AI in Medicine Lab

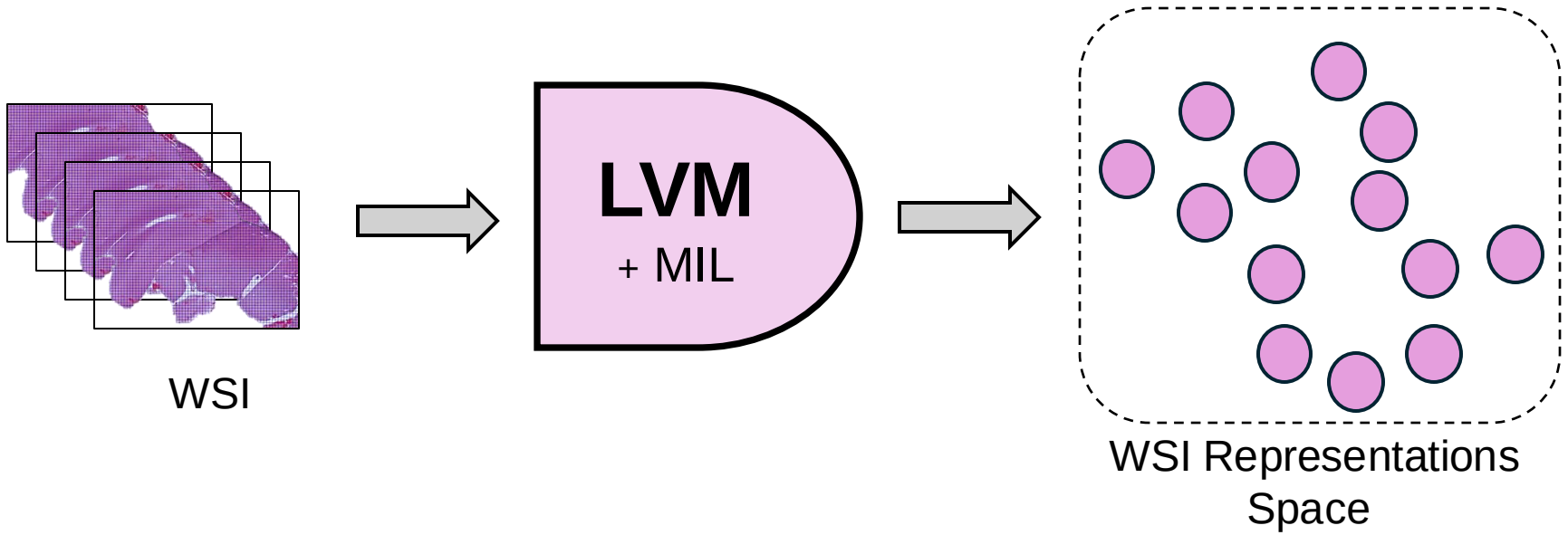
Boltzmann Semantic Score: A Semantic Metric to Evaluate Large Vision Models using Large Language Models

Ali Khajegili Mirabadi, Katherine Rich, Hossein Farahani, Ali Bashashati

The University of British Columbia

The 13th International Conference on Learning Representations

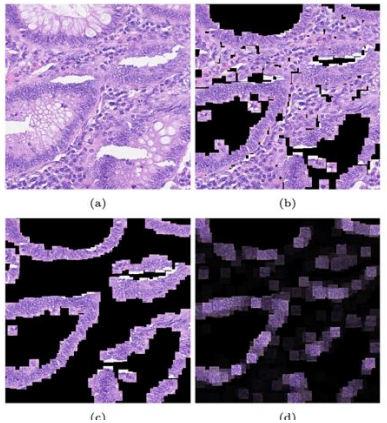
Question:



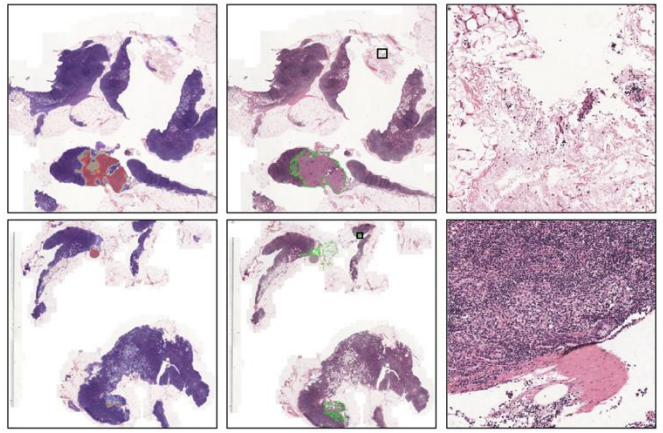
Do Large Vision Models (LVMs) extract **semantically relevant** features similar to those identified by human experts?

