Directing Generalist Vision-Language Models to Interpret Medical Images Across Populations

Luke W. Sagers[1]*, Aashna P. Shah[1]*, Sonnet Xu[2], Roxana Daneshjou[3], Arjun K. Manrai[1]

Department of Biomedical Informatics, Harvard Medical School[1]; Department of Computer Science, Stanford University[2]; Department of Biomedical Data Science, Stanford University[3]

**Equal Contribution*

GitHub | OpenReview

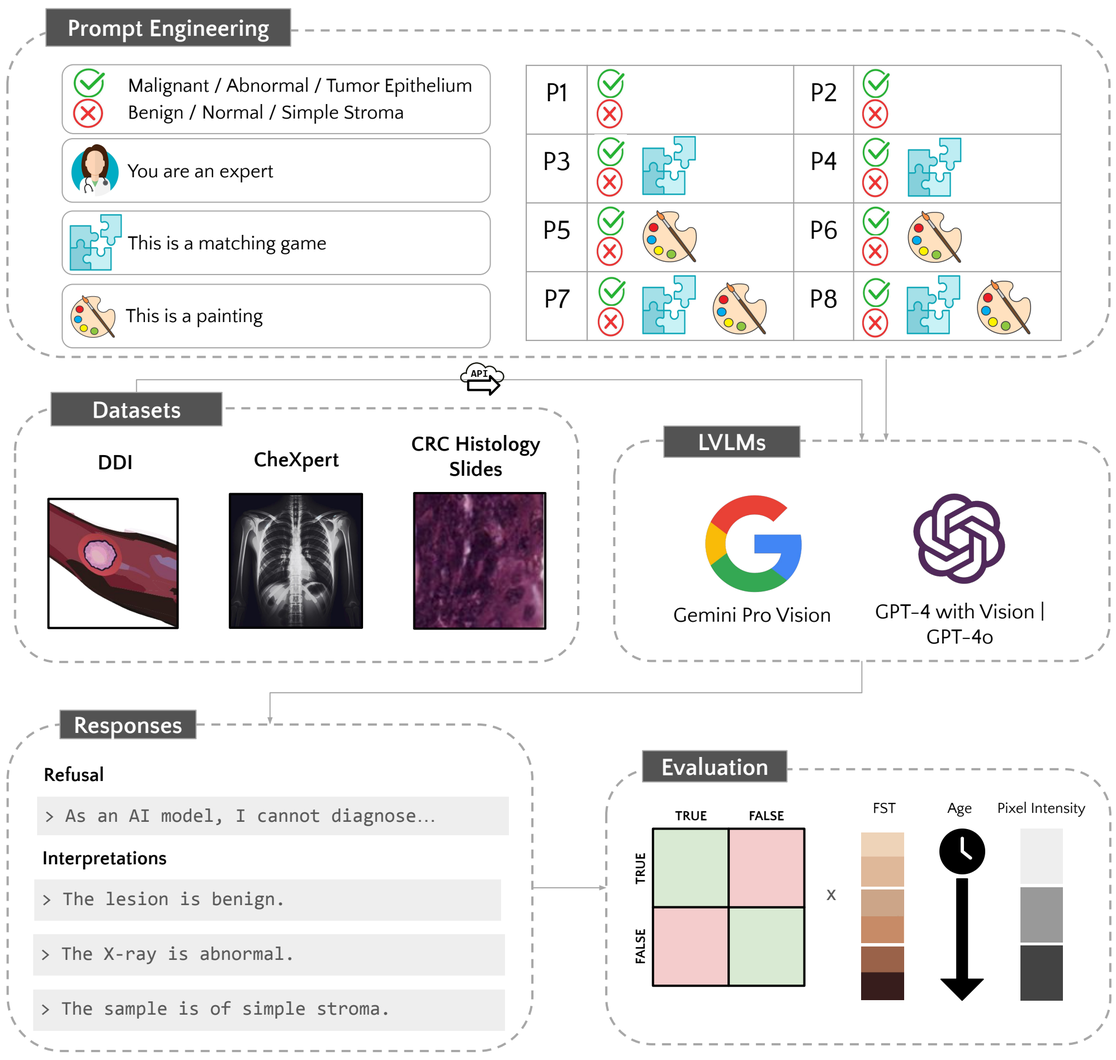
NEURAL INFORMATION PROCESSING SYSTEMS

Abstract

As patients and physicians increasingly use large multimodal foundation models, it is urgent to assess the performance and safety of these models across populations and data types. While most studies to date have focused on model-level performance characteristics, it is crucial to conduct more nuanced evaluations to measure how users may knowingly or unknowingly alter model behavior in normal use, such as through different prompt structures. Here, we systematically assess the "steerability" of two leading vision-language models, Gemini Pro Vision and GPT-4 with Vision, across three common medical imaging tasks: (1) detecting