Insights into a radiology-specialised multimodal large language model with sparse autoencoders

Kenza Bouzid, Shruthi Bannur, Felix Meissen, Daniel C. Castro, Anton Schwaighofer, Javier Alvarez-Valle, Stephanie L. Hyland

Interpreting MAIRA-2 internals

MAIRA-2: MLLM for chest X-ray reporting

What is the Multimodal Large Language Model (MLLM) MAIRA-2 [1]*?*

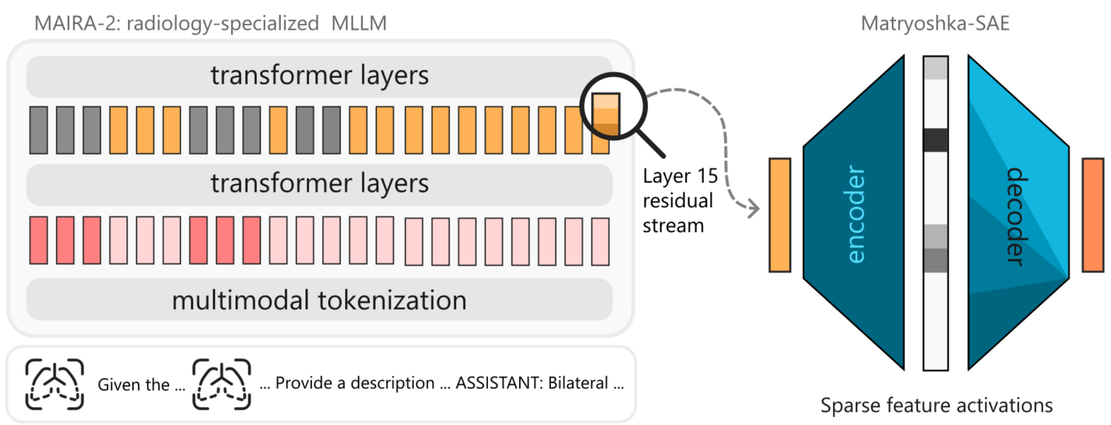
- MLLM that generates the "**findings**" of a chest X-ray report.
- Uses current image, prior imaging, prior report, indication, etc.
- Backbones: LLM (Vicuna 7B 1.5), vision encoder (RAD-DINO [2]).

Why is it important to interpret MAIRA-2?

- Understand concepts learned by MAIRA-2 and how they are used.
- Improve model trust in a high-stakes domain such as healthcare.
- A path towards more transparent and controllable radiology models.

SAE training

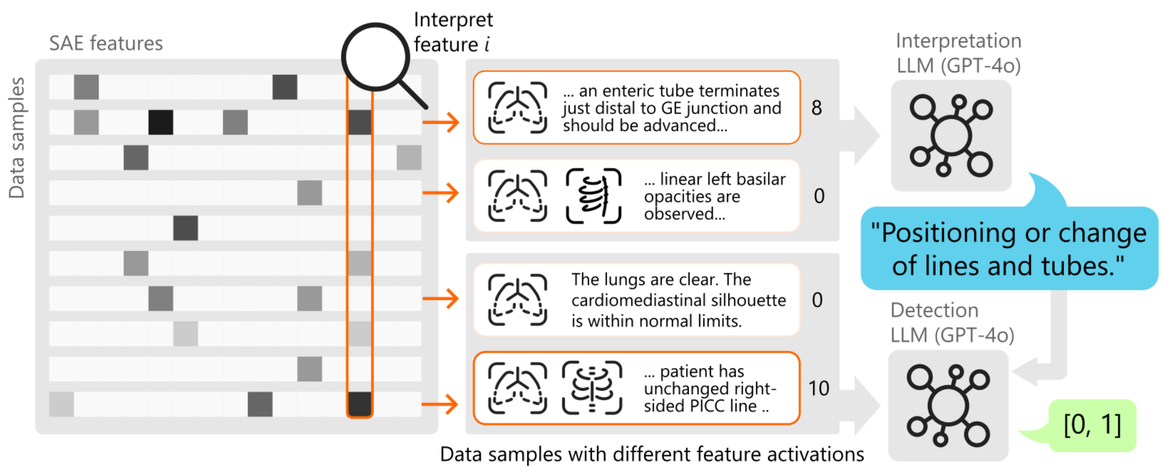
- Fit a **Matryoshka BatchTopk SAE [3]** on MAIRA-2 activations (size 4096).
- Using expansion factor *ef*=4 and *k*=256, resulting in 16,384 SAE features.