What people do now:

- Gradient Visualization or Attention Score Visualization:
  - ➢ Comparison with Expert Annotations
  - ➢ Expert Evaluations of the Heatmaps



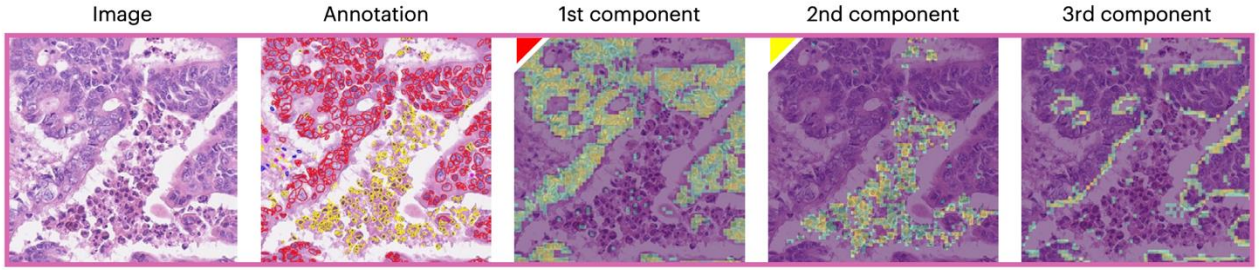
Instance Attention Score Visualization[1]



Attention Heatmaps and Comparison with Expert Annotations[2]

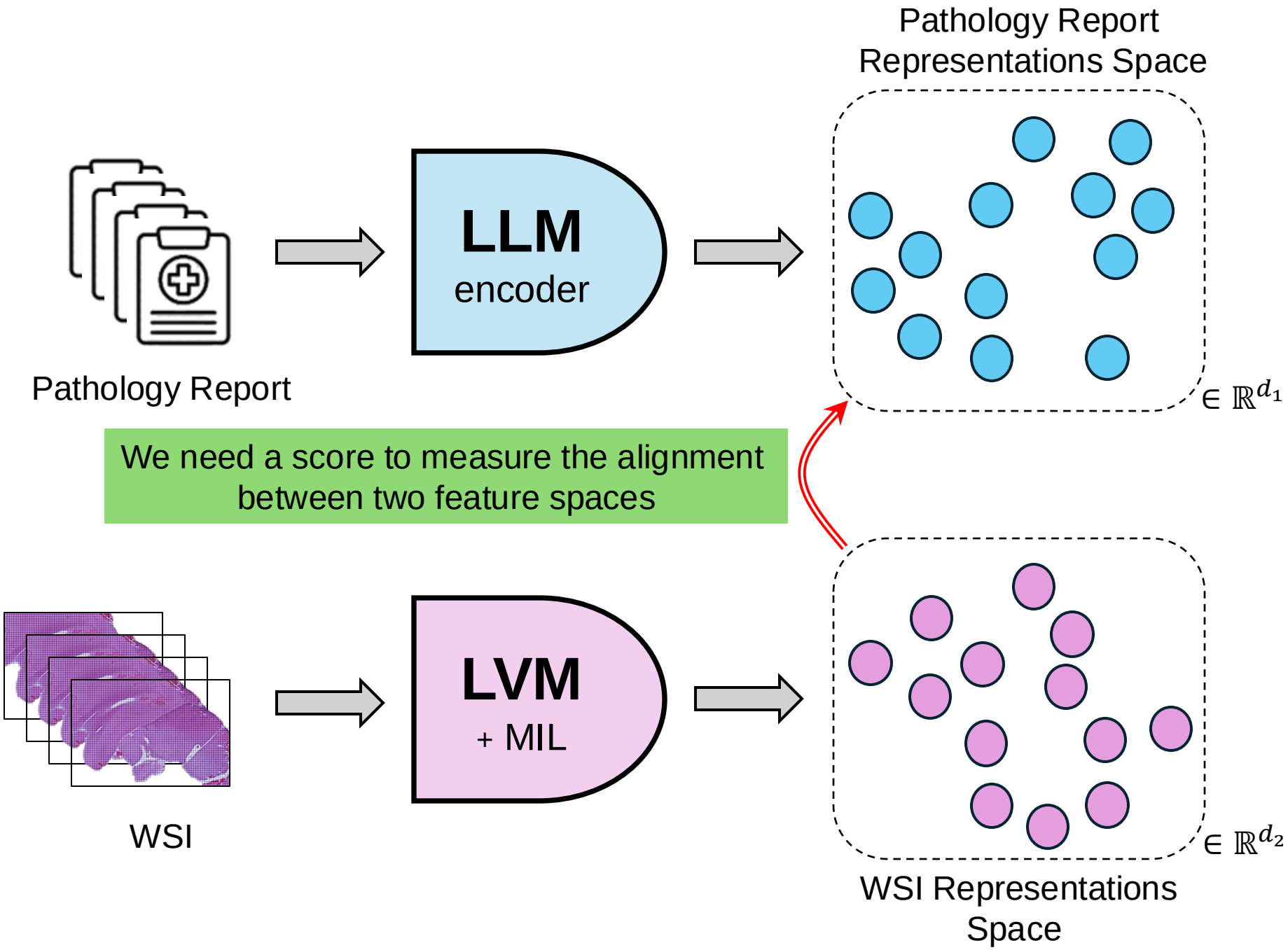
Problems with this approach:

- **Subjective**
- **Subject to Variability**
- **Small Sample Size**
- **Limited to Certain Cancers**



PCA on Tile Features and Compare with Annotations[3]

Solution:


Pathology Report
Representations Space
LLM
encoder
Pathology Report
$\in \mathbb{R}^{d_1}$
We need a score to measure the alignment between two feature spaces
LVM
+ MIL
WSI
$\in \mathbb{R}^{d_2}$
WSI Representations Space

Challenges & Questions:

Q1: Can LLMs encode Pathology Reports properly?

Q2: How to measure the structural similarity of two spaces with non-identical dimensions?

