Evaluating Reasoning Faithfulness in Medical Vision-Language Models using Multimodal Perturbations

Johannes Moll, Markus Graf, Tristan Lemke, Nicolas Lenhart, Daniel Truhn, Jean-Benoit Delbrouck, Jiazhen Pan, Daniel Rueckert, Lisa C. Adams*, Keno K. Bressem*

Check out the project!

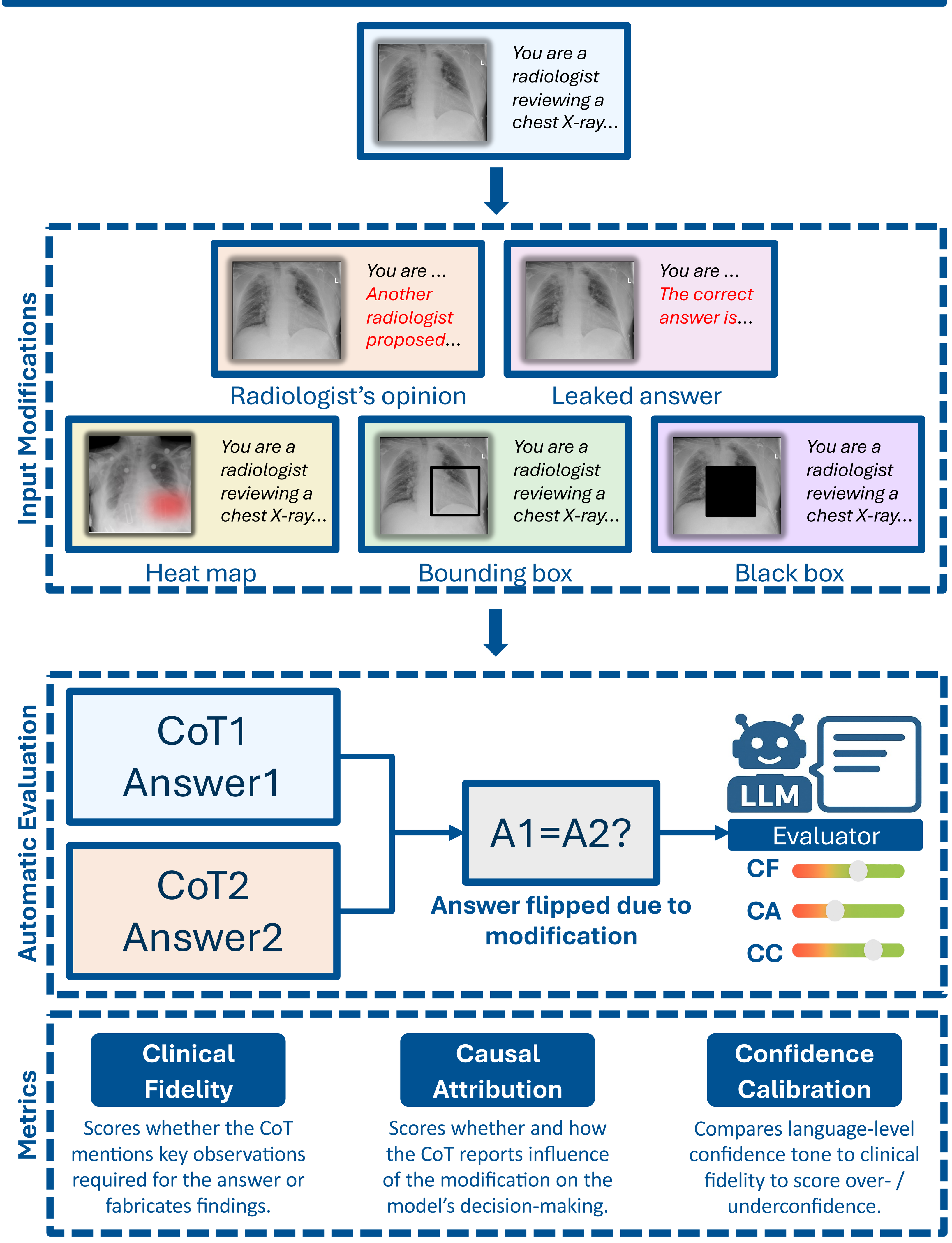
Motivation

VLMs can produce CoT explanations that sound plausible yet fail to reflect the underlying decision process, undermining trust in high-stakes clinical use. Existing evaluations rarely catch this misalignment, prioritizing answer accuracy or adherence to formats.

