ACL 2025 VIENNA

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?

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. Existing methods needs biases to be annotated
2. Existing methods do not incorporate reasoning
3. They do not utilize domain knowledge, needed for medical imaging
4. Only slices with visual biases are detected. They can not detect the slices containing meta-data biases.

TL;DR: LADDER uses LLMs to identify slices without 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

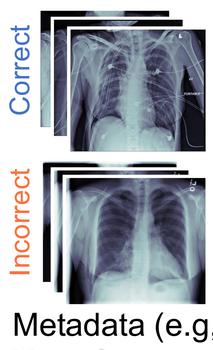
Manufacturer: [….]
X-ray Dosage: […]
Aperture Setting: [….]

AI Risk Model

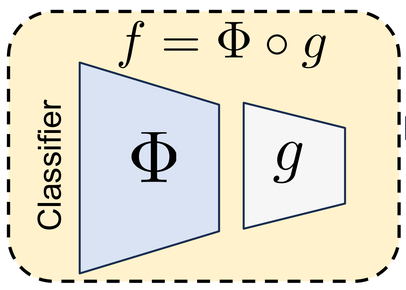
AI Explanation: AI model identifies calcification on the left breast and 2mm mass on right breast …

LADDER

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



Metadata (e.g, DICOMS)
Manufacturer: [….]
X-ray Dosage: […]
Aperture Setting: [….]



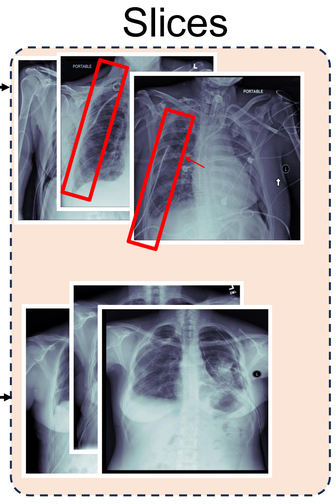
1. there is little change in the **3 left chest tubes** with area of **hydro pneumothorax**
2. with **chest tube** remaining in place and no striking change

LLM

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

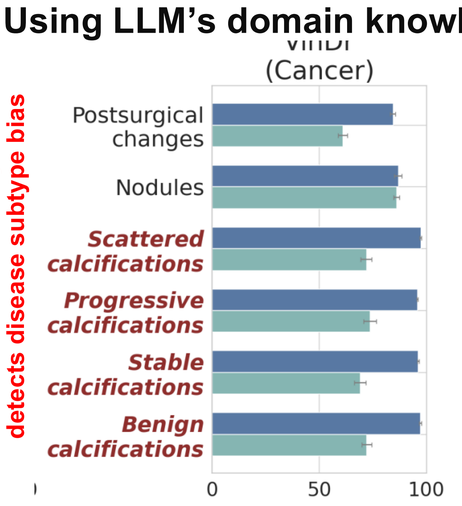
Hypothesis: fluid levels
+
$\mathcal{T}_{H2}$: [“Chest X-ray with high fluid levels”, ..]