Zero-shot LLM representation in Pathology:

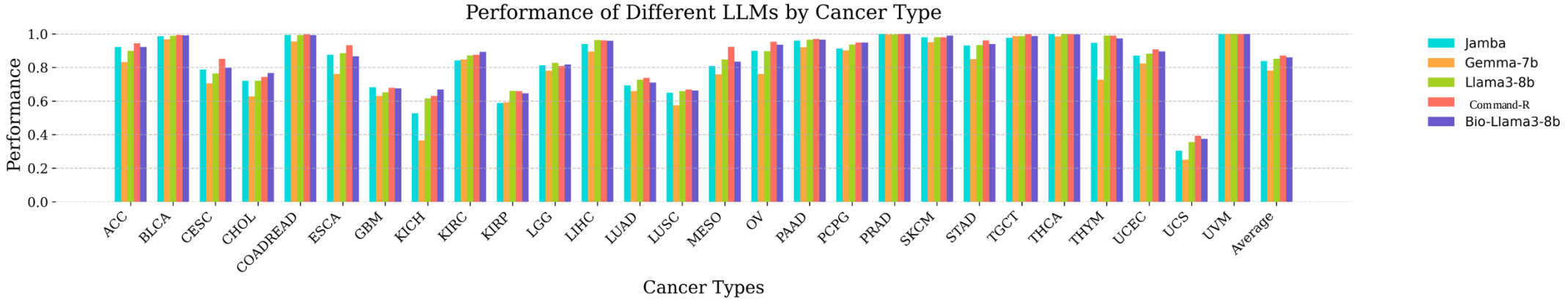
- 5 Tasks with 5 LLMs:
  - Information Retrieval in Organ-Independent and Organ-Specific settings with original and perturbed texts
  - Survival prediction

Command-R UMAP plot of TCGA: ~9,500 pathology reports

The C-index of RSF for Survival Prediction

LLM	BRCA	GBM	KIRC	KIRP	LGG	LUAD	LUSC	UCEC
Command-R	$0.622_{\pm 0.02}$	$0.537_{\pm 0.03}$	$0.722_{\pm 0.04}$	$0.743_{\pm 0.10}$	$0.643_{\pm 0.05}$	$0.611_{\pm 0.04}$	$0.547_{\pm 0.03}$	$0.602_{\pm 0.08}$
Gemma-7b	$0.603_{\pm 0.04}$	$0.512_{\pm 0.03}$	$0.707_{\pm 0.02}$	$0.634_{\pm 0.11}$	$0.611_{\pm 0.08}$	$0.570_{\pm 0.03}$	$0.543_{\pm 0.04}$	$0.600_{\pm 0.04}$
Jamba	$0.625_{\pm 0.05}$	$0.501_{\pm 0.02}$	$0.689_{\pm 0.06}$	$0.745_{\pm 0.09}$	$0.639_{\pm 0.06}$	$0.595_{\pm 0.06}$	$0.545_{\pm 0.01}$	$0.617_{\pm 0.07}$
Llama3-8b	$0.629_{\pm 0.05}$	$0.521_{\pm 0.04}$	$0.713_{\pm 0.03}$	$0.759_{\pm 0.07}$	$0.607_{\pm 0.07}$	$0.585_{\pm 0.07}$	$0.520_{\pm 0.06}$	$0.580_{\pm 0.08}$
Bio-Llama3-8b	$0.627_{\pm 0.07}$	$0.537_{\pm 0.03}$	$0.709_{\pm 0.05}$	$0.726_{\pm 0.08}$	$0.587_{\pm 0.07}$	$0.583_{\pm 0.03}$	$0.548_{\pm 0.04}$	$0.621_{\pm 0.04}$
Average	$0.621_{\pm 0.01}$	$0.522_{\pm 0.02}$	$0.708_{\pm 0.01}$	$0.721_{\pm 0.05}$	$0.617_{\pm 0.02}$	$0.589_{\pm 0.02}$	$0.541_{\pm 0.01}$	$0.604_{\pm 0.02}$

LLMs' top-1 accuracy in organ-specific setting with original text



Zero-shot LLM representation in Pathology:

- 5 Tasks with 5 LLMs:
  - Information Retrieval in Organ-Independent and Organ-Specific settings with original and perturbed texts
  - Survival prediction

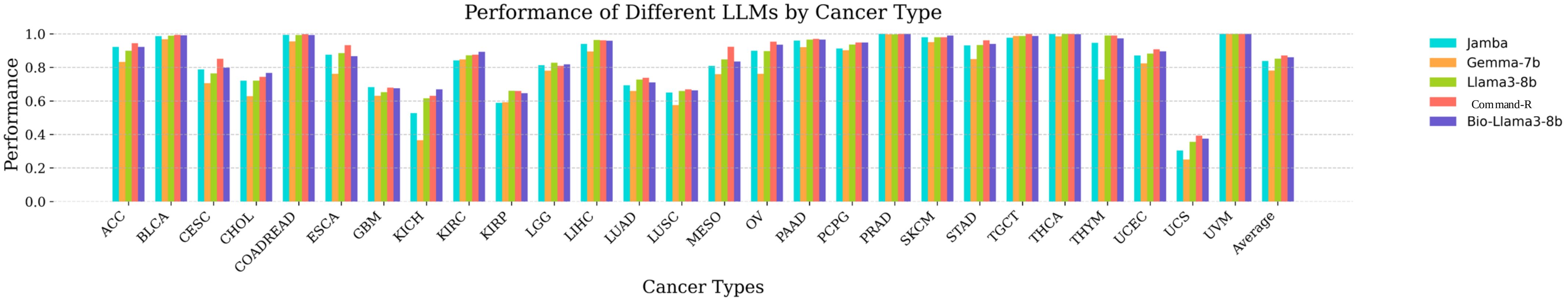
Command-R UMAP plot of TCGA: ~9,500 pathology reports

