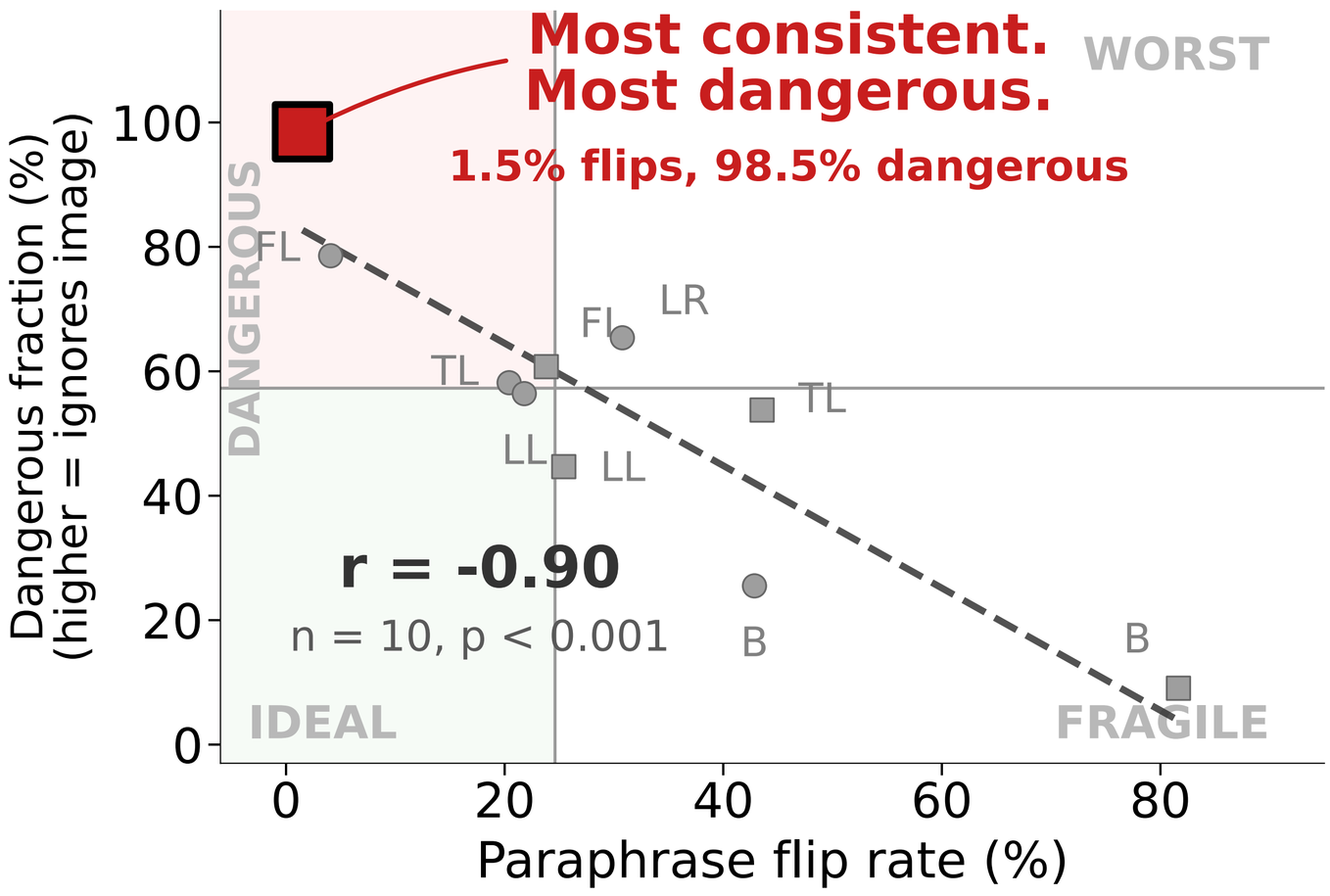
Consistency is not safety

Binesh Sadanadan (Advisor: Vahid Behzadan)
SAIL Lab, University of New Haven · CHIL 2026 Doctoral Symposium

Measuring, diagnosing, and fixing paraphrase failures in medical vision-language models.



The finding

The most consistent model in the study is the **most dangerous**.

Across 10 model × dataset settings, the paraphrase flip rate and the share of consistent-but-image-blind (*Dangerous*) predictions correlate $r = -0.89$ ($p < 0.001$).

The lowest-flip model, LLaVA-Rad on PadChest (1.5% flips), has the *highest* Dangerous share, **98.5%**.

The dissertation: measure, diagnose, mitigate, evaluate, monitor

1. Measure

not buy consistency

PSF-Med: 26,850 questions, 3 countries; 6 VLMs flip 6.3% to 40.4%

2. Diagnose

because it ignores the image

LLaVA-Rad: 1.5% flips but 99.1% text-only agreement

3. Mitigate

0.10% of weights

14.6% → 4.4% (−69.6%, p=0.002)

4. Evaluate

attention is not grounding

$r = -0.89$; attention vs. causal $\rho \approx -0.05$ to -0.17

5. Monitor

gates admit text-answerable cases

flip AUROC = 0.711

6 VLMs evaluated after setting aside two degenerate yes-bias cells.

Low flip rate is read as safe. It is not.

A medical VLM can give different yes/no answers to one clinical question reworded. The field treats consistency as a safety signal.

But a model can be perfectly consistent by answering from the text and ignoring the image.

Future directions

- Extend the layer-17 register gate to other architectures (LLaMA-based, encoder-only).
- Train for grounding directly: objectives that penalize invariance to pathology occlusion.
- Deployment-time monitoring with formal coverage guarantees.

Contributions

- **PSF-Med benchmark:** 26,850 questions across US, Spain, and Vietnam, with a rubric-based equivalence audit.
- The **robustness-grounding trade-off**, across 6 models and 3 datasets.
- **Mechanistic diagnosis:** SAE feature 3818 at layer 17, a clinical query register gate.
- **SAE-guided LoRA:** 0.10% of parameters, −69.6% flip rate.
- A **four-quadrant safety taxonomy** separating real safety from illusory safety.
- An **entropy → flip bridge** for deployment-time monitoring (AUROC 0.711).

Takeaway

Evaluate medical VLMs on **two axes**: does it answer the same way, and does the answer come from the image. Consistency without verified grounding is a false sense of safety.

PSF-Med + papers