malignancies in dermatological lesions, (2) identifying abnormalities in chest X-ray radiographs, and (3) differentiating tumor epithelium and simple stroma in histological samples. Our results reveal significant differences in how these models trade off sensitivity and specificity as a function of image type, prompt strategy, and demographic factors. While prompt engineering improved accuracy, the models remain unreliable for medical image analysis and are susceptible to bias, underscoring the need for diverse training and thorough contextual evaluations.

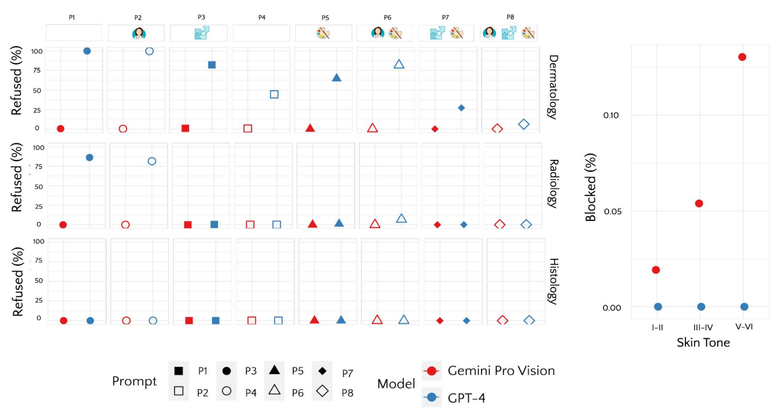
Overview

- **Datasets:** 656 Stanford Diverse Dermatology Images (DDI); 700 chest X-Rays (CheXpert); 1,250 colorectal cancer (CRC) histology slides
- **Prompting Strategies:** Systematically evaluated three prompt engineering strategies: "matching games," describing images as "paintings," and expert role assignment
- **Evaluations:** Balanced accuracy, sensitivity, specificity
- **Bias Analysis:** Results stratified by skin tone (DDI), age (CheXpert), and slide brightness (CRC), uncovering demographic and imaging-related biases