The C-index of RSF for Survival Prediction

LLM	[illegible]	GBM	KIRC	KIRP	LGG	LUAD	LUSC	UCEC
Command-R	[illegible]	$0.537_{\pm[illegible]}$	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Gemma-7b	[illegible]	$0.512_{\pm0.03}$	$0.707_{\pm0.02}$	$0.634_{\pm0.11}$	$0.611_{\pm0.08}$	$0.570_{\pm0.03}$	$0.543_{\pm0.04}$	$0.600_{\pm0.04}$
Jamba	$0.625_{\pm0.05}$	$0.501_{\pm0.02}$	$0.689_{\pm0.06}$	$0.745_{\pm0.09}$	$0.639_{\pm0.06}$	$0.595_{\pm0.06}$	$0.545_{\pm0.01}$	$0.617_{\pm0.07}$
Llama3-8b	$0.629_{\pm0.05}$	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Bio-Llama3-8b	$0.627_{\pm0.07}$	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Average	$0.621_{\pm0.01}$	$0.522_{\pm0.02}$	$0.708_{\pm0.01}$	$0.721_{\pm0.05}$	$0.617_{\pm0.02}$	[illegible]	$0.541_{\pm0.01}$	[illegible]

Q1: Can LLMs encode Pathology Reports properly?
Answer: **Yes, their performance is promising!**

LLMs' top-1 accuracy in organ-specific setting with original text

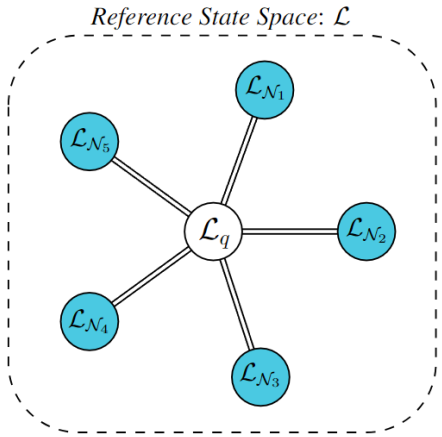


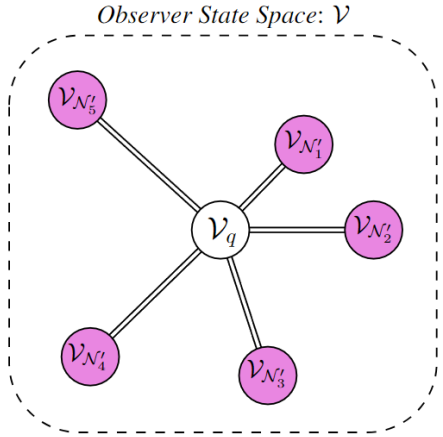
Challenges & Questions:

Q1: Can LLMs encode Pathology Reports properly?

Q2: How to measure the structural similarity of two spaces with non-identical dimensions?

Boltzmann Semantic Score: Theory



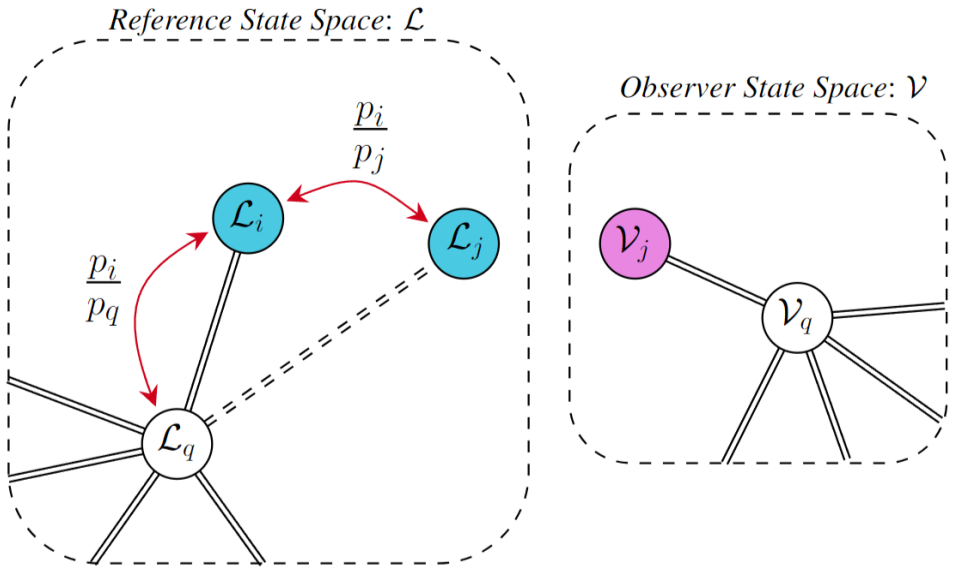


$\mathcal{L}_{\mathcal{N}_i}$: $\mathcal{N}_i$–th neighboring state of q

$\mathcal{L}_q$: the central state q

$\mathcal{V}_{\mathcal{N}'_i}$: $\mathcal{N}'_i$–th neighboring state of q

$\mathcal{V}_q$: the central state q



$\mathcal{L}_i$: the reference matching state

$\mathcal{L}_j$: the corresponding non-matching estimated state by Φ in $\mathcal{L}$

