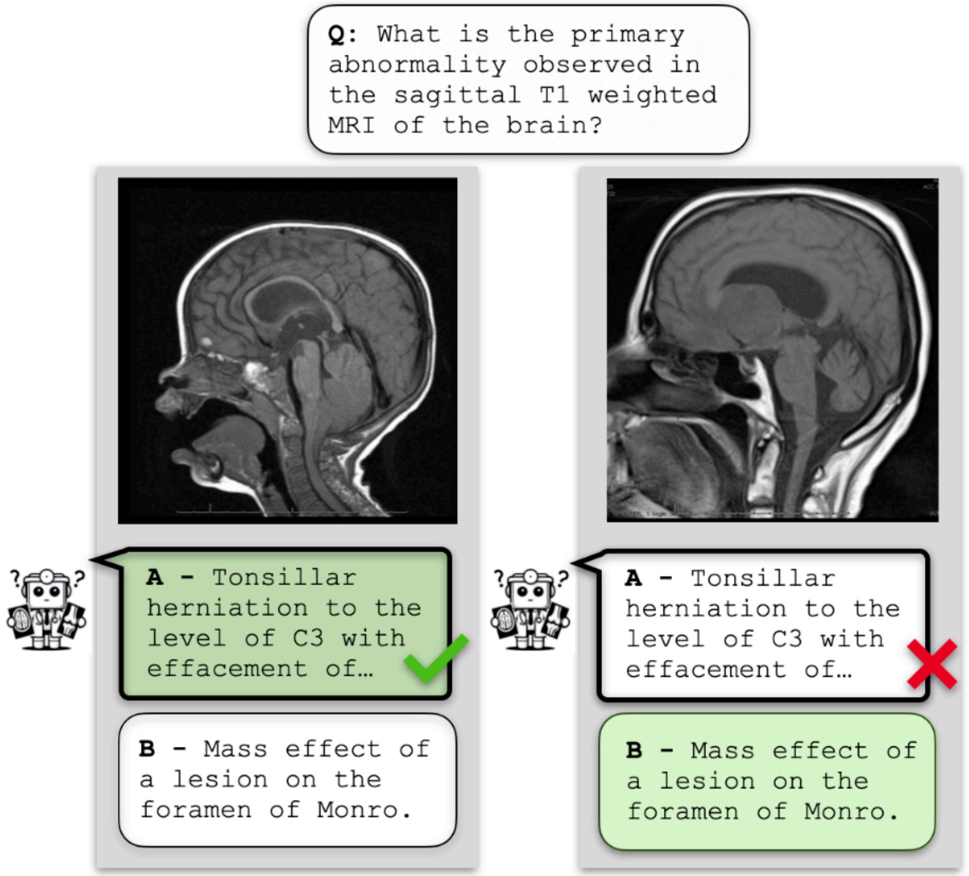
⚠

AI's performance is worse than random guessing!

How does MediConfusion work?

One question with two options
Two confusing images
Different answers

Indiv. score: total correct answers
Confusion: samples with the same answers
Set score: Correct answer to both



Q: What is the primary abnormality observed in the sagittal T1 weighted MRI of the brain?

A - Tonsillar herniation to the level of C3 with effacement of… ✓

B - Mass effect of a lesion on the foramen of Monro.

A - Tonsillar herniation to the level of C3 with effacement of… ✗

B - Mass effect of a lesion on the foramen of Monro.

