Stable wrongness in medical vision-language models: counterfactual answer sensitivity explained

25 November 2026 · 7 min read · blogs.iictp.org/counterfactual-answer-sensitivity-medical-vlm-stable-wrongness-explained

Short answer: an arXiv preprint submitted on 22 November 2026 by Qi Wu and colleagues shows that medical vision-language models can give wrong yes/no answers that are confident and repeatable even when the patient image is removed or mismatched, a failure the authors call stable wrongness. They propose Counterfactual Answer Sensitivity (CAS), which tests whether an answer still holds when the visual evidence is replaced, and report a macro-averaged AUROC gain of 0.150 for error detection across 7 models and 3 benchmarks. This is an arXiv preprint. It has not been peer reviewed, and its numbers are the authors' own claims.

TL;DR The short version

- Medical vision-language models can give confident, repeatable wrong yes/no answers even when the patient image is removed, a failure the authors call stable wrongness.
- Their counterfactual test checks whether an answer changes when the image is blanked or replaced, and beats sampling-based error detection across 7 models and 3 benchmarks.
- It covers only closed yes/no questions in an arXiv preprint, but shows that high accuracy can come from patterns in the question text.

Read online, listen to it, or change the text size: https://blogs.iictp.org/counterfactual-answer-sensitivity-medical-vlm-stable-wrongness-explained

What is stable wrongness?

Sampling-based uncertainty methods treat disagreement between repeated answers as a sign of error. The abstract argues that this works only if errors are unstable. If a model answers "yes" every time because of a text shortcut, agreement is high and the method is falsely reassured.

How CAS works

- **Counterfactual interventions.** Replace the patient image with text only, a black image or a corrupted image.
- **Evidence sensitivity.** If a committed yes/no answer does not change when visual evidence is invalidated, the answer was not supported by the image.
- **Reliability score.** CAS shifts estimation from answer stability to evidence sensitivity.

What is new compared with earlier approaches

The comparison is with the hallucination-aware calibration (HAC) family, which relies on sampling stability. CAS asks a different question: does the answer depend on the image at all?

Results reported by the source

- Across 7 vision-language models and 3 benchmarks, CAS outperforms the best HAC-family scores in all evaluated settings.
- The macro-averaged AUROC gain for error detection is +0.150.
- Mechanistic analysis shows CAS recovers hidden errors in the region where HAC is confident; the strongest standalone signal is the counterfactual shift of the decision boundary.

Limitations and open questions

- It covers closed-ended yes/no questions; open-ended reporting is a different problem.
- Benchmarks are not clinical deployments, and the abstract does not name the models.
- Extra model calls per question add cost. Sources were read on 25 November 2026. Products, rules and standards change, so check the linked originals before you rely on any detail.

How to apply it in an evaluation

1. For each yes/no test question, create a text-only and a blacked-out version.
2. Flag cases where the answer does not change.
3. Treat those as shortcut-driven even if the answer is correct.
4. Report the share of such cases alongside accuracy.

What it means for clinicians and buyers

A high accuracy score on closed-ended questions can come from patterns in the question text. Ask a vendor whether results hold when the image is removed. See AI for TB screening on chest X-rays for an imaging system with a defined workflow.

Frequently asked questions

- **What is stable wrongness?** Confident, repeatable wrong answers that persist even when the image evidence is removed or mismatched.
- **What is CAS?** A test that checks whether an answer still holds when visual evidence is invalidated.
- **What gain does it report?** A macro-averaged AUROC gain of 0.150 across 7 models and 3 benchmarks.
- **Is it peer reviewed?** No, it is a preprint.

Related reading

- AI for TB screening: chest X-ray, MedGemma, MedSigLIP

- SAFE-Triage
- AI Practitioner certification

> Related credential: AI Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. Wu Q, Liu X, Wang A, Wang R, Chen W, Wang W, Yao Q. Rethinking Closed-Ended Medical VLM Evaluation: From Stable Wrongness to Counterfactual Answer Sensitivity. arXiv 2611.29592, 22 Nov 2026 (preprint). arxiv.org/abs/2611.29592

Published by IICTP, an independent technical credential body, on 25 November 2026. The current version of this post is at https://blogs.iictp.org/counterfactual-answer-sensitivity-medical-vlm-stable-wrongness-explained. More guides: blogs.iictp.org.