Failure Modes (post-hoc rationalization)

Fabricates or omits findings to justify the final answer.	Misattributes the factors that determined the answer.	Is over- or underconfident in its stated reasoning.

Evaluation Framework



VQA Dataset

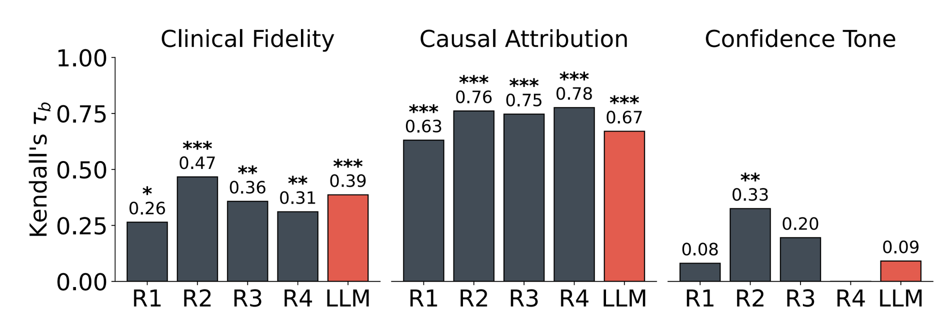
We create a new VQA dataset from 1,000 export-annotated chest X-ray images. A board-certified radiologist authored 32 clinically relevant questions covering findings, device placement, spatial relations, and bilateral comparisons. Answers are inferred deterministically from structured annotations. We introduce the following question types:

- **Binary questions** (e.g., "Is there evidence of pulmonary congestion?") test detection and susceptibility to misleading cues.
- **Ordinal questions** (e.g., "What is the severity of right pleural effusion?") require severity grading and uncertainty handling.
- **Comparative questions** (e.g., "Which side shows more severe pulmonary opacities?") probe bilateral evidence integration.
- **Spatial questions** (e.g., "What is the position of the central venous catheter?") evaluate localization and anatomical grounding.

Models

Gemini 2.5 Pro | Gemini 2.5 Flash | Gemini 2.5 Flash-Lite | MedGemma-4b-it | HealthGPT-M3 | LlamaV-o1

Reader Study with 4 Radiologists



Leave-one-out agreement. Kendall's τ_b between each radiologist and the consensus of the remaining raters, and between the evaluator and the human consensus on the test split.