Individual score: 1
Confusion: 1
Set score: 0

The idea behind finding image pairs

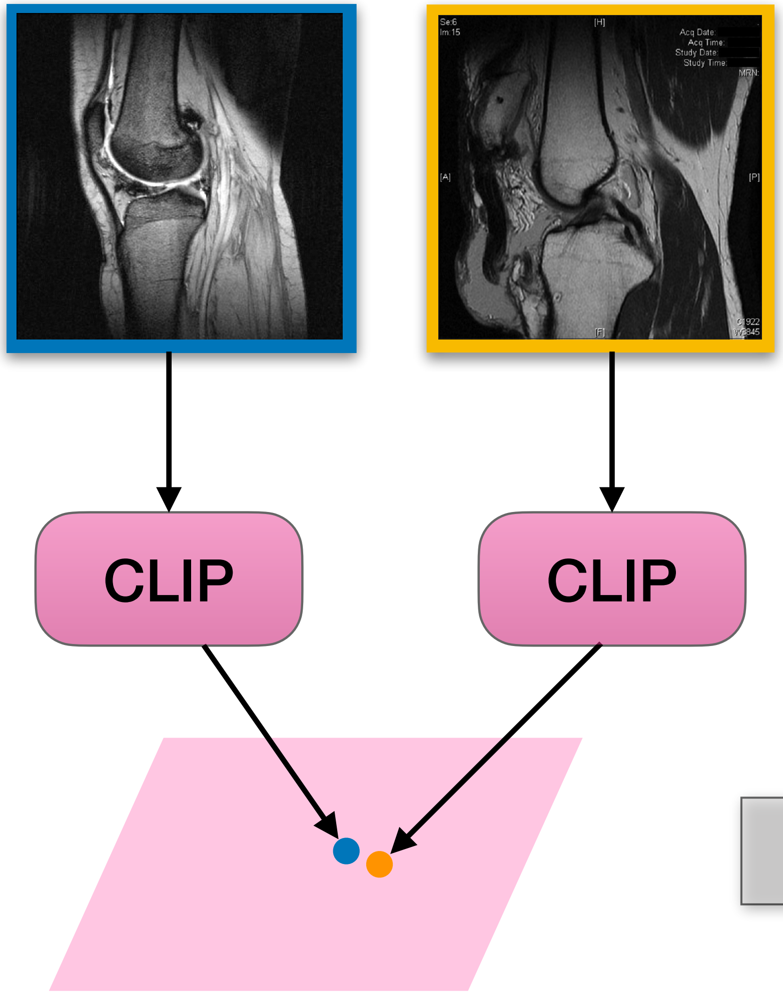
Background: CLIP

Provides embeddings for text and image

Image encoder of many MLLMs

Trained with a contrastive loss to align text and image embedding

Clearly different images



Highly similar encoding

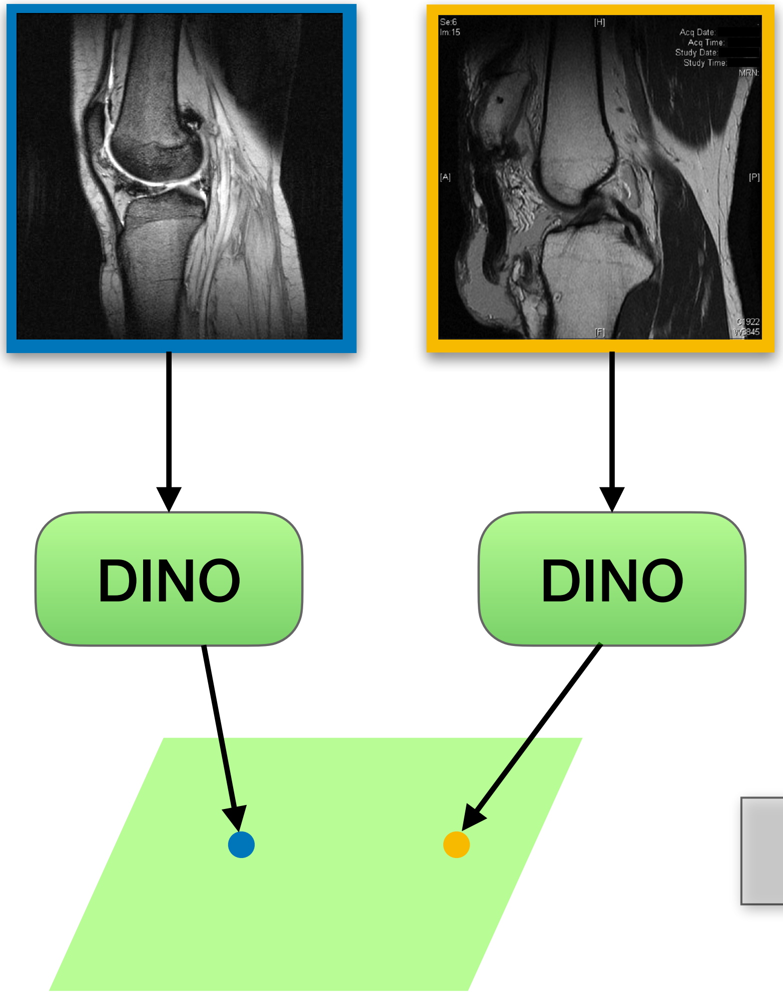
BioMedCLIP: Finetuned for medical applications