$$\mathcal{L}(\mathbf{x}) = \sum_{m_j \in \mathcal{M}} \|\mathbf{x} - \underbrace{(\mathbf{W}^{\text{dec}}_{1:m_j}\mathbf{f}(\mathbf{x})_{1:m_j} + \mathbf{b}^{\text{dec}})}_{\text{reconstruction using first } m_j \text{ features}}\|_2^2 + \alpha\mathcal{L}_{\text{aux}}$$

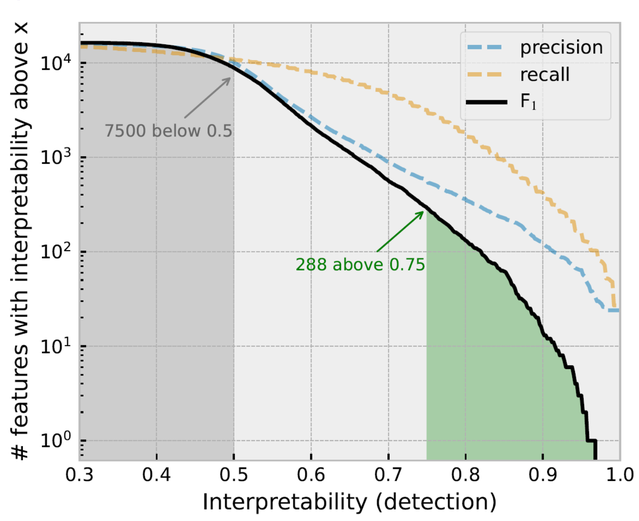
Automated interpretations with LLMs

- Large scale auto interpretability of 99.5% (16,299) of the 16,384 SAE features.
- Show dataset exemplars with different activation strengths to GPT-4o [4].
- Use detection scoring [4] to evaluate the quality of automated interpretations .



Interpretable features exist, but are rare

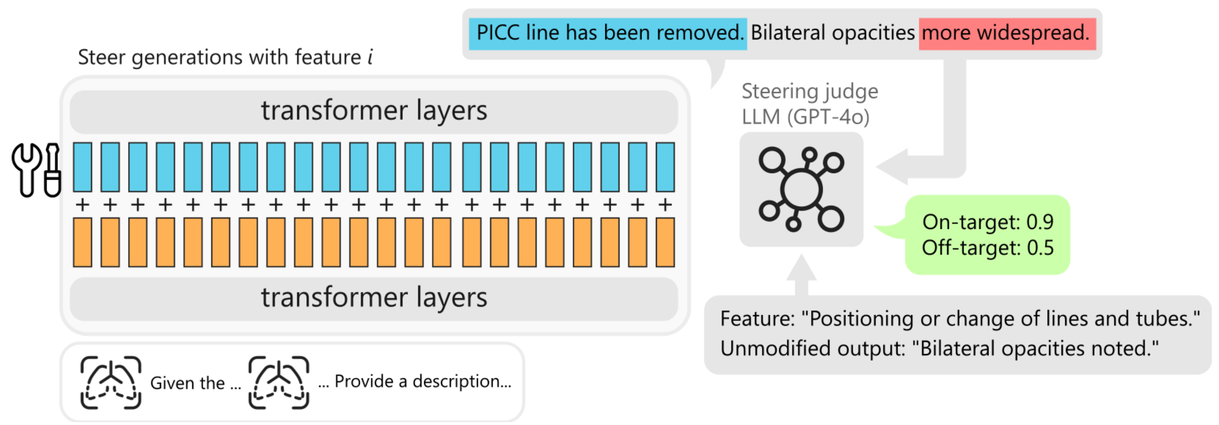
- We discover fine-grained concepts in MAIRA-2.
- Among 16,384 features, 288 (1.8%) score above 0.75.
- 7,500 features (46%) score below 0.5 (random performance).
- Many SAE features are not interpretable (yet!).



