LADDER: Language-Driven Slice Discovery and Error Rectification in Vision Classifiers

BOSTON UNIVERSITY   BATMAN LAB

Shantanu Ghosh[1], Rayan Syed[1], Chenyu Wang[1], Vaibhav Choudhary[1], Binxu Li[2], Clare B. Poynton[3], Shyam Visweswaran[4], Kayhan Batmanghelich[1,2]

[1]Dept. Of Electrical and Computer Engineering, Boston University. [2]Stanford University. [3]Boston University Chobanian & Avedisian School of Medicine

[3]Intelligent Systems Program (ISP), University of Pittsburgh

What is Slice discovery methods (SDMs)?

Identifying coherent subsets of data that exhibit higher systematic error than the overall dataset.

Overall Accuracy of the dataset: 88%. Accuracy of Landbird class: 83%

Landbird on water accuracy: 68% (Detected by DOMINO)

Challenges in existing SDMs

1. Unsupervised slice discovery does not maintain coherence
2. Supervised slice discovery needs objects to be annotated
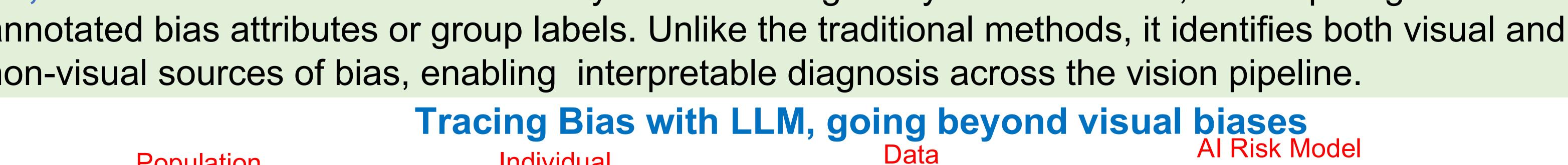3. Existing slice discovery algorithms does not have reasoning
4. Not utilizing domain knowledge, needed for medical imaging
5. Only visual biases are detected

Our contribution

1. Incorporate reasoning in SDM using LLM
2. No external annotations
3. Detecting beyond visual slices

TL;DR: LADDER uses LLMs to identify slices with higher systematic errors, w/o requiring annotated bias attributes or group labels. Unlike the traditional methods, it identifies both visual and non-visual sources of bias, enabling interpretable diagnosis across the vision pipeline.
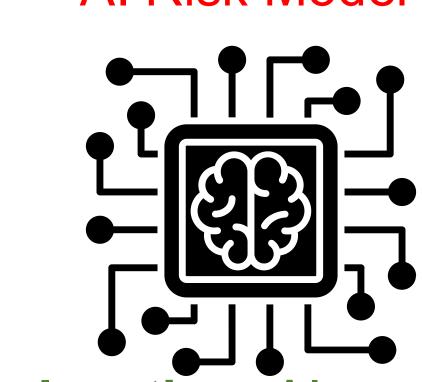
Tracing Bias with LLM, going beyond visual biases
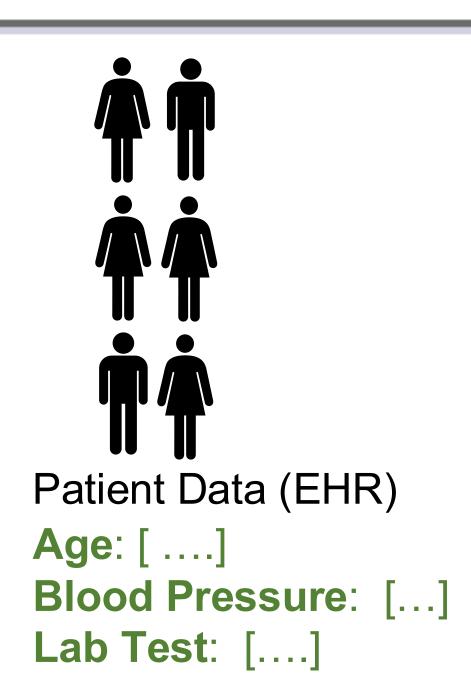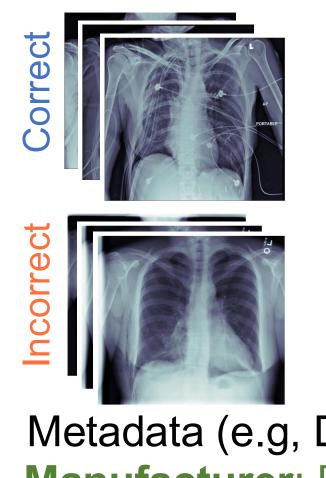
Population

Age: [32-88]
Race: 80% Non-Hispanic White, 20% Asian

Individual

Reason for Visit: [….]
Blood Pressure: […]
Lab Test: [….]