Background: DINO

Provides robust image representations

Its embeddings can capture visual details

Clearly different images



DINO

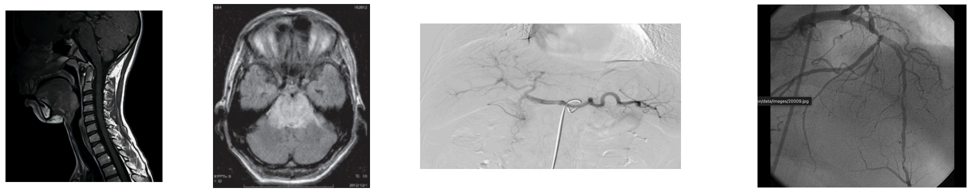
DINO

Dissimilar encodings

Discovering confusing pairs

Pick a dataset

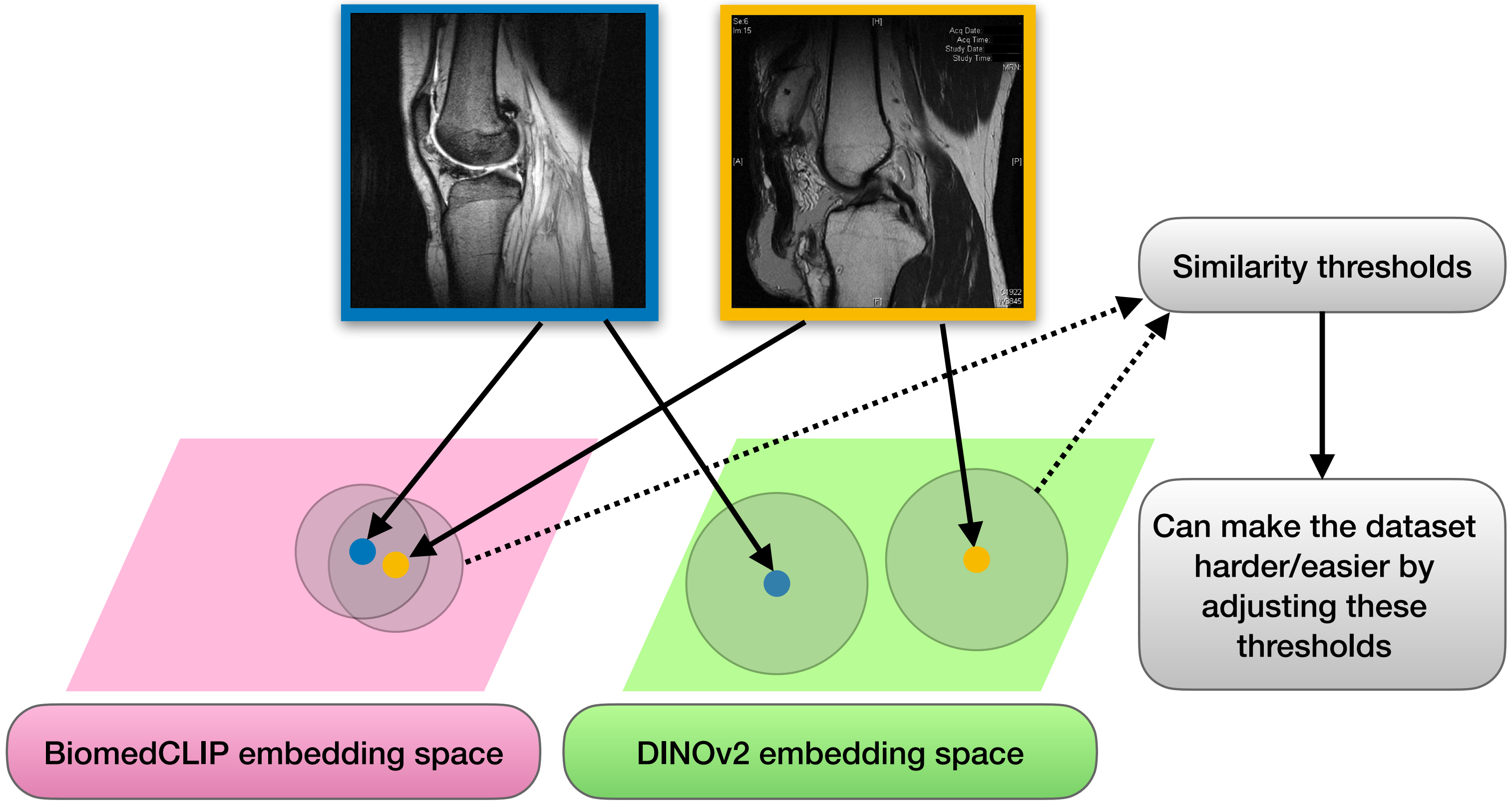
ROCOv2



Search for images with:

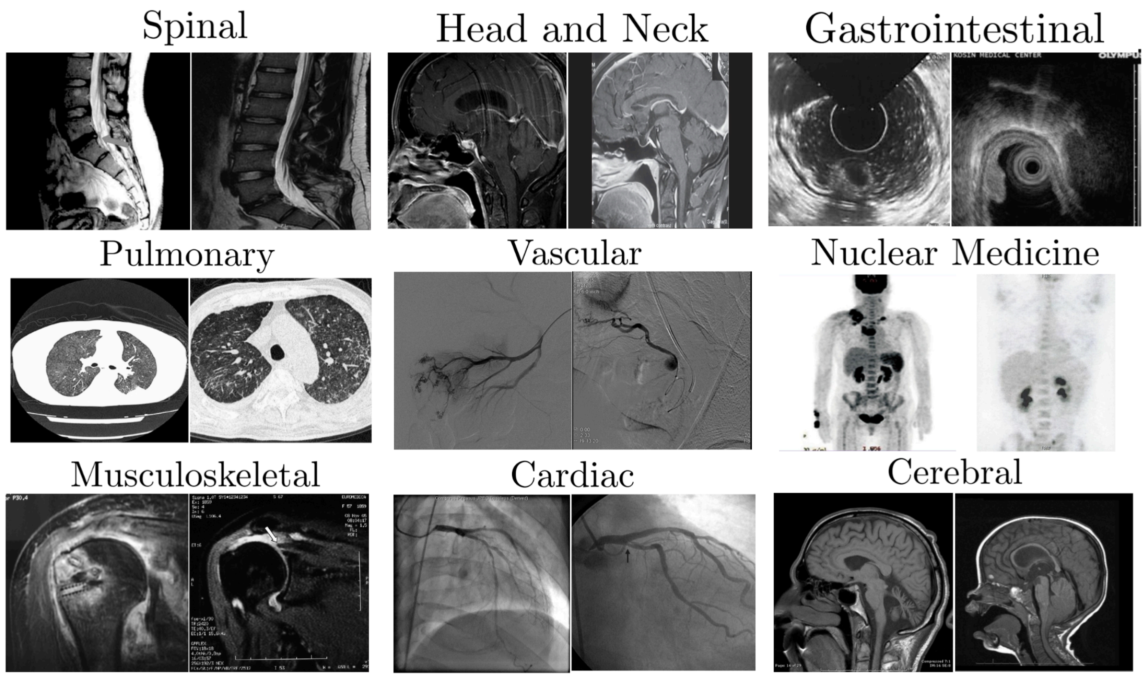
Similar encoding

Clear visual differences



Similarity thresholds

Can make the dataset harder/easier by adjusting these thresholds



VQA Generation

Image captions

Gradient-echo-based MRI from a patient with recurrent t-GCT…

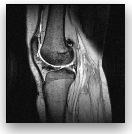
Sagittal MRI scan showing complete rupture of the …

Prompt

Your task is to create a two-choice question based on the above captions for which the answer is different for the two images.

LLM

Q: What is the primary pathology in the MRI scan?