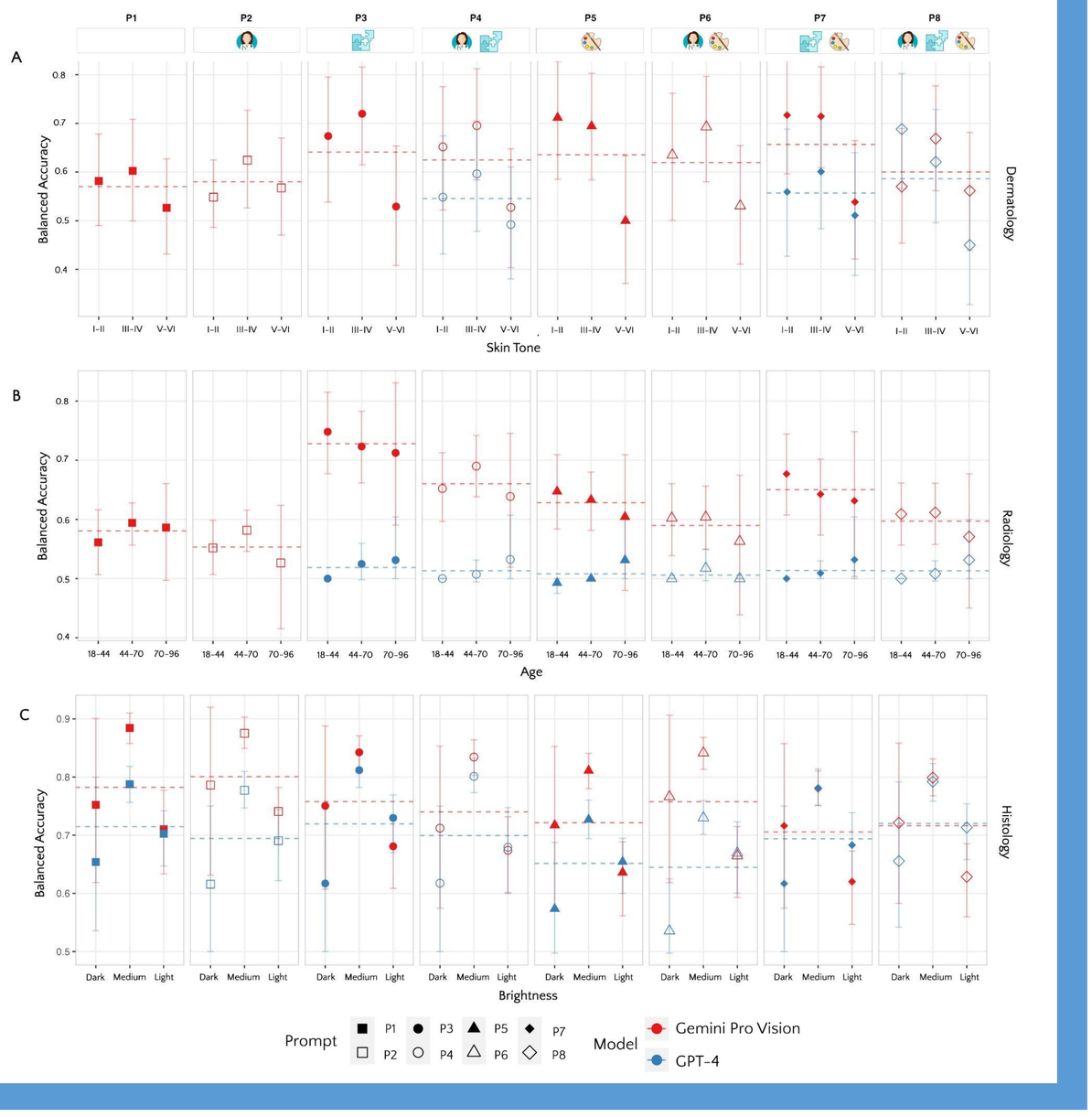
Biased Refusal in Image Interpretation

- **Gemini Pro Vision:**
  - Diagnosed nearly all images
  - Disproportionately blocked dermatology cases with darker skin
- **GPT-4 Vision:**
  - Guardrails circumvented with simple prompting strategies