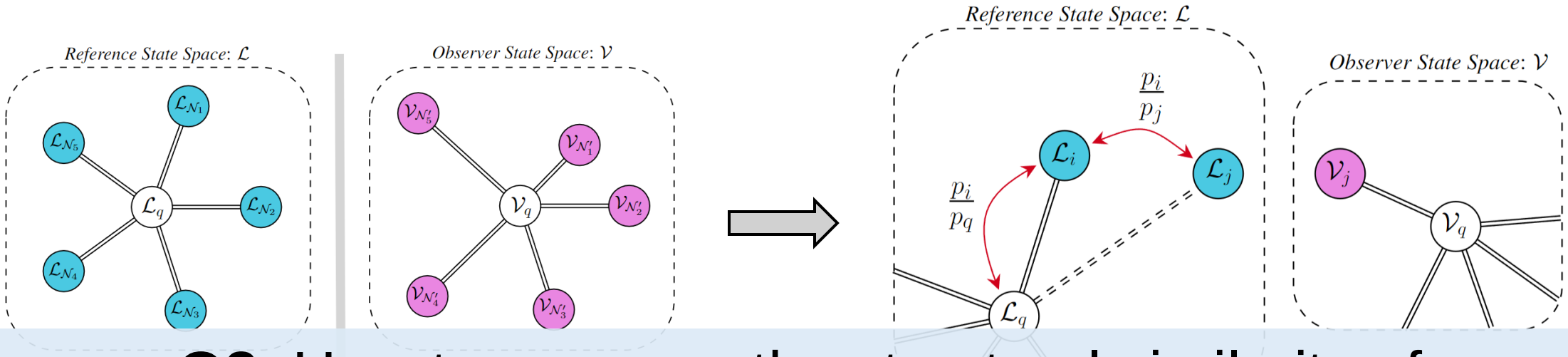
$\mathcal{V}_j$: the estimated state j by Φ in $G_{\mathcal{V},q}$

Boltzmann Factor: $$\frac{p_i}{p_q} = \exp\left(-\frac{\Delta\mathcal{E}}{kT}\right) = \exp\left(-\frac{\|\mathcal{L}_i - \mathcal{L}_q\|_2}{\sqrt{d_1}}\right)$$

Second-order Boltzmann Factor: $$b_{i;j|q} := \frac{p_i}{p_j} \cdot \frac{p_i}{p_q}$$

$$\mathcal{B}_q = \frac{\sum_{(i,j)\in\mathbb{A}} b_{i;j|q}}{\sum_{(i,j)\in\mathbb{A}\cup\mathbb{D}} b_{i;j|q}} = \frac{\sum_{(i,j)\in\mathbb{A}\cup\mathbb{D}} b_{i;j|q} - \sum_{(i,j)\in\mathbb{D}} b_{i;j|q}}{\sum_{(i,j)\in\mathbb{A}\cup\mathbb{D}} b_{i;j|q}} = 1 - \frac{\sum_{(i,j)\in\mathbb{D}} b_{i;j|q}}{\sum_{(i,j)\in\mathbb{A}\cup\mathbb{D}} b_{i;j|q}}$$

Boltzmann Semantic Score: Theory



Q2: How to measure the structural similarity of two spaces with non-identical dimensions?

Answer: **Boltzmann Semantic Score!**

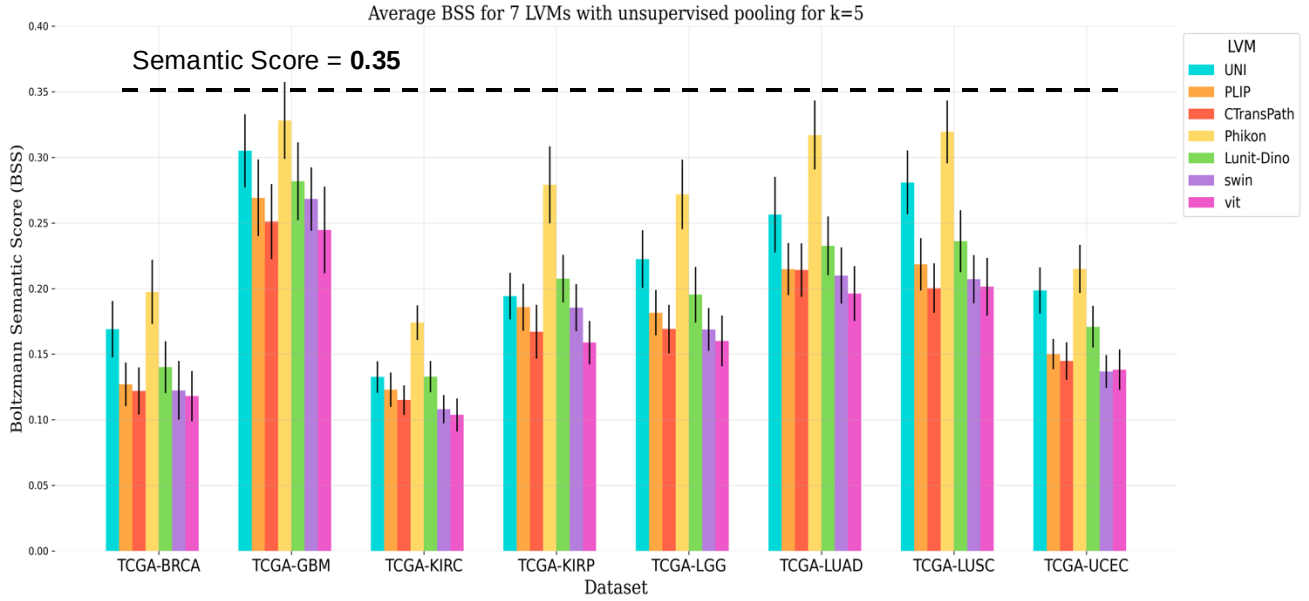
Boltzmann Factor: $$\frac{p_i}{p_q} = \exp\left(-\frac{\Delta\mathcal{E}}{kT}\right) = \exp\left(-\frac{\|\mathcal{L}_i - \mathcal{L}_q\|_2}{\sqrt{d_1}}\right)$$