A: Recurrent t-GCT … ✓ ✗

B: Complete rupture of … ✗ ✓

Radiologist feedback

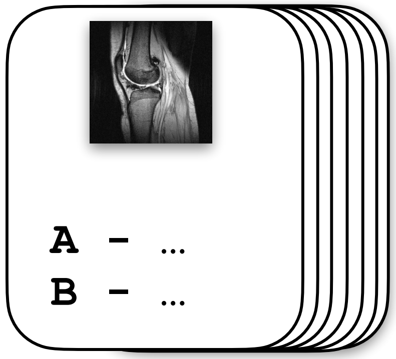
We need to filter the questions

Quality

Correctness

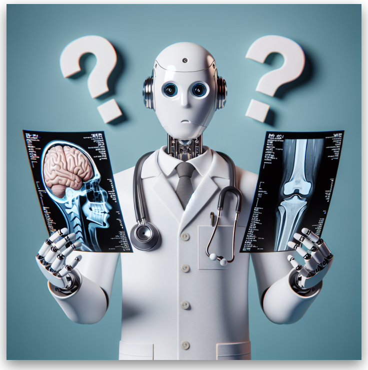
Relevance

Confusing pairs



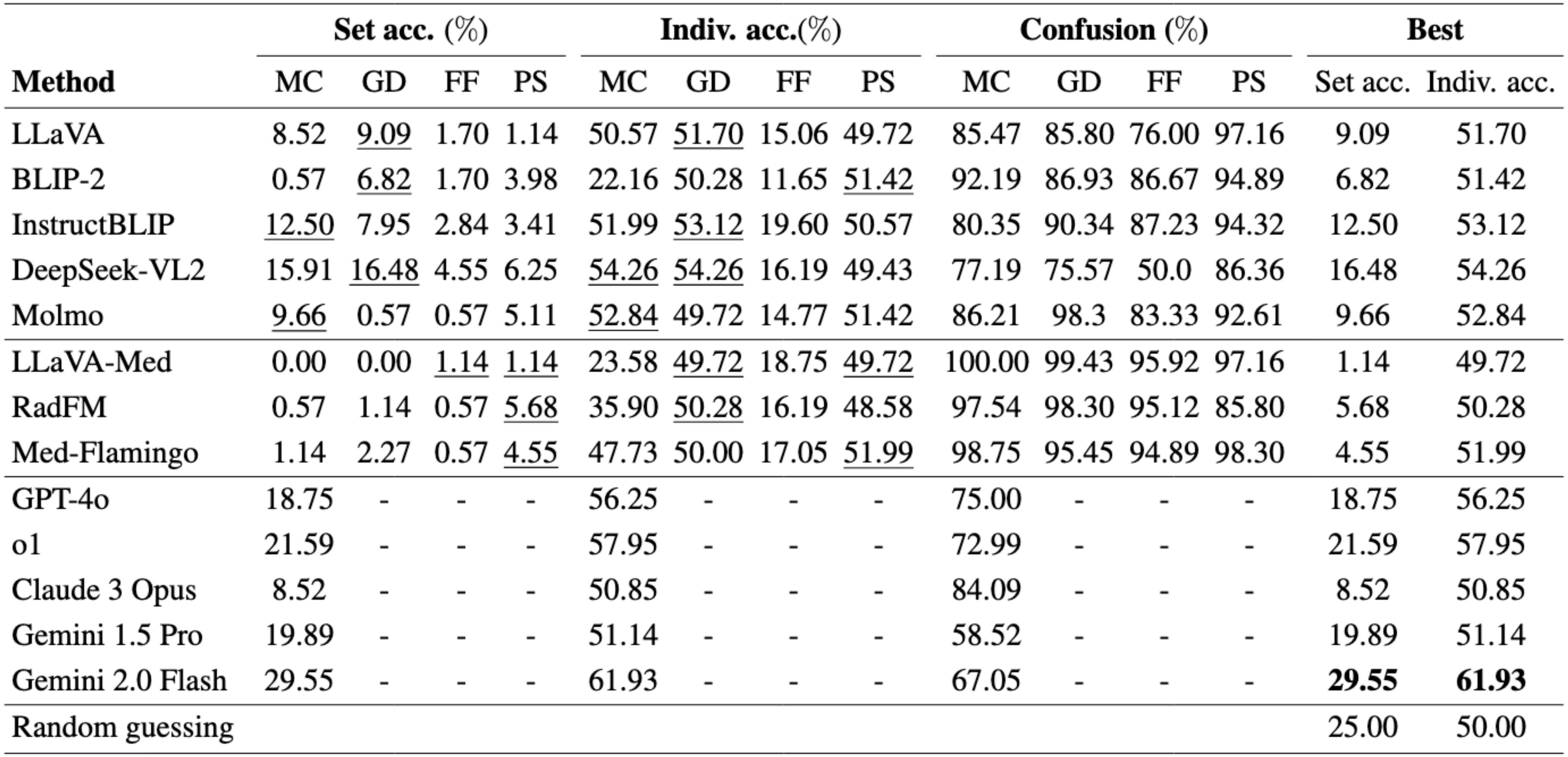
Radiologist

Filtering + editing



MEDICONFUSION

Performance



	Set acc. (%)				Indiv. acc.(%)				Confusion (%)				Best	
Method	MC	GD	FF	PS	MC	GD	FF	PS	MC	GD	FF	PS	Set acc.	Indiv. acc.
LLaVA	8.52	9.09	1.70	1.14	50.57	51.70	15.06	49.72	85.47	85.80	76.00	97.16	9.09	51.70
BLIP-2	0.57	6.82	1.70	3.98	22.16	50.28	11.65	51.42	92.19	86.93	86.67	94.89	6.82	51.42
InstructBLIP	12.50	7.95	2.84	3.41	51.99	53.12	19.60	50.57	80.35	90.34	87.23	94.32	12.50	53.12