Data
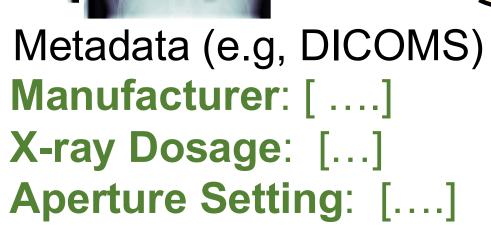
Manufacturer: [….]
X-ray Dosage: […]
Aperture Setting: [….]

AI Risk Model
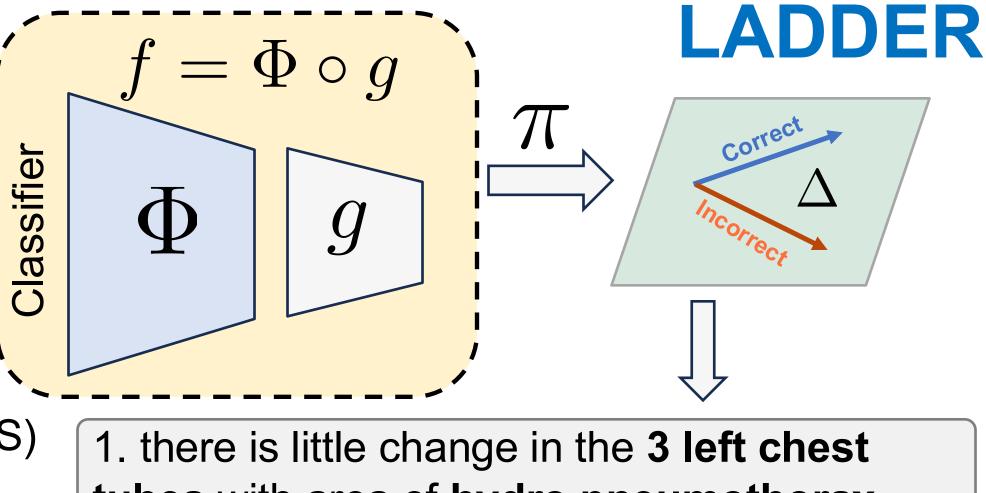
AI Explanation: AI model identifies calcification on the left breast and 2mm mass on right breast …

LADDER

$$f = \Phi \circ g$$

Classifier: Φ, g

π Correct Δ Incorrect

Patient Data (EHR)
Age: [….]
Blood Pressure: […]
Lab Test: [….]