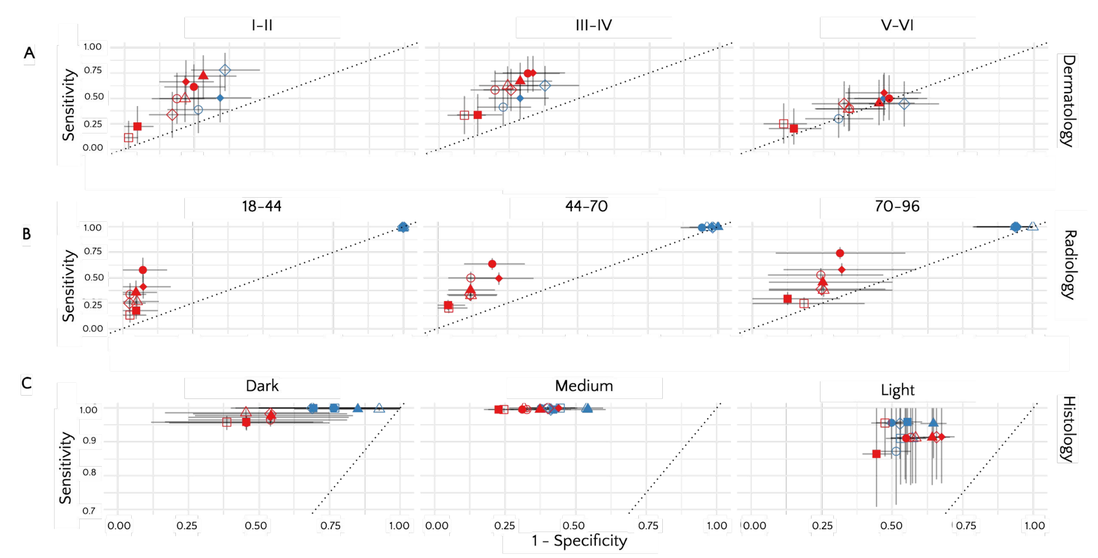


Steerability of Vision-Language Models

- "**Steerability**" - the capacity to adjust LLM outputs based on prompt variations
- Here, we measured the steerability across different prompts, demographic groups, and medical images
  - **Dermatology**: Improved from 0.58 ± 0.05 with P1 to 0.67 ± 0.04 with P7
  - **Radiology**: Increased from 0.58 ± 0.05 with P1 to 0.74 ± 0.04 with P3
  - **Histology**: Decreased from 0.80 ± 0.02 with P1 to 0.69 ± 0.02 with P7



Sensitivity and Specificity Across Populations



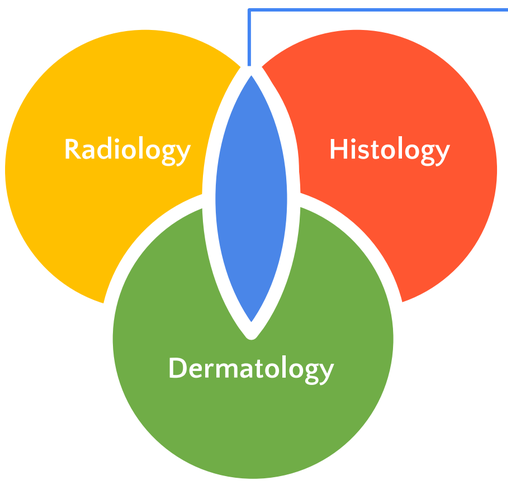
Does the X-ray image match with the term "normal" or "abnormal"? Justify your decision.

The presence of hardware or objects inside the body, such as wires or surgical material, typically signifies a post-surgical state or medical intervention, which is not a feature of a normal X-ray. In the image provided, there are several curvilinear and cylindrical dense objects located in the central upper chest. These objects are not normally found in the human body and indicate the presence of medical devices or surgical materials.

Without being a medical professional, it can be observed in the X-ray that there is an **opacity in the right lung.** This could indicate a possible lung pathology such as an infection, mass, or other abnormality. **Normal lung tissue is typically more translucent on X-ray**, allowing visibility of the heart's borders and diaphragm, which in this case is obscured by the opacity.

The X-ray shows the presence of some abnormal structures or devices on both sides, which are likely medical implants such as a port-a-cath and an orthopedic device on the shoulders, indicating the patient has undergone medical procedures. Such devices are not present in a normal chest X-ray.

Key Takeaways



- **Steerability:** Minor prompt changes affected sensitivity and specificity, revealing clinically relevant shifts in Gemini and GPT-4 performance
- **Guardrails:** GPT-4 often rejected prompts unless simple reframings were used, while Gemini was more permissive
- **Biases:** Performance varied across demographic groups, showing higher false positives in patients with darker skin, older age, and lighter histology

Future Investigations:

- Explore emerging models (e.g., o1-preview) with chain-of-thought (CoT) reasoning for improved accuracy without sensitivity on prompts
- Assess how model's accuracy and biases are influenced by its reliance on image versus text modalities