DeepSeek-VL2	15.91	16.48	4.55	6.25	54.26	54.26	16.19	49.43	77.19	75.57	50.0	86.36	16.48	54.26
Molmo	9.66	0.57	0.57	5.11	52.84	49.72	14.77	51.42	86.21	98.3	83.33	92.61	9.66	52.84
LLaVA-Med	0.00	0.00	1.14	1.14	23.58	49.72	18.75	49.72	100.00	99.43	95.92	97.16	1.14	49.72
RadFM	0.57	1.14	0.57	5.68	35.90	50.28	16.19	48.58	97.54	98.30	95.12	85.80	5.68	50.28
Med-Flamingo	1.14	2.27	0.57	4.55	47.73	50.00	17.05	51.99	98.75	95.45	94.89	98.30	4.55	51.99
GPT-4o	18.75	-	-	-	56.25	-	-	-	75.00	-	-	-	18.75	56.25
o1	21.59	-	-	-	57.95	-	-	-	72.99	-	-	-	21.59	57.95
Claude 3 Opus	8.52	-	-	-	50.85	-	-	-	84.09	-	-	-	8.52	50.85
Gemini 1.5 Pro	19.89	-	-	-	51.14	-	-	-	58.52	-	-	-	19.89	51.14
Gemini 2.0 Flash	29.55	-	-	-	61.93	-	-	-	67.05	-	-	-	**29.55**	**61.93**
Random guessing													25.00	50.00