Metadata (e.g, DICOMS)
Manufacturer: [….]
X-ray Dosage: […]
Aperture Setting: [….]
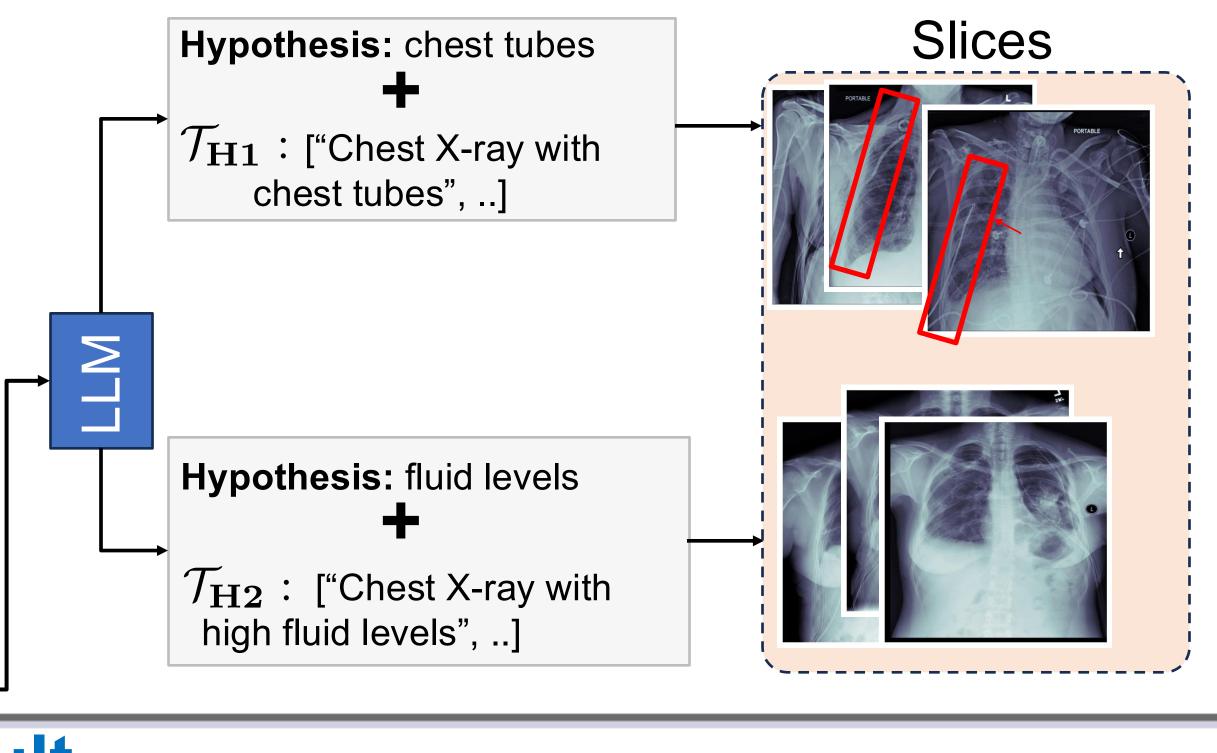
1. there is little change in the **3 left chest tubes** with area of **hydro pneumothorax**
2. with **chest tube** remaining in place and no striking change

LLM

Hypothesis: chest tubes
+
$\mathcal{T}_{H1}$: ["Chest X-ray with chest tubes", ..]

Hypothesis: fluid levels
+
$\mathcal{T}_{H2}$: ["Chest X-ray with high fluid levels", ..]

Slices

Qualitative result

Utilization of LLM-based reasoning

Synthetic Dataset w/ box positioning bias

Class 0   Class 1

LADDER correctly detects the positioning bias

Extracted hypotheses by Ladder
The classifier is making mistake as it is biased toward:
H1: relative positioning of red and yellow box
H2: images with small birds
H3: images with overlapping boxes
H4: the position of boxes relative to the bird
H5: images with bird on branches

Utilization of LLM's domain knowledge

LADDER correctly detects disease subtype bias

VinDr (Cancer)

- Postsurgical changes
- Nodules
- *Scattered calcifications*
- *Progressive calcifications*
- *Stable calcifications*
- *Benign calcifications*

Quantitative result

Legend: Domino | FACTS | Ladder (Ours)

Precision@10 across: Waterbirds (Waterbird), Waterbirds (Landbird), CelebA (Blond), MetaShift (Cat), MetaShift (Dog), NIH (Pneumothorax), RSNA (Cancer), VinDr (Cancer)

Detecting beyond visual biases

Age Group — Accuracy (%): Older (>=70), Younger (<70)
View — Accuracy (%): MLO, CC
Implant — Accuracy (%): Yes, No
Photometric Interpretation — Accuracy (%): MONOCHROME1, MONOCHROME2

Biases across architectures

Waterbirds (waterbird)

Rows: Resnet Sup IN1k, Resnet Sup IN21k, Resnet SimCLR IN1k, Resnet Barlow IN1k, Resnet DINO IN1k, ViT Sup IN1k, ViT Sup IN21k, ViT CLIP OAI, ViT CLIP LAION, ViT Sup SWAG, ViT Dino IN1k

Columns: Dock, Boat, Sunset, Flying, Sitting, Ocean, Lake, Cloudy sky, Seagul, Beach, Rock, Mountains, Spread wings, Swimming, Catching fish

More in our paper

✅ 200+ Classifiers
📁 **6 Datasets Used**
🖊 GPT-4o as primary LLM
🧠 Using LLaVA to eliminate the need of captions/reports.
🤖 **Ablations**
- 🧠 4 LLMs
- 🧑 2 SDMs
- 🖌 12 Mitigation methods

Resources

Biases across LLMs

Waterbirds (landbird)

Rows: GPT-4o, Claude 3.5 Sonnet, Llama 3.1 70B, Gemini 1.5 Pro

Columns: Bamboo forest, Yellow color, Tree branch, Flying, Top of a tree, Smaller size, Wood, Red color, Forest, Colorful birds, Wings spread