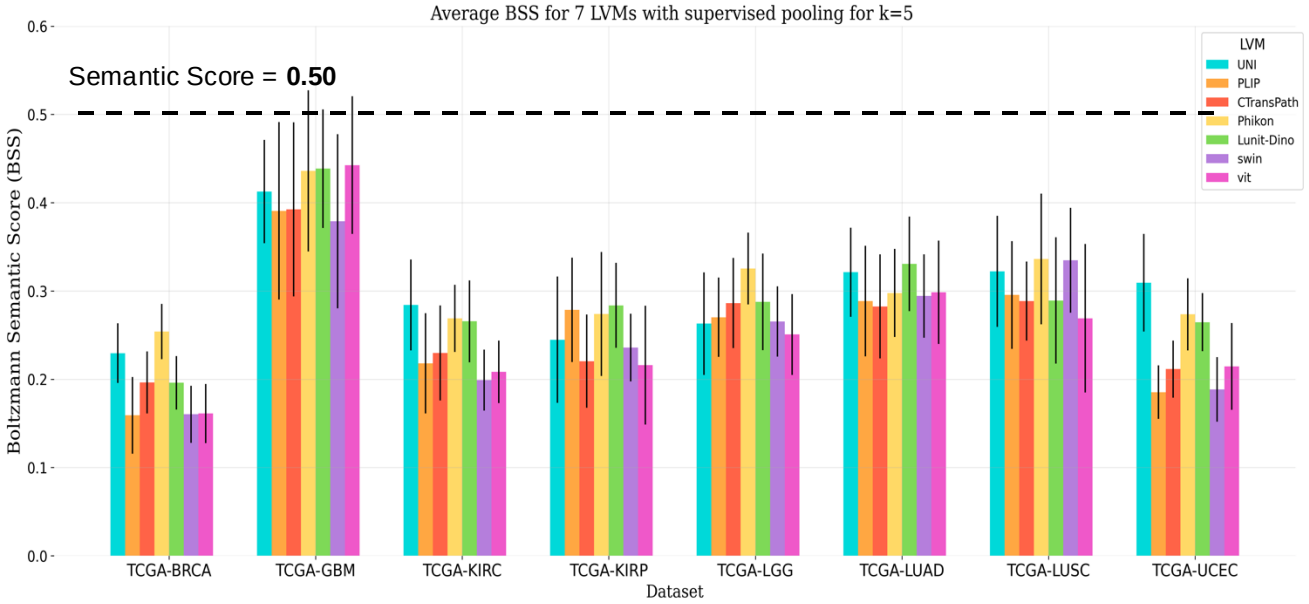
Second-order Boltzmann Factor: $$b_{i;j|q} := \frac{p_i}{p_j} \cdot \frac{p_i}{p_q}$$

$$\mathcal{B}_q = \frac{\sum_{(i,j)\in\mathbb{A}} b_{i;j|q}}{\sum_{(i,j)\in\mathbb{A}\cup\mathbb{D}} b_{i;j|q}} = \frac{\sum_{(i,j)\in\mathbb{A}\cup\mathbb{D}} b_{i;j|q} - \sum_{(i,j)\in\mathbb{D}} b_{i;j|q}}{\sum_{(i,j)\in\mathbb{A}\cup\mathbb{D}} b_{i;j|q}} = 1 - \frac{\sum_{(i,j)\in\mathbb{D}} b_{i;j|q}}{\sum_{(i,j)\in\mathbb{A}\cup\mathbb{D}} b_{i;j|q}}$$