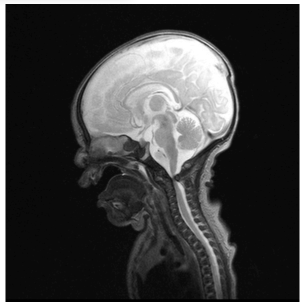
Failure Modes

1

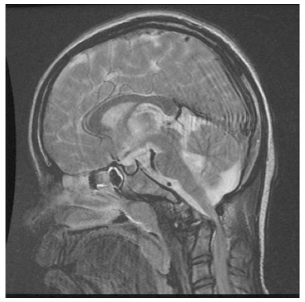
Normal/variant anatomy vs. pathology

What is the primary cause of severe spinal cord compression in this image?

Anterior subluxation of C1 vertebra relative to C2



C1-C2 instability

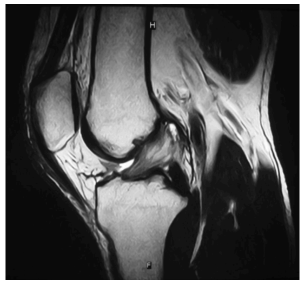


2

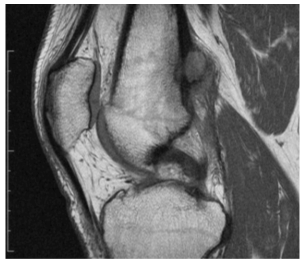
Lesion signal characteristics

What is the signal intensity of the abnormality observed on the T2-weighted images?

High signal intensity



Low signal intensity

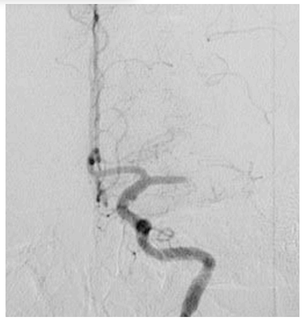


3

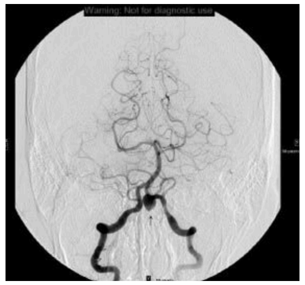
Vascular conditions

What specific vascular pathology is observed in the image?

Total occlusion of the left middle cerebral artery



Presence of a left PICA aneurysm

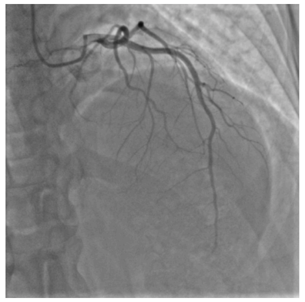


4

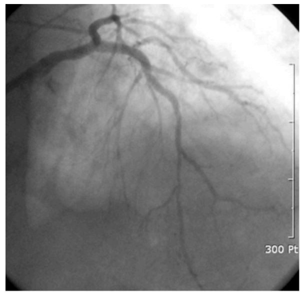
Medical devices

What is the condition of the left anterior descending artery close to the apical region?

Critical narrowing with flow cessation



Successfully treated with stent implantation



Thank you for your attention!