Steering MAIRA-2

Steering and steering evaluation

- Steer generations of MAIRA-2 using discovered concept vectors [5].
- Evaluate on- and off-target effects of steering using LLMs [6].



Steering success depends on the feature

- Feature steering produces more off-target changes than on-target ones.
- Some off-target effects reflect the underlying correlations in the dataset.
- In 35% of cases across all features, steering leads to no observable changes.

	Feature 1599: ***Describing findings without comparison to prior images.***
Original generation	Compared with the prior study, lung volumes are lower, causing bronchovascular crowding. However, no focal consolidation, pleural effusion, or pneumothorax. Cardiomediastinal and hilar silhouettes are unchanged.
Steered generation	Lung volumes are low, causing bronchovascular crowding. No focal consolidation, pleural effusion, or pneumothorax. Cardiomediastinal and hilar silhouettes are unremarkable.
	Feature 6412: ***Detection of pleural effusions on imaging studies.***
Original generation	The heart size is normal. The mediastinal and hilar contours are unremarkable. The pulmonary vascularity is normal. Patchy ill-defined opacity is noted within the left lung base, which may represent an area of developing infection. No pleural effusion or pneumothorax is present. There are no acute osseous abnormalities.
Steered generation	The heart size is normal. The mediastinal and hilar contours are unremarkable. The pulmonary vascularity is normal. There is a small left pleural effusion. Minimal streaky opacity in the left lung base may reflect atelectasis, though infection cannot be completely excluded. There is no pneumothorax. No acute osseous abnormality is seen.
	Feature 10643: ***Immediate notification of findings by telephone upon discovery.***
Original generation	Since _, small right pleural effusion is unchanged, right pleural catheter is in unchanged position. The lungs are clear. The cardiomediastinal silhouette, hilar contours, and pleural surfaces are normal. No pneumothorax.
Steered generation	Since _, a right pleural drainage catheter is in place. A small right apical pneumothorax is seen. A small right pleural effusion is seen. A left pleural effusion is small. The lungs are clear. The heart size is normal. Tips in the oesophagus are noted.

On-target effects Off-target effects

Mechanistic interpretability of MAIRA-2 is challenging

- SAE training is complex and engineering-intensive.
- Automated interpretability in specialized domains is quite difficult.
- Only a small set of SAE features are interpretable.
- Feature splitting and composition lead to many redundant descriptions.
- Steering shows low success rates and sometimes causes adverse effects.
- Scalable evaluation is complex and requires careful case selection.

References

aka.ms/maira-2-sae

1. Bannur, S., Bouzid, K., Castro, D. C., Schwaighofer, A., Thieme, A., Bond-Taylor, S., ... & Hyland, S. L. (2024). MAIRA-2: Grounded radiology report generation. *arXiv:2406.04449* https://huggingface.co/microsoft/maira-2
2. Pérez-García, F., Sharma, H., Bond-Taylor, S., Bouzid, K., Salvatelli, V., Ilse, M., ... & Oktay, O. (2025). Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*. https://huggingface.co/microsoft/rad-dino
3. Bussmann, B., Nabeshima, N., Karvonen, A., and Nanda, N. (2025). Learning Multi-Level Features with Matryoshka Sparse Autoencoders. *arXiv:2503.17547*
4. Paulo, G., Mallen, A., Juang, C., and Belrose, N. (2024). Auto-matically Interpreting Millions of Features in Large Language Models. *arXiv:2410.13928*.
5. Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez,J. J., Mini, U., and MacDiarmid, M. (2023). Steering language models with activation engineering. *arXiv:2308.10248*
6. Wu, Z., Arora, A., Geiger, A., Wang, Z., Huang, J., Jurafsky,D., Manning, C. D., and Potts, C. (2025). AxBench: Steering LLMs? Even Simple Baselines Outperform Sparse Autoencoders. *arXiv:2501.17148*

Microsoft Research, Health Futures UK

Kenza Bouzid | LinkedIn
kenza.bouzid@microsoft.com

Microsoft