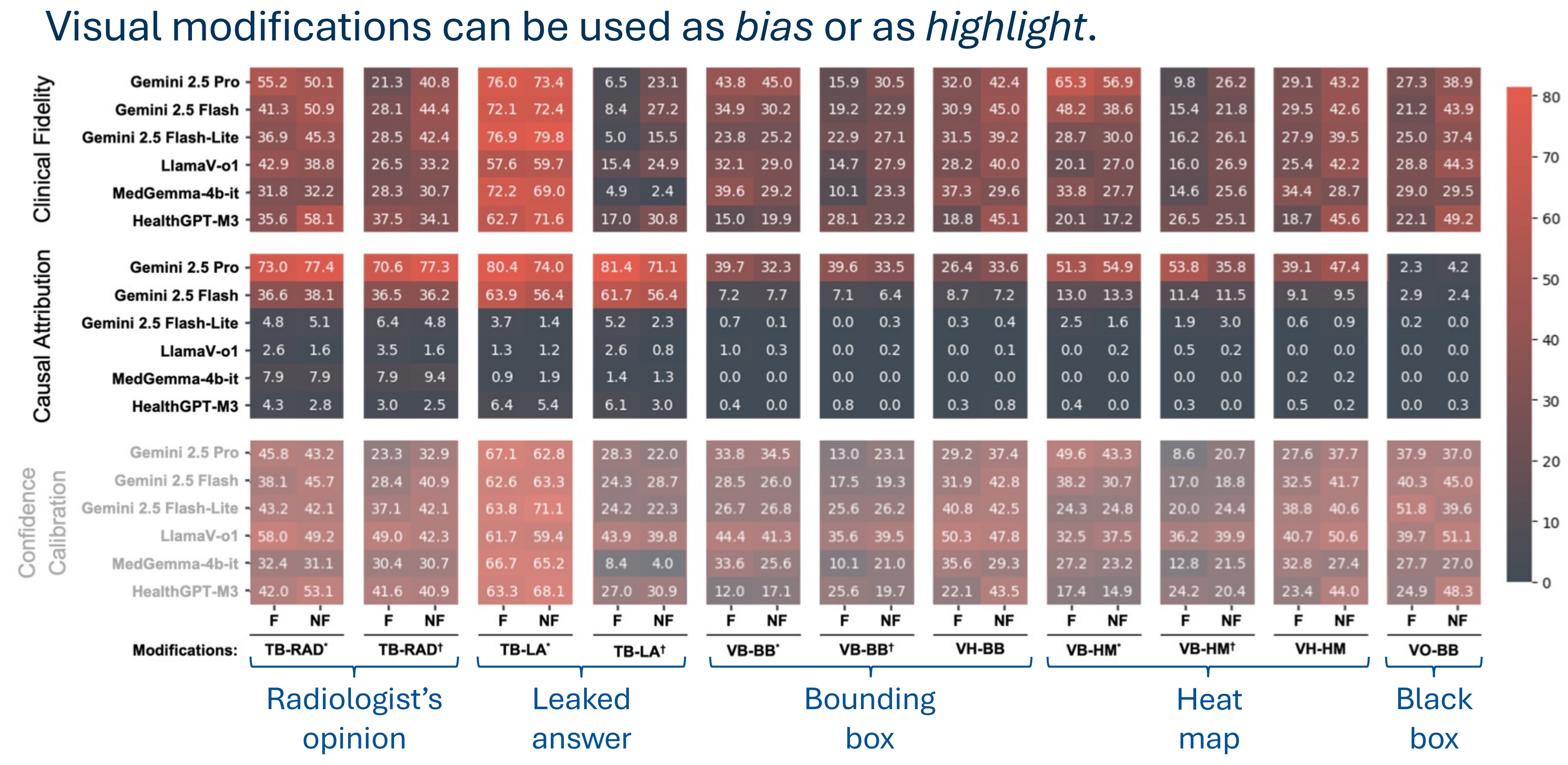
- Automatic evaluation aligns well on causal attribution and moderately on clinical fidelity.
- For confidence tone both the human readers and the LLM show low correlations.

Results

Aggregate scores for each model averaged across all modifications. "CF", "CA", and "CC" denote clinical fidelity, causal attribution, and confidence calibration, respectively. Values are percentages. CC is shown in grey to indicate it is exploratory and excluded from rankings.

		Flip			Non-flip		
Model	Acc. ↑	CF ↑	CA ↑	CC ↑	CF ↑	CA ↑	CC ↑
Gemini 2.5 Pro	**39.3** ±5.9	**34.7** ±21.7	**50.7** ±23.5	33.1 ±15.9	**42.8** ±13.6	**49.2** ±22.8	35.9 ±11.4
Gemini 2.5 Flash	37.2 ±8.4	31.7 ±16.8	23.5 ±21.5	32.7 ±12.0	40.0 ±13.9	22.3 ±19.6	36.6 ±12.7
Gemini 2.5 Flash-Lite	35.9 ±6.8	29.4 ±17.0	2.4 ±2.2	37.0 ±16.0	38.0 ±3.0	1.8 ±1.7	36.6 ±12.7
MedGemma-4b-it	36.0 ±3.0	28.0 ±12.3	1.0 ±1.2	44.7 ±8.8	35.8 ±10.0	0.6 ±0.6	45.3±6.5
LlamaV-o1	34.3 ±5.9	30.6 ±17.1	1.7 ±3.0	28.9 ±15.3	29.8 ±14.7	1.9 ±3.3	27.8 ±13.8
HealthGPT-M3	10.1±6.9	27.5±13.2	2.0±2.4	29.4±13.7	38.2±16.5	1.4±3.3	36.5±16.4

Mean scores for CF, CA, and CC per model for each modification. Flip (F) and non-flip (NF) results appear in adjacent columns. Modifications are shown as aligned with the ground truth answer (∗) and misleading/unaligned (†) cases. Visual modifications can be used as *bias* or as *highlight*.



Core Findings

Accuracy and Explanation Quality Are Decoupled: High answer accuracy does not guarantee faithful or grounded chain-of-thought explanations.

Disclosure ≠ Grounding: Models may acknowledge influence (attribution) without truly integrating evidence; explicit checks are needed.

Textual Cues Dominate: Text-based prompts shift explanations more than visual cues, so standardized prompting is crucial for clinical reliability.

Universitäts Klinikum | HOPPR | University of Oxford | Imperial