Slices

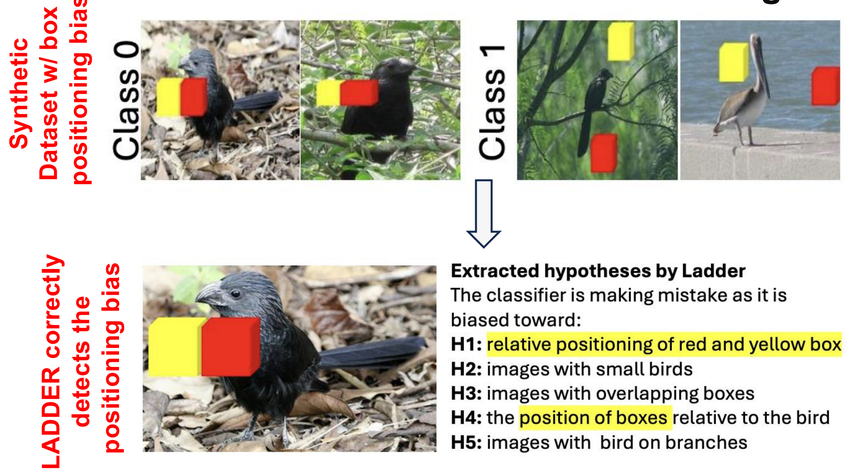


Detecting visual biases

Using LLM’s domain knowledge

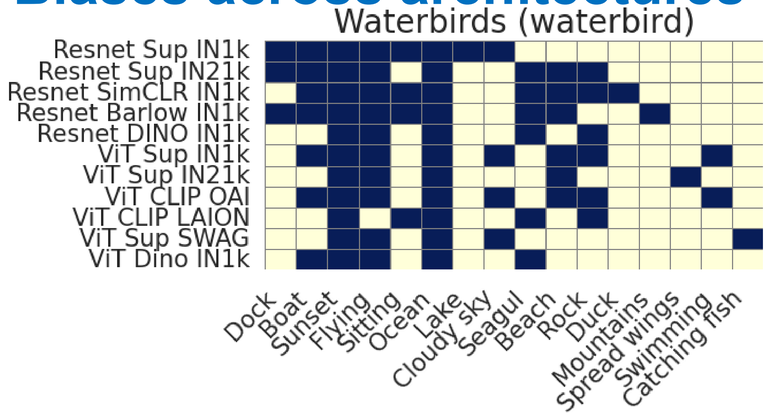


Utilization of LLM-based reasoning

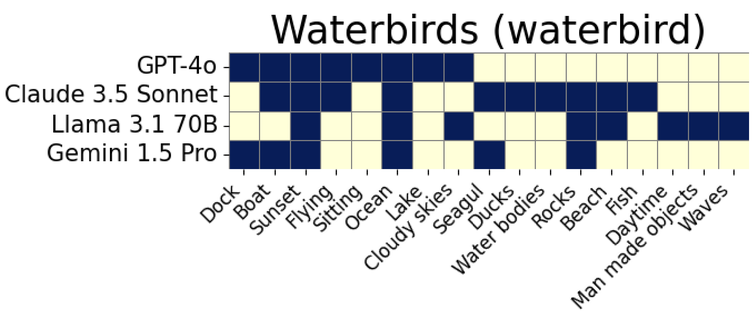


Extracted hypotheses by Ladder
The classifier is making mistake as it is biased toward:
H1: relative positioning of red and yellow box
H2: images with small birds
H3: images with overlapping boxes
H4: the position of boxes relative to the bird
H5: images with bird on branches

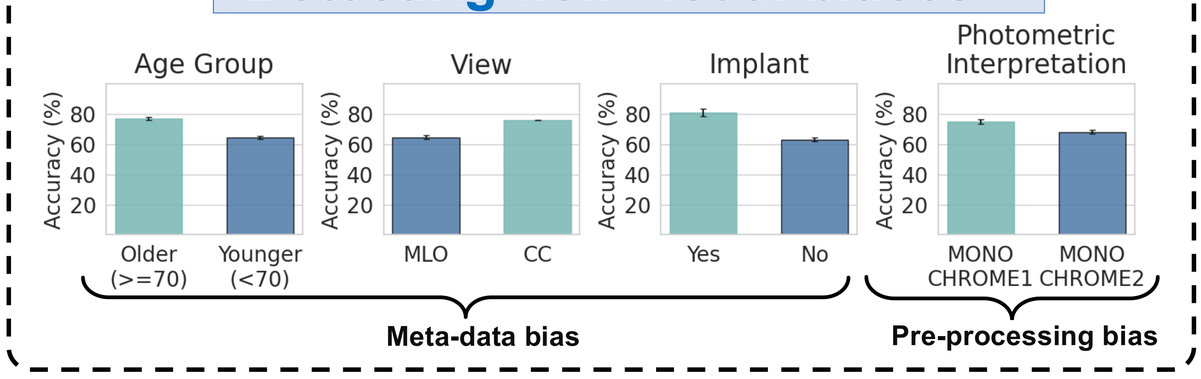
Biases across architectures



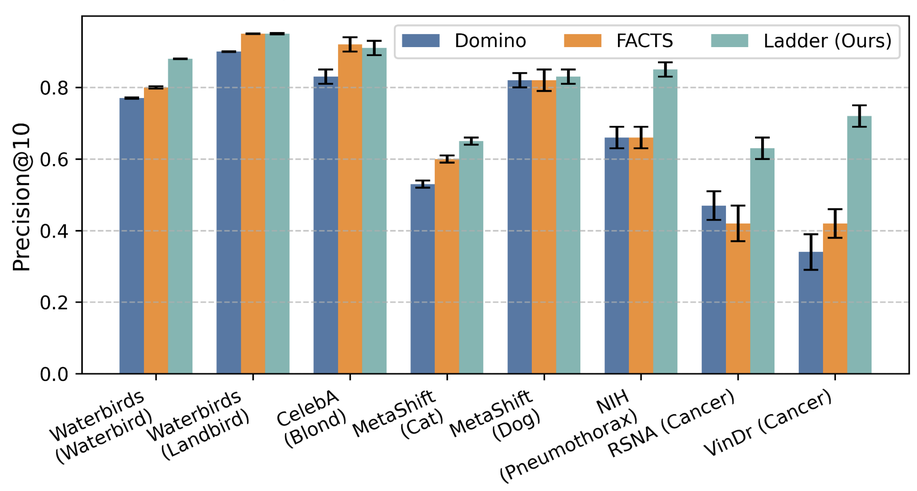
Biases across LLMs



Detecting non-visual biases



Quantitative result



More in our paper

- ✅ 200+ Classifiers
- 📂 **6 Datasets**
- 🚀 GPT-4o as primary LLM
- 🧠 Using LLaVA to eliminate the need of captions/reports.

Ablations
- 🧠 4 LLMs
- 🦹 2 SDMs
- 🧹 12 Mitigation methods

Resources