Boltzmann Semantic Score: Benchmark

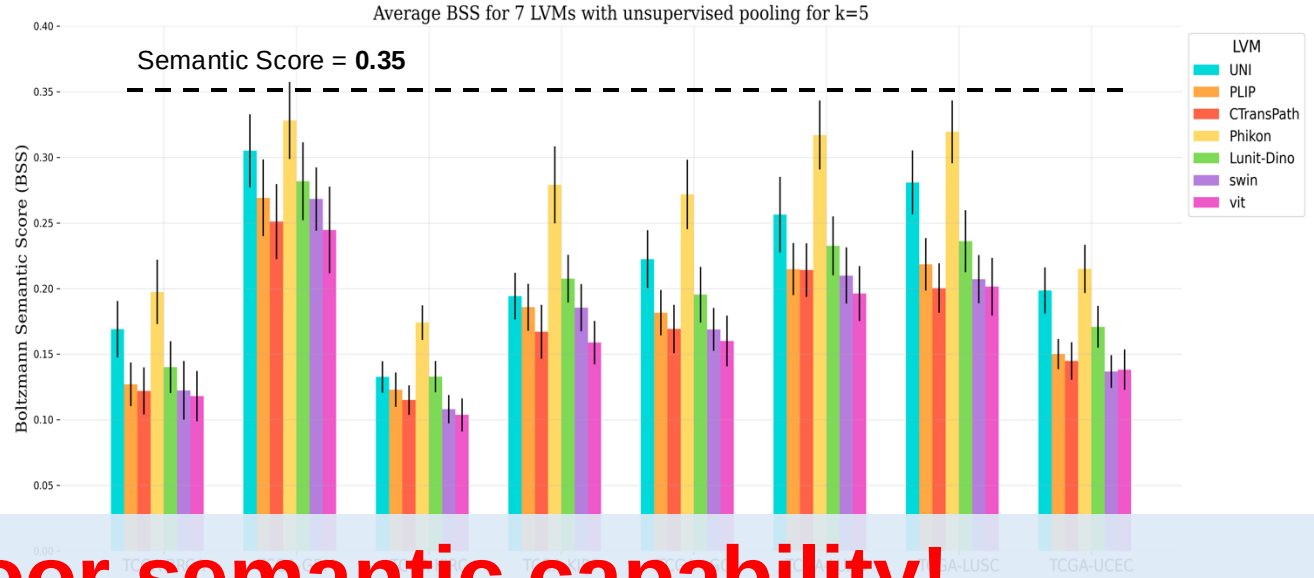
- 7 LVMs were benchmarked using BSS
  - UNI
  - Phikon
  - PLIP
  - CTransPath
  - Lunit-Dino
  - SwinT
  - ViT
- Two setting for instance aggregation:
  - Supervised Pooling with AbMIL
  - Unsupervised Pooling with Mean-pooling
- The BSS reported here is average on 5 LLMs as the references





Boltzmann Semantic Score: Benchmark

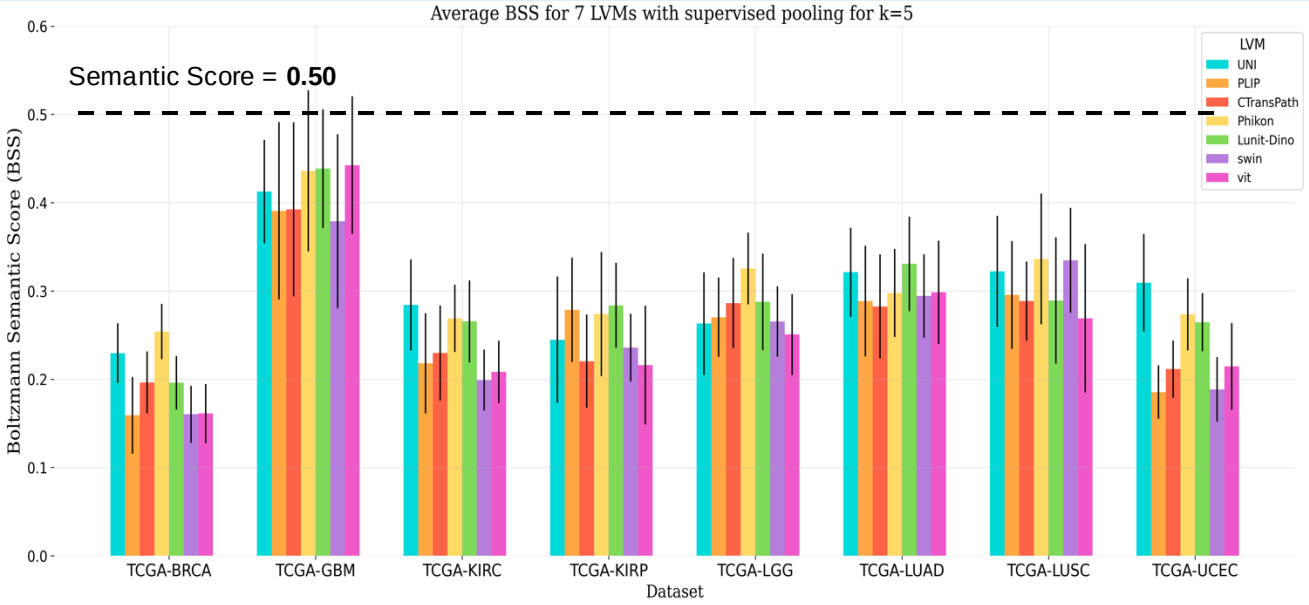
- 7 LVMs were benchmarked using BSS
  - UNI
  - Phikon
  - PLIP
  - CTransPath
  - Lunit-Dino
  - SwinT
  - ViT



LVMs suffer from poor semantic capability!

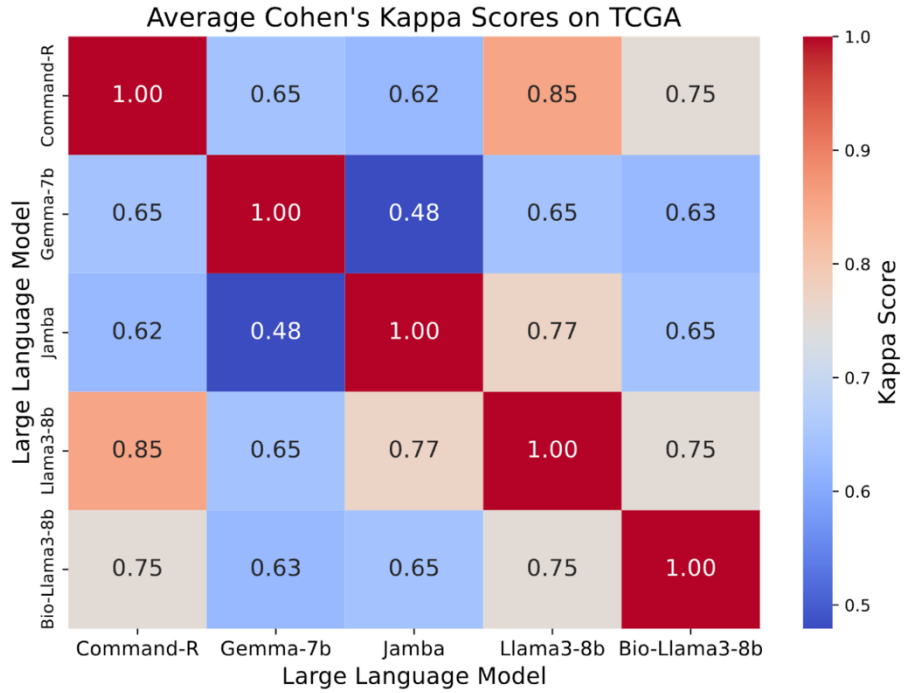
- Two setting for instance aggregation:
  - Supervised Pooling with AbMIL
  - Unsupervised Pooling with Mean-pooling

- The BSS reported here is average on 5 LLMs as the references



Boltzmann Semantic Score: Reliability

- Two Experiments:
  - ➢ LLMs' consensus in ranking LVMs using Cohen's Kappa
  - ➢ Downstream task metric predictability



(a) Correlation between BSS and top-1 Accuracy in Information Retrieval

LLM	GBM		KIRC		KIRP		LGG		LUAD		LUSC	
	r	p-value	r	p-value	r	p-value	r	p-value	r	p-value	r	p-value
Command-R	0.335	$4.6e^{-03}$	0.605	$2.9e^{-08}$	0.267	$2.5e^{-02}$	0.274	$2.2e^{-02}$	0.830	$6.4e^{-19}$	0.557	$5.4e^{-07}$
Gemma-7b	0.386	$9.6e^{-04}$	0.483	$2.3e^{-05}$	0.475	$3.3e^{-05}$	0.309	$9.2e^{-03}$	0.879	$1.4e^{-23}$	0.459	$6.4e^{-05}$
Jamba	0.302	$1.1e^{-02}$	0.583	$1.2e^{-07}$	0.247	$4.0e^{-02}$	0.259	$3.0e^{-02}$	0.838	$1.5e^{-19}$	0.517	$4.6e^{-06}$
Llama3-8b	0.293	$1.4e^{-02}$	0.550	$8.2e^{-07}$	0.326	$5.9e^{-03}$	0.227	$5.9e^{-02}$	0.807	$3.3e^{-17}$	0.540	$1.4e^{-06}$
Bio-Llama3-8b	0.355	$2.6e^{-03}$	0.572	$2.4e^{-07}$	0.323	$6.5e^{-03}$	0.231	$5.5e^{-02}$	0.834	$2.9e^{-19}$	0.575	$1.9e^{-07}$

(b) Correlation between BSS and C–index in Survival Prediction

LLM	BRCA		GBM		KIRC		KIRP		LGG		LUAD		LUSC		UCEC	
	r	p-value	r	p-value	r	p-value	r	p-value	r	p-value	r	p-value	r	p-value	r	p-value
Command-R	0.344	$1.9e^{-11}$	−0.093	$9.6e^{-01}$	0.273	$1.1e^{-07}$	0.150	$2.4e^{-03}$	−0.027	$6.9e^{-01}$	0.233	$5.4e^{-06}$	0.099	$3.2e^{-02}$	0.367	$6.6e^{-13}$
Gemma-7b	0.307	$2.3e^{-09}$	−0.030	$7.1e^{-01}$	0.245	$1.8e^{-06}$	0.202	$7.1e^{-05}$	−0.017	$6.2e^{-01}$	0.223	$1.3e^{-05}$	0.081	$6.6e^{-02}$	0.352	$6.4e^{-12}$
Jamba	0.344	$2.0e^{-11}$	−0.080	$9.3e^{-01}$	0.257	$5.7e^{-07}$	0.144	$3.4e^{-03}$	−0.031	$7.2e^{-01}$	0.209	$3.9e^{-05}$	0.089	$4.8e^{-02}$	0.352	$5.7e^{-12}$
Llama3-8b	0.350	$7.8e^{-12}$	−0.085	$9.4e^{-01}$	0.272	$1.1e^{-07}$	0.151	$2.4e^{-03}$	−0.036	$7.5e^{-01}$	0.218	$1.9e^{-05}$	0.096	$3.6e^{-02}$	0.370	$4.4e^{-13}$
Bio-Llama3-8b	0.347	$1.2e^{-11}$	−0.087	$9.5e^{-01}$	0.287	$2.3e^{-08}$	0.131	$7.1e^{-03}$	−0.034	$7.4e^{-01}$	0.224	$1.1e^{-05}$	0.112	$1.8e^{-02}$	0.373	$2.9e^{-13}$

Conclusion:

Do Large Vision Models (LVMs) extract **semantically relevant** features similar to those identified by human experts?
Answer: LVMs suffer from poor semantic capability!

Q1: Can LLMs encode Pathology Reports properly?
Answer: **Yes, zero-shot LLMs are promising!**

Q2: How to measure the structural similarity of two spaces with non-identical dimensions?
Answer: **Yes, Boltzmann